Università Commerciale "Luigi Bocconi"
PhD School

PhD program in: Statistics
Cycle: XXXV
Disciplinary Field: SECS-S/01

Hierarchical structures in Bayesian Statistics

Advisor: Igor Prünster
Co-Advisor: Antonio Lijoi

PhD Thesis by
Filippo Ascolani
ID number: 3110343

Year: 2024

Abstract

The amazing growth of computational power, storage capacity and data sources opens new exciting frontiers in the processing and analysis of data. This brings up new challenges when modeling phenomena with complex dependence structures, as both statisticians and applied researchers must deal with high dimensional problems and provide accurate inference with a reasonable computational effort. As a consequence, a tradeoff among complexity of the model, its interpretability and accuracy, and availability of efficient algorithms is a difficult, yet crucial, issue in modern data analysis. A theoretical understanding of widely applied methodologies and algorithms is therefore vital to provide convincing guarantees for the quality of the inference.

A natural framework to address the above issues is provided by Bayesian inference. Indeed it combines principled modeling and coherent learning methodology with the availability of sampling schemes and other computational algorithms. In particular, this thesis will focus mostly (though not exclusively) on discrete Bayesian Nonparametric models, which allow for extremely flexible learning mechanisms that can capture complex features of the phenomenon of interest. However, the presence of an infinite dimensional parameter space makes the mathematical and methodological investigation more demanding.

Within this framework, this thesis follows three distinct, but related, directions: (i) modelling complex dependence structures (e.g. time series, multi-samples data...) via a Bayesian nonparametric approach, (ii) mathematical investigation of the resulting inferential procedures, complemented by the proposal of methods for measuring and tuning dependence and proving frequentist asymptotic properties, (iii) rigorous analysis of the computational algorithms employed for posterior inference with the aforementioned structures, with a focus on high dimensional problems. A unifying thread shared by all these lines of research is the study of the specific probabilistic structure considered: indeed, the choice of a particular dependence structure (more specifically hierarchical models), which is often selected through modelling considerations (prior information, domain-specific knowledge, etc.), requires the investigation of the associated inferential and computational properties. Indeed, different specifications may have significantly different levels of analytical tractability and the performance of routinely used MCMC algorithms (e.g. gradient-based methods, Gibbs samplers) may greatly vary.

Foundations of Bayesian learning are discussed in the first Chapter, with a focus on the predictive viewpoint; the relevance of hierarchical structures is also emphasized. Chapters 2 and 3 discuss various classes of hierarchical models, based on different nonparametric priors; both theoretical and methodological aspects are presented. The last Chapter, finally, deals with the computational challenges arising in high dimensional hierarchical models.

Contents

Acknowledgements

Sarebbe un compito arduo menzionare tutte le persone che hanno contribuito a questa tesi e, anche qualora vi riuscissi, una mera pagina non renderebbe loro giustizia. Mi riprometto dunque di ringraziare costoro di persona, ben conscio che quello che ho ricevuto nella mia vita dipende dagli altri molto più di quanto dipenda da me. Mi limito qui a ringraziare coloro ai quali, a causa dei miei limiti empirici, non posso stringere la mano e rivolgere direttamente i miei ringraziamenti.

A Paolo, innanzitutto e soprattutto. Mi piace pensare che, quando alla presa con decisioni importanti, la mia urna di Pólya sia stata rinforzata da te: spero che avremo occasione un giorno di discutere assieme queste pagine.

A Maria Teresa, inoltre. Sebbene non avessimo gli stessi interessi, da te ho visto il vero significato della parola passione: raramente ho imparato di più che dalle tue discussioni.

A Carluccio, poi. Alla mia età, mentre io studiavo spensierato, tu eri in ritirata in mezzo alla neve. Non ci siamo mai incontrati, se non nei racconti, ma spero che in noi tu viva quello che non hai potuto vivere.

Infine i miei pensieri vanno a Ernestina Tarchetti, Mario Pollarolo, Carlo Giuliani, Alfredo Quarantelli, Don Mario Cuniberto, Pierpaolo Merkel, Luciano e Guido Ascolani.

Nulla dies umquam memori vos eximet aevo.

Chapter 1

Introduction to Bayesian statistics: the role of exchangeability

1.1 Introduction

A common way to introduce Bayesian statistics is to say that parameters are treated as random quantities, in constrast with the classical setting where they are unknown, but fixed. This entails that a Bayesian model can be written as

$$X_i \mid \theta \overset{\text{iid}}{\sim} f(x \mid \theta), \quad \theta \sim \pi(\theta), \tag{1.1}$$

where $f(x \mid \theta)$ is the *likelihood* function and $\pi(\theta)$ is the *prior* distribution. Thus, Bayesian inference becomes the study of the *posterior* distribution, which is the law of θ given the data $X_{1:n} = (X_1, \dots, X_n)$ and represents the update of the prior beliefs with the collected information. Thanks to Bayes' Theorem, under suitable regularity conditions, the posterior density can be easily computed as

$$\pi(\theta \mid X_{1:n}) \propto f\left(X_{1:n} \mid \theta\right) \pi(\theta). \tag{1.2}$$

Within this perspective, the relevance of formula (1.2) is that it allows to perform the fundamental inversion: from the effects (i.e. the observations) we want to deduce the causes (i.e. the parameters). Many books start from representation (1.1) to introduce Bayesian methodologies (e.g. Robert (2007); Ghosal and Van der Vaart (2017)) and there are valid reasons to do so: for example, from a decision theoretic standpoint, Bayesian estimators enjoy optimal properties (see Chapters 2, 8 and 9 of Robert (2007)). Moreover, since a Bayesian analysis is based on (1.2), it automatically satisfies the Likelihood principle (e.g. Chapter 1 in Robert (2007)), which is often seen as a natural requirement.

Nevertheless, this is not the only way to look at Bayesian models and certainly not the way I was introduced to the Bayesian way. According to this perspective, in a sense that we will make precise, Bayesian statistics becomes free of Bayes' Theorem: actually, even the formulation (1.1) becomes no more the starting point, but rather the consequence of a deeper principle, both philosophically and mathematically. The goal of this Chapter is to present formally such viewpoint, which I find extremely fascinating: it is therefore a small tribute to all the professors that introduced me to this world, namely (in chronological order) Raffaele Argiento, Matteo Ruggiero, Antonio Lijoi, Igor Prünster and Sonia Petrone. The textbook closest to this point of view is probably Regazzini (1996), which unfortunately has never been translated into English.

The central theme is *prediction*. We are given a set of observations $X_{1:n} = x_{1:n}$, that we assume have been collected under the same experimental conditions, and we want to predict the $(n+1)$-th, i.e. X_{n+1}. Within the classical setting, the usual assumptions of independence and identical distribution make this task conceptually challenging, since the predictive distribution $p(x_{n+1} \mid x_{1:n})$ strictly speaking does not depend on $x_{1:n}$. Instead, notice that this does not happen for Bayesian models as in (1.1), since

$$p(x_{n+1} \mid x_{1:n}) = \int f(x_{n+1} \mid \theta)\pi(\mathrm{d}\theta \mid x_{1:n}), \tag{1.3}$$

which depends on $X_{1:n} = x_{1:n}$ through the posterior distribution. Thus, a classical statistician would probably define a likelihood depending on some parameter θ, choose a suitable estimator $\hat{\theta} = \hat{\theta}(X_{1:n})$ and use it to predict new values: therefore X_{n+1} depends on $X_{1:n}$ only through the estimation of the parameters. An alternative, which we follow here, is to modify the probabilistic assumption on the data, making $p(x_{n+1} \mid x_{1:n})$ directly depend on $x_{1:n}$, as in (1.3). This is where the notion of *exchangeability* comes into play.

We say that a vector $X_{1:n} = (X_1, \dots, X_n)$ is exchangeable if for any permutation π we have that $\left(X_{\pi(1)}, \dots, X_{\pi(n)}\right)$ has the same distribution of $X_{1:n}$. A sequence $\{X_i\}_i$ is exchangeable if $(X_1, \dots, X_n)$ is exchangeable for every n. Loosely speaking, exchangeability means that order does not matter. In the case of binary variables, i.e. $X_i \in \{0, 1\}$, it implies that only the frequencies of 0s and 1s matter, rather than the specific occurrences: vectors of n elements with the same number of successes yield the same distribution. With a statistical perspective, exchangeability seems a reasonable way to formalize the loose expression “under the same experimental conditions”: observing a value at position i or j, with $i, j = 1, \dots, n$, should not affect our inference. Clearly, independence and identical distribution implies exchangeability, but the reverse implication does not hold. Consider observations generated according to model (1.1), which are only conditionally independent, and notice that

$$p(X_1, \dots, X_n) = \int \prod_{i=1}^{n} f(X_i \mid \theta)\pi(\mathrm{d}\theta) = \int \prod_{i=1}^{n} f(X_{\pi(i)} \mid \theta)\pi(\mathrm{d}\theta) = p\left(X_{\pi(1)}, \dots, X_{\pi(n)}\right), \tag{1.4}$$

for every n and permutation π. Thus the sequence $\{X_n\}_n$ with law given by (1.1) is exchangeable. Incredibly enough, also the converse holds: an exchangeable sequence $\{X_i\}_i$ is such that it can be represented as in (1.1). This is the content of the celebrated de Finetti’s Theorem (de Finetti, 1929, 1937; Hewitt and Savage, 1955): an exchangeable distribution can be (uniquely) written as a mixture of independent laws. Therefore, parameters θ arise directly by the assumption of exchangeability (which relates only to the observables) and the prior is exactly the mixing measure. We can say more: an exchangeable law is characterized by the predictive distributions $p(x_{n+1} \mid x_{1:n})$, see Theorem 4 below. Therefore in this sense every Bayesian analysis concerns prediction: a family of (coherent) predictive distributions implies an exchangeable law, which in turn implies representation (1.1) and the existence of θ. This Bayesian focus on prediction, even when inference on the parameters is of interest, is not a novel idea, yet still receives considerable attention, see e.g. Fong et al. (2021); Holmes and Walker (2023); Fortini and Petrone (2023). In the next Sections we will discuss in details de Finetti’s Theorem and its implications for parametric and nonparametric inference: its extension to more involved settings (i.e. partial exchangeability) and the connection with hierarchical modelling is discussed. For a stimulating account of de Finetti’s contributions to Probability and Statistics we refer to Cifarelli and Regazzini (1996), while Kingman (1978) provides an excellent review on the uses of exchangeability. For a more introductory and divulgative treatment, see Diaconis and Skyrms (2018).

1.2 de Finetti's Theorem

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, where the sequence of random variables $\{X_n\}_n$ is defined. The state space, also called *sampling* space, is denoted by $\mathbb{X}$ and hereafter assumed to be Polish, i.e. homeomorphic to a separable complete metric space. The resulting Borel σ-algebra on $\mathbb{X}$ is denoted by $\mathcal{X}$. We define in the usual way the product spaces $\left(\mathbb{X}^{(n)}, \mathcal{X}^{(n)}\right)$ and $\left(\mathbb{X}^{(\infty)}, \mathcal{X}^{(\infty)}\right)$. We will not dwell too much on measure theoretic details, throughout this document we think of $\mathbb{X}$ either being equal to $\mathbb{R}^d$ or to a discrete (finite or countable) space: however, notice that a minimal amount of assumptions on the sampling space is required for de Finetti's Theorem to hold, see Dubins and Freedman (1979) for counterexamples on non standard spaces.

We call $P(\mathbb{X})$ the space of probability measures on $\mathbb{X}$, endowed with $\mathcal{P}(\mathbb{X})$, the smallest σ-algebra on $P(\mathbb{X})$ that makes the maps

$$m_B \,:\, P(\mathbb{X}) \to \mathbb{R}_+, \quad m_B(P) = P(B)$$

measurable, for every $B \in \mathcal{X}$. The definition of a σ-algebra is necessary to construct probability measures on $P(\mathbb{X})$, that will play the role of prior distributions. For every $P \in P(\mathbb{X})$ we define the associated product measures with $P^{(n)} \in P(\mathbb{X}^{(n)})$ and $P^{(\infty)} \in P(\mathbb{X}^{(\infty)})$, whose measurability can be easily proven. Finally, we call *random probability measure* any measurable function P from $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in $\left(P(\mathbb{X}), \mathcal{P}(\mathbb{X})\right)$. We are now ready to state de Finetti's Theorem.

Theorem 1. *The sequence $\{X_n\}_n$ is exchangeable if and only if there exists a probability measure Q on $\left(P(\mathbb{X}), \mathcal{P}(\mathbb{X})\right)$ such that for every n we have*

$$\mathbb{P}\left(X_1 \in A_1, \ldots, X_n \in A_n\right) = \int_{P(\mathbb{X})} \prod_{i=1}^{n} P(A_i) Q(dP), \tag{1.5}$$

for every $A \in \mathcal{X}^{(\infty)}$. Moreover

$$\frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(\cdot) \to Q(\cdot) \tag{1.6}$$

weakly almost surely, as $n \to \infty$.

Equation (1.5) is the formal mathematical translation of the usual representation (1.1): exchangeable sequences can be written as mixture of independent and identically distributed (i.i.d.) sequences, i.e. there exists a random probability measure conditional to which observations are i.i.d., as in (1.1). Notice that if Q is degenerate over a finite dimensional space, we recover the parametric setting. For example, the case

$$Q\left(\left\{P \in P(\mathbb{X}) \,:\, P(\mathrm{d}x) = N(\theta, 1)\mathrm{d}x,\, \theta \in \mathbb{R}\right\}\right) = 1$$

corresponds to the Bayesian model with Gaussian likelihood and a prior on the location parameter θ. Moreover, the limit (1.6) implies that the *de Finetti measure*, or *prior*, can be recovered as the weak limit of the empirical distribution. Thus parameters arise by assumptions on the observables and are identifiable, in the sense that are measurable with respect to the σ-algebra generated by the sequence of observations: loosely speaking, knowing the entire sequence of datapoints is equivalent to knowing the realisation of Q. Notice that, as discussed in (1.4), representation (1.5) can be easily shown to imply exchangeability, so that only the converse implication is of interest. Before providing a formal proof of Theorem 1, we provide some intuition on why it should hold, considering the simple case of binary observations.

1.2.1 de Finetti's Theorem for binary data

Here we assume that $\mathbb{X} = \{0, 1\}$, so that de Finetti's Theorem can be written in the following way.

Theorem 2. *The sequence $\{X_n\}_n$ is exchangeable if and only if there exists a probability measure Q on the unit interval such that*

$$\mathbb{P}\left((X_1, \dots, X_n) = (x_1, \dots, x_n)\right) = \int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} Q(d\theta),$$

for every $(x_1, \dots, x_n) \in \{0, 1\}^n$. Moreover $\frac{1}{n}\sum_{i=1}^n \delta_{X_i}(\cdot) \to Q(\cdot)$ weakly almost surely, as $n \to \infty$.

Therefore, when dealing with binary sequences, a probability of success θ is sampled from Q and the n observations are obtained by drawing with replacement from a un urn with proportion given by θ and $1-\theta$. This is the setting where de Finetti's Theorem has been proven for the first time (de Finetti, 1929), but here we consider the arguments given in Diaconis and Freedman (1980b). More accessible accounts can be found in Heath and Sudderth (1976) and Chapter 7 of Diaconis and Skyrms (2018).

Fix $n \in N$ and let $(X_1, \dots, X_n) \in \{0, 1\}^n$ be an exchangeable random vector. This means that vectors with the same number of successes yield the same probability, i.e.

$$\mathbb{P}\left(X_1 = x_1, \dots, X_n = x_n \mid S_n = r\right) = \begin{cases} \frac{1}{\binom{n}{r}} & \text{if } \sum_{i=1}^n x_i = r \\ 0 & \text{else} \end{cases},$$

where $S_n = \sum_{i=1}^n X_i$. The interpretation is that, conditional on observing r successes, the vector $(X_1, \dots, X_n)$ is obtained by sampling without replacement from an urn with r balls with label 1. Therefore, denoting with $P_r(x_1, \dots, x_n)$ the distribution above, which does not depend on the specific exchangeable law under consideration, we have

$$\begin{aligned} \mathbb{P}\left(X_1 = x_1, \dots, X_n = x_n\right) &= \sum_{r=0}^n P_r(x_1, \dots, x_n)\mathbb{P}(S_n = r) \\ &= \frac{1}{\binom{n}{\sum_{i=1}^n x_i}}\mathbb{P}(S_n = \sum_{i=1}^n x_i), \end{aligned}$$

so that the the distribution of $(X_1, \dots, X_n)$ is a mixture of urn sampling schemes, where the mixing measure is given by the law of the number of successes. Similarly, if $(X_1, \dots, X_N)$ is exchangeable and $n \leq N$, denoting $\sum_{i=1}^n x_i = r$ we have

$$\mathbb{P}\left(X_1 = x_1, \dots, X_n = x_n\right) = \sum_{s=r}^N \mathbb{P}(X_1 = x_1, \dots, X_n = x_n \mid S_N = s)\mathbb{P}(S_N = s).$$

Reasoning as before, conditional on observing s successes out of N trials, the vector $(X_1, \dots, X_n)$ is obtained by sampling n times without replacement from an urn with s balls with label 1 and $N - s$ balls with labels 0. Then by exchangeability we obtain

$$\begin{aligned}\mathbb{P}(X_1 = x_1, \dots, X_n = x_n \mid S_N = s) &= \mathbb{P}(X_1 = 1, \dots, X_r = 1, X_{r+1} = 0, \dots, X_n = 0 \mid S_N = s) \\ &= \frac{s}{N}\frac{s-1}{N-1}\cdots\frac{s-r+1}{N-r+1}\frac{N-s}{N-r}\cdots\frac{N-s-n+r+1}{N-n+1}.\end{aligned}$$

If n and r are fixed, but N grows (i.e. we are closer to an exchangeable sequence), sampling with or without replacement from an urn become very similar. More formally, Theorem 4 in Diaconis and Freedman (1980b) shows that

$$\left|\frac{s}{N}\frac{s-1}{N-1}\cdots\frac{s-r+1}{N-r+1}\frac{N-s}{N-r}\cdots\frac{N-s-n+r+1}{N-n+1} - \left(\frac{s}{N}\right)^r\left(\frac{N-s}{N}\right)^{n-r}\right| \leq 4\frac{n}{N}$$

uniformly over r and s. Thus we obtain

$$\mathbb{P}\left(X_1 = x_1, \dots, X_n = x_n\right) = \sum_{s=r}^{N}\left(\frac{s}{N}\right)^r\left(\frac{N-s}{N}\right)^{n-r}\mathbb{P}(S_N = s) + o_N(1),$$

where $o_N(1)$ is a function $g(N)$ such that $g(N) \to 0$ as $N \to \infty$. Writing $\theta = s/N$ and denoting with $\mu_N(\theta)$ the discrete probability measure supported on $\{1/N, 2/N, \dots, 1\}$ such that $\mu_N(\theta) = \mathbb{P}(S_N = N\theta)$, we have

$$\mathbb{P}\left(X_1 = x_1, \dots, X_n = x_n\right) = \int_0^1 \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\mu_N(\mathrm{d}\theta) + o_N(1). \tag{1.7}$$

Therefore, in order to prove Theorem 2 it suffices to prove that $\mu_N(\mathrm{d}\theta)$ converges weakly to some probability measure Q, which is exactly equivalent to $\frac{1}{N}\sum_{i=1}^N \delta_{X_i}(\cdot) \to Q(\cdot)$ weakly almost surely, as $N \to \infty$. We will show this in the next Section, by proving a Strong Law of Large Numbers for exchangeable sequences. Notice that representation (1.7) implies that de Finetti's Theorem holds approximately if $(X_1, \dots, X_N)$ is an exchangeable vector and n is much smaller than N. For more details on representation of finitely exchangeable laws, see Diaconis and Freedman (1980b).

1.2.2 Laws of Large Numbers for exchangeable sequences

Laws of Large Numbers for independent random variables are a crucial component of the study of statistical methods in the classical sense. A similarly prominent role is played by the corresponding laws for exchangeable sequences: the effect of dependence across the random variables is given by the convergence to a non degenerate random variable.

For every $n \in \mathbb{N}$ a measurable function $f : \mathbb{X}^{(\infty)} \to \mathbb{R}$ is *n-symmetric* if for every permutation π of $\{1, \dots, n\}$ we have

$$f(x) = f\left(x_{\pi(1)}, \dots, x_{\pi(n)}, x_{n+1}\right), \dots x \in \mathbb{X}^{(\infty)}.$$

We denote by $\mathcal{S}_n \in \mathcal{X}^{(\infty)}$ the σ-algebra generated by n-symmetric functions. It is clear that a $(n+1)$-symmetric function is also n-symmetric, so that $\mathcal{S}_{n+1} \subset \mathcal{S}_n$. We also denote with $\mathcal{S} = \lim_{n\to\infty}\mathcal{S}_n = \cap_{n=1}^{\infty}\mathcal{S}_n$ the smallest σ-algebra on $\mathbb{X}^{(\infty)}$ that makes the n-symmetric functions for every $n \in \mathbb{N}$ measurable. We are ready to state the Law of Large Numbers for exchangeable sequences.

Theorem 3. *Let $\{X_n\}_n$ be an exchangeable sequence and $\varphi : \mathbb{X} \to \mathbb{R}$ a measurable function with $\mathbb{E}[|\varphi(X_1)|] < \infty$. Then there exists a random variable $\tilde{\varphi}$ such that*

$$\frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) \to \tilde{\varphi},$$

almost surely and in L^1, as $n \to \infty$. Moreover $\tilde{\varphi} = \mathbb{E}[\varphi(X_1) \mid \mathcal{S}]$ almost surely.

Proof. Let f be a n-symmetric function and take $j \in \{1, \dots, n\}$. Then by exchangeability we have

$$\mathbb{E}\left[\varphi(X_j)f(X)\right] = \mathbb{E}\left[\varphi(X_1)f(X_j, X_2, X_{j-1}, X_1, X_{j+1}, \dots)\right] = \mathbb{E}\left[\varphi(X_1)f(X)\right],$$

where the last equality follows by n-symmetricity. This means that

$$\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left[\varphi(X_j)f(X)\right] = \mathbb{E}\left[\varphi(X_1)f(X)\right],$$

which implies that

$$\int_A \left(\frac{1}{n}\sum_{j=1}^{n} \varphi(X_j)\right) \mathrm{d}\mathbb{P} = \int_A \varphi(X_1)\, \mathrm{d}\mathbb{P} = \int_A \mathbb{E}\left[\varphi(X_1) \mid \mathcal{S}_n\right] \mathrm{d}\mathbb{P}, \quad A \in \mathcal{S}_n$$

where the latter equality holds by definition of conditional expectation. Note that $\frac{1}{n}\sum_{j=1}^{n} \varphi(X_j)$ is $\mathcal{S}_n$-measurable, so that

$$\frac{1}{n}\sum_{j=1}^{n} \varphi(X_j) = \mathbb{E}\left[\varphi(X_1) \mid \mathcal{S}_n\right]$$

almost surely. Denote $Y_n = \mathbb{E}\left[\varphi(X_1) \mid \mathcal{S}_n\right]$ and notice that by the Law of Iterated Expectation we have

$$\mathbb{E}\left[Y_n \mid \mathcal{S}_{n+1}\right] = \mathbb{E}\Big[\mathbb{E}\left[\varphi(X_1) \mid \mathcal{S}_n\right] \mid \mathcal{S}_{n+1}\Big] = \mathbb{E}\left[\varphi(X_1) \mid \mathcal{S}_{n+1}\right] = Y_{n+1},$$

so that $\{Y_n\}_n$ is a reversed martingale with respect to $\{\mathcal{S}_n\}_n$. By the convergence theorem for reversed martingales it holds

$$\frac{1}{n}\sum_{i=1}^{n} \varphi(X_i) = \mathbb{E}\left[\varphi(X_1) \mid \mathcal{S}_n\right] \to \mathbb{E}\left[\varphi(X_1) \mid \mathcal{S}\right],$$

almost surely and in L^1, as $n \to \infty$. □

Let $\varphi = \mathbb{1}_B$. Then Theorem (3) states that

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{X_j}(B) \to \mathbb{E}[\mathbb{1}_B \mid \mathcal{S}] = \mathbb{P}(X_1 \in B \mid \mathcal{S}). \tag{1.8}$$

Moreover, it is easy to show that $\mathcal{S}$ belongs to the tail σ-algebra, i.e. $\mathcal{S} \subset \mathcal{T} = \cap_{n=1}^{\infty}\sigma\left(\{X_i\}_{i\geq n}\right)$, which implies $\mathbb{P}(X_1 \in B \mid \mathcal{S})$ being $\mathcal{T}$-measurable. If the observations X_i's are independent and identically distributed, which is a particular case of exchangeability, by Kolmogorov's 0-1 Law we have $\mathbb{P}(X_1 \in B \mid \mathcal{S}) \in \{0,1\}$ and $\mathbb{P}(X_1 \in B \mid \mathcal{S})$ is degenerate, recovering the standard Law of Large Numbers.

Convergence as in (1.8), though holding for every $B \in \mathcal{X}$, is not enough to prove weak convergence of the empirical measure, that we denote with $\hat{P}_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_j}(\cdot)$. The additional steps strongly depend on the Polishness of $\mathbb{X}$, as detailed in the next corollary.

Corollary 1. *If $\{X_n\}_n$ is an exchangeable sequence on a Polish space, then there exists a random probability measure such that $\hat{P}_n \to \tilde{P}$ weakly almost surely, as $n \to \infty$. Moreover $\tilde{P}(B) = \mathbb{P}(X_1 \in B \mid \mathcal{S})$ almost surely.*

Proof. Since $\mathbb{X}$ is Polish, there exists a sequence of uniformly continuous and bounded functions such that $P_n \to P$ weakly if and only if $\int g_k(x)P_n(\mathrm{d}x) \to \int g_k(x)P(\mathrm{d}x)$ for every $k = 1, 2, \dots$. Moreover, measurability of $\hat{P}_n$ with respect to $\mathcal{P}(\mathbb{X})$ can be easily shown, since the sets $\{P \in P(\mathbb{X}) : P(A) \in C\}$, for every $A \in \mathcal{X}$ and C Borel set of $\mathbb{R}$, form a generating class for $\mathcal{P}(\mathbb{X})$.

Since $\mathbb{X}$ is Polish, there exists a version $\tilde{P}$ of the conditional probability distribution of X_1 given $\mathcal{S}$, such that $\tilde{P}(B) = \mathbb{P}(X_1 \in B \mid \mathcal{S})$. Therefore

$$\mathbb{E}\left[g_k(X_1) \mid \mathcal{S}\right] = \int g_k(x)\tilde{P}(\mathrm{d}x)$$

almost surely for every $k = 1, 2, \dots$. Therefore, by applying Theorem (3) to every k we have

$$\int g_k(x)\hat{P}_n(\mathrm{d}x) = \frac{1}{n}\sum_{i=1}^{n} g_k(X_i) \to \mathbb{E}\left[g_k(X_1) \mid \mathcal{S}\right] = \int g_k(x)\tilde{P}(\mathrm{d}x)$$

almost surely as $n \to \infty$. Thus the result follows. □

An immediate consequence of Corollary 1 is the proof of Theorem 2: indeed it implies weak convergence of measures $\mu_N(\mathrm{d}\theta)$ in (1.7), almost surely as in $N \to \infty$. In the next Section we prove de Finetti's Theorem for a generic sampling space.

1.2.3 Proof of de Finetti's Theorem

The original proof for an arbitrary sampling space is given in Hewitt and Savage (1955). Here instead, we follow the same reasoning of Kingman (1978), which is based on the Strong Law of Large Numbers given in Theorem 3. For an argument based on approximating exchangeable sequences with finite exchangeable vectors, see Diaconis and Freedman (1980b).

Proof of Theorem 1. Let $\tilde{P}$ be the conditional probability distribution of X_1 given $\mathcal{S}$, i.e. $\tilde{P}(B) = \mathbb{P}(X_1 \in B \mid \mathcal{S})$. Such $\tilde{P}$ exists since $\mathbb{X}$ is Polish. By Corollary 1, for every $n \in N$ and $A_1, \dots, A_n \in \mathcal{X}$ we have

$$\begin{aligned}\prod_{i=1}^{n} \tilde{P}(A_i) &= \mathbb{E}\left[\prod_{i=1}^{n} \tilde{P}(A_i) \mid \tilde{P}\right] = \mathbb{E}\left[\prod_{i=1}^{n} \lim_{N\to\infty} \hat{P}_N(A_i) \mid \tilde{P}\right] \\ &= \lim_{N\to\infty} \mathbb{E}\left[\prod_{i=1}^{n} \hat{P}_N(A_i) \mid \tilde{P}\right],\end{aligned}$$

where the latter equality holds by Dominated Convergence Theorem. Notice that

$$\prod_{i=1}^{n} \hat{P}_N(A_i) = \prod_{i=1}^{n} \frac{1}{N} \sum_{i=1}^{N} \delta_{X_i}(A_i) = \frac{1}{N^n} \prod_{i=1}^{n} \sum_{i=1}^{N} \delta_{X_i}(A_i).$$

Denote with $\mathcal{C}$ the set of ordered samples $(j_1, \dots, j_n)$ of n elements in $\{1, \dots, N\}$ with possible repetitions. Similarly, define with $\mathcal{D}$ the subset of $\mathcal{C}$ with no repetitions, i.e. $j_1 \neq \dots \neq j_n$. Thus we can write

$$\prod_{i=1}^{n} \hat{P}_N(A_i) = \lim_{N\to\infty} \frac{1}{N^n} \sum_{\mathcal{D}} \prod_{i=1}^{n} \delta_{X_{j_i}}(A_i) + \lim_{N\to\infty} \frac{1}{N^n} \sum_{\mathcal{C}\setminus\mathcal{D}} \prod_{i=1}^{n} \delta_{X_{j_i}}(A_i)$$

and

$$\prod_{i=1}^{n} \tilde{P}(A_i) = \lim_{N\to\infty} \frac{1}{N^n} \sum_{\mathcal{D}} \prod_{i=1}^{n} \mathbb{E}\left[\delta_{X_{j_i}}(A_i) \mid \tilde{P}\right] + \lim_{N\to\infty} \frac{1}{N^n} \sum_{\mathcal{C}\setminus\mathcal{D}} \prod_{i=1}^{n} \mathbb{E}\left[\delta_{X_{j_i}}(A_i) \mid \tilde{P}\right]. \tag{1.9}$$

Notice moreover that

$$\frac{1}{N^n} \sum_{\mathcal{C}\setminus\mathcal{D}} \mathbb{E}\left[\prod_{i=1}^{n} \delta_{X_{j_i}}(A_i) \mid \tilde{P}\right] \leq \frac{\text{Card}\,(\mathcal{C}\setminus\mathcal{D})}{N^n} = \frac{N^n - N(N-1)\cdots(N-n+1)}{N^n} \to 0, \tag{1.10}$$

as $N \to \infty$. Combining (1.9) and (1.10), by exchangeability we have

$$\begin{aligned}
\prod_{i=1}^{n} \tilde{P}(A_i) &= \lim_{N\to\infty} \frac{1}{N^n} \sum_{\mathcal{D}} \mathbb{E}\left[\delta_{X_{j_1}}(A_1) \cdots \delta_{X_{j_n}}(A_n) \mid \tilde{P}\right] \\
&= \lim_{N\to\infty} \frac{1}{N^n} \sum_{\mathcal{D}} \mathbb{P}\left(X_{j_1} \in A_1, \dots X_{j_n} \in A_n \mid \tilde{P}\right) \\
&= \lim_{N\to\infty} \frac{\text{Card}(\mathcal{D})}{N^n} \mathbb{P}\left(X_1 \in A_1, \dots X_n \in A_n \mid \tilde{P}\right) \\
&= \mathbb{P}\left(X_1 \in A_1, \dots X_n \in A_n \mid \tilde{P}\right).
\end{aligned}$$

Therefore, if Q is the law of $\tilde{P}$, we have

$$\begin{aligned}
\mathbb{P}\left(X_1 \in A_1, \dots, X_n \in A_n\right) &= \int_{P(\mathbb{X})} \mathbb{P}\left(X_1 \in A_1, \dots X_n \in A_n \mid \tilde{P}\right) Q(\mathrm{d}\tilde{P}) \\
&= \int_{P(\mathbb{X})} \prod_{i=1}^{n} \tilde{P}(A_i) Q(\mathrm{d}\tilde{P}),
\end{aligned}$$

as desired. □

Uniqueness of representation (1.5) follows by uniqueness of the weak limit, since Q is given by the limit of the empirical distribution.

1.3 Consequences of de Finetti's Theorem

As mentioned before, de Finetti's Theorem and in particular representation (1.5) yield a profound philosphical meaning. Statistical models as in (1.1) follow as a consequences of the exchangeability assumption: thus, the existence of parameters arise by a suitable homogeneity requirement on the observables. It is beyond the scope of this thesis (and my personal knowledge) to discuss in detail the philosophical consequences and issues of such viewpoint: we refer to Section 2 of Cifarelli and Regazzini (1996) for an extensive treatment of de Finetti's interpretation of the representation theorem and the links with the problem of induction.

The second part of the Theorem, i.e. the convergence (1.6) of the empirical measure, says that the prior measure can be derived by an infinite sequence of exchangeable observations. Thus, specifying a Bayesian model is equivalent to specify how prediction is performed, i.e. a family of transition kernels

$$p_n(A; x_{1:n}) = \mathbb{P}\left(X_{n+1} \in A \mid X_1 = x_1, \dots, X_n = x_n\right). \tag{1.11}$$

The next theorem, due to Fortini et al. (2000), shows that, at least in principle, providing suitable predictive distributions is equivalent to exchangeability.

Theorem 4. *A sequence of transition kernels $\{p_n\}_n$ as in* (1.11) *identifies the law of an exchangeable sequence $\{X_n\}_n$ if and only if*

1. $p_n(A; x_{1:n}) = p_n\left(A; x_{\pi(1)}, \dots, x_{\pi(n)}\right)$ *for every $n \in \mathbb{N}$, $A \in \mathcal{X}$ and permutation π of $\{1, \dots, n\}$.*

2. *For every $A, B \in \mathcal{X}$ it holds*

$$\int_A p_{n+1}(B; x_1, \dots, x_n, x_{n+1})\, p_n(dx_{n+1}; x_{1:n}) = \int_B p_{n+1}(A; x_1, \dots, x_n, x_{n+1})\, p_n(dx_{n+1}; x_{1:n}).$$

While the first requirement of Theorem 4, i.e. that prediction should not depend on the order of the collected datapoints, is easily satisfied, the second one is less intuitive. It can be rewritten as

$$\mathbb{P}\left(X_{n+1} \in A, X_{n+2} \in B \mid X_{1:n} = x_{1:n}\right) = \mathbb{P}\left(X_{n+1} \in B, X_{n+2} \in A \mid X_{1:n} = x_{1:n}\right)$$

for every $A, B \in \mathcal{X}$ and it is in general hard to come up with sequences of predictive distributions that satisfy such condition. Therefore the usual route, as we will see in the next Sections, is to exploit de Finetti's Theorem and pass through the definition of a prior law Q on $\left(P(\mathbb{X}), \mathcal{P}(\mathbb{X})\right)$. This justifies the usual way of formulating a Bayesian model as in (1.1). An alternative, that we do not consider here, is to consider a different (weaker) type of dependence among the observations that allows to define a predictive rule in a simpler way, see e.g. Holmes and Walker (2023); Fortini and Petrone (2023); Berti et al. (2023).

1.4 Parametric setting and Bayes' Theorem

As explained in the Introduction and justified in the previous Section, it is customary to start a Bayesian analysis by explicitly defining a joint law of the parameters and the observations. Formally, denoting with Θ the parameter space, which we assume to be Polish with Borel σ-

algebra $\mathcal{B}$, the pair (X, θ) has joint distribution

$$\mathbb{P}(X \in A, \theta \in B) = \int_B P_\theta(A)\Pi(\mathrm{d}\theta), \tag{1.12}$$

where $\Pi \in P(\Theta)$ is the *prior* distribution and $P_\theta \in P(\mathbb{X})$ is the *likelihood.* In the previous Sections we had $\Theta = P(\mathbb{X})$ and $\mathcal{B} = \mathcal{P}(\mathbb{X})$, so that $P_\theta(A) = \theta(A)$, for every $\theta \in \Theta$. We use a different notation here, since we want to restrict to the parametric finite dimensional case. The following discussion is based mostly on Chapters $2 - 3$ of Regazzini (1996) and Chapter 1 of Szabó and van der Vaart (2023).

In order to formalize representation (1.12) we need to define the notion of a *Markov kernel* Q from $(\Theta, \mathcal{B})$ to $(\mathbb{X}, \mathcal{X})$ as a map $Q : \Theta \times \mathcal{X} \to [0, 1]$ such that

(i) The map $A \to Q(\theta, A)$ is a probability measure for every $\theta \in \Theta$.

(ii) The map $\theta \to Q(\theta, A)$ is measurable for every $A \in \mathcal{X}$.

Thus we assume that the statistical model $\{P_\theta\,;\, \theta \in \Theta\}$ is such that $Q(\theta, A) = P_\theta(A)$ is a Markov kernel from $(\Theta, \mathcal{B})$ to $(\mathbb{X}, \mathcal{X})$, which formalizes P_θ being the conditional distribution of X given θ. Requirement (ii) of the definition of Markov kernel makes the right hand side of (1.12) well defined. It is possible to show that there exists a suitable probability space on which the pair (X, θ) can be constructed: see Szabó and van der Vaart (2023) for details.

Thus, given a statistical model $\{P_\theta\,;\, \theta \in \Theta\}$ defined as a Markov kernel and a prior distribution Π, we define the *posterior distribution* as a Markov kernel $Q(x, B) = \Pi(B \mid x)$ from $(\mathbb{X}, \mathcal{X})$ to $(\Theta, \mathcal{B})$ such that

$$\mathbb{P}(X \in A, \theta \in B) = \int_A \Pi(B \mid x)P(\mathrm{d}x), \tag{1.13}$$

where $P(A) = \mathbb{P}(X \in A) = \int_\Theta P_\theta(A)\pi(\mathrm{d}x)$ is the marginal distribution of $\mathbb{X}$ induced by the joint distribution (1.12). Notice that representation (1.13) looks very similar to representation (1.12), with opposite roles played by X and θ: therefore the posterior distribution is defined through a suitable disintegration of the joint distribution of the pair (X, θ). It is possible to show that a posterior distribution exists as soon as Θ is Polish, see Theorem 1.3 in Szabó and van der Vaart (2023). Ancillarly, notice that this justifies taking $\Theta = P(\mathbb{X})$, since the latter is a Polish space if $\mathbb{X}$ is Polish.

The problem now becomes how to compute the posterior distribution as defined in (1.13): the relevance of parametric models and Bayes' Theorem stems from this issue. Indeed, assume there exist a σ-finite measure μ on $(\mathbb{X}, \mathcal{X})$ and a measurable map $(x, \theta) \to p_\theta(x)$ such that

$$P_\theta(A) = \int_A p_\theta(x)\mu(\mathrm{d}x) \tag{1.14}$$

for every $A \in \mathcal{X}$. If this holds we say that the statistical model is *dominated*: usual choices for μ are the Lebesgue measure on $\mathbb{R}^d$ or the counting measure. In this case Bayes' formula states

$$\Pi(B \mid x) = \begin{cases} \frac{\int_B p_\theta(x)\Pi(\mathrm{d}\theta)}{\int_\Theta p_\theta(x)\Pi(\mathrm{d}\theta)} & \text{if } \int_\Theta p_\theta(x)\Pi(\mathrm{d}\theta) > 0 \\ 0 \quad \text{else} \end{cases} \tag{1.15}$$

The next theorem shows that (1.15) defines a version of the posterior distribution; we use the term *version* to emphasize that the posterior distribution is uniquely defined only up to a null set, with respect to the marginal P. Notice moreover that if $\int_\Theta p_\theta(x)\Pi(\mathrm{d}\theta) = 0$ the choice of the value of $\Pi(B \mid x)$ is irrelevant, as the following proof shows. The following proof is well known, we follow the lines of Proposition 1.8 in Szabó and van der Vaart (2023).

Theorem 5. *Let a σ-finite measure μ on $(\mathbb{X}, \mathcal{X})$ and a measurable map $(x,\theta) \to p_\theta(x)$ be such that* (1.14) *holds for every $A \in \mathcal{X}$. Then formula* (1.15) *gives a version of the posterior distribution* $\Pi(B \mid x)$.

Proof. It is not difficult to prove that $(x, B) \to \Pi(B \mid x)$ is a Markov kernel: in particular, requirement (ii) follows by Fubini's Theorem. By (1.12) and again Fubini's Theorem we have

$$\begin{aligned}\mathbb{P}(X \in A, \theta \in B) &= \int_B P_\theta(A)\Pi(\mathrm{d}\theta) = \int_B \int_A p_\theta(x)\mu(\mathrm{d}x)\Pi(\mathrm{d}\theta) \\ &= \int_A \int_B p_\theta(x)\Pi(\mathrm{d}\theta)\mu(\mathrm{d}x),\end{aligned}$$

for every $A \in \mathcal{X}$ and $B \in \mathcal{B}$. In particular, with $B = \Theta$, we have that $p(x) = \int_\Theta p_\theta(x)\mu(\mathrm{d}x)$ is the density of P with respect to μ. Therefore, the set $N = \{x : p(x) = 0\}$ is such that $P(N) = 0$. Therefore, for every $x \in N^c$ we can write

$$\int_B p_\theta(x)\Pi(\mathrm{d}\theta) = \Pi(B \mid x)p(x),$$

with $\Pi(B \mid x)$ as in (1.15). Therefore we can write

$$\mathbb{P}(X \in A, \theta \in B) = \int_{A\cap N^c} \Pi(B \mid x)p(x)\mu(\mathrm{d}x) + \int_{A\cap N}\int_B p_\theta(x)\Pi(\mathrm{d}\theta)\mu(\mathrm{d}x).$$

Since $P(N) = 0$, disintegration (1.13) holds. □

Formula (1.15) is thus the cornerstone of Bayesian parametric models. However, if the support of Π is large, e.g. it is equal to $P(\mathbb{X})$, the assumption of dominated statistical model is often not satisfied. For instance let Θ be the set of discrete probability measures on $\mathbb{R}$ and $P_\theta(A) = \theta(A)$, with θ of the form

$$\theta = \sum_{i\geq 1} W_i \delta_{Z_i}, \tag{1.16}$$

where $\{W_i\}_i$ are random probability weights and Z_i are random atoms sampled i.i.d. by a fixed distribution $Q_0 \in P(\mathbb{R})$. Thus, fixing θ means specifying weights and atoms. It is clear that P_θ is not dominated by the Lebesgue measure, for every θ. Moreover, for every discrete measure μ there exists θ such that the supports of μ and θ are disjoint. Therefore formula (1.15) can not be applied to models as in (1.16) and we need to rely on other tools to obtain the posterior distribution. This is the topic of the next Section.

1.5 Bayesian nonparametrics and the Dirichlet process

Reiterating from the previous discussion, when the statistical model is dominated we can rely on Bayes' formula (1.15) to obtain the posterior distribution. When the model is not dominated, and this usually happens when Θ is infinite dimensional, two problems arise: first of all it is a challenging task to define a meaningful probability measure (i.e. the *prior*) on an infinite dimensional space, moreover the computation of the posterior distribution requires ad hoc reasoning. In this Section we focus on the setting induced by the Dirichlet process (Ferguson, 1973), which solves at once the aforementioned issues. We will then have the first example of Bayesian nonparametric model: we will call it *discrete*, because its realizations are almost surely discrete probability measures. Throughout this thesis we will fit into this framework, but other prossibilities are available (e.g. models based on Gaussian processes, see Williams and Rasmussen (2006)). Notice that the term *nonparametric* refers not to a lack of parameters, but rather to infinitely many of them, in order to span an infinite dimensional space.

1.5.1 Constructing infinite dimensional priors through projections

Combining the notations used in the previous Sections, our parametric space is $\Theta = P(\mathbb{X})$ and we want to define a probability measure $Q \in P(\Theta)$ whose realizations are denoted by $P \in P(\mathbb{X})$. Therefore we can say that P is a random probability measure, i.e. a measurable function from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(P(\mathbb{X}), \mathcal{P}(\mathbb{X}))$.

Notice that we can equivalently describe P as a stochastic process over sets, i.e. $\{P(A) : A \in \mathcal{X}\}$, so that specifying the law of the prior means specifying the law of the process. In particular, for every ordered collection of sets $(A_1, \dots, A_k)$, with $A_i \in \mathcal{X}$, we can define

$$Q_{A_1,\dots,A_k}(C) = \mathbb{P}\left((P(A_1), \dots, P(A_k)) \in C\right)$$

for every $C \in \mathcal{B}\left([0,1]^k\right)$, where $\mathcal{B}\left([0,1]^k\right)$ denotes the Borel σ-algebra on $[0,1]^k$. Spanning over the sets $(A_1, \dots, A_k)$ we obtain the collection $Q = \{Q_{A_1,\dots,A_k} : A_1, \dots, A_k \in \mathcal{X}, k \geq 1\}$. It is easy to prove that Q satisfies the following properties:

$(P1)$ If π is a permutation of $\{1, \dots, k\}$ and $\pi C = \left\{(x_{\pi(1)}, \dots, x_{\pi(k)}) : (x_1, \dots, x_k) \in C\right\}$ then

$$Q_{A_1,\dots,A_k}(C) = Q_{A_{\pi(1)},\dots,A_{\pi(1)}}(\pi C)$$

for every $C \in \mathcal{B}\left([0,1]^k\right)$.

$(P2)$ $Q_{\mathbb{X}}(C) = \delta_1(C)$ for every for every $C \in \mathcal{B}([0,1])$.

$(P3)$ For every refinement $(B_1, \dots, B_n)$ of $A_1, \dots, A_k)$, i.e.

- $(B_1, \dots, B_n)$ is a partition of $\mathbb{X}$ into $\mathcal{X}$-sets;
- Any set $A_j = \cup_{(j)} B_i$, where $(j) = \{i \in \{1, \dots, n\} : B_i \subset A_j\}$;

then

$$Q_{A_1,\dots,A_k}(C) = Q_{B_1,\dots,B_n}\left(\left\{x \in [0,1]^n : \left(\sum_{(1)} x_i, \dots, \sum_{(k)} x_i\right) \in C\right\}\right)$$

for every $C \in \mathcal{B}\left([0,1]^k\right)$.

$(P4)$ For every $\{A_n\}_n$ in $\mathcal{X}$ monotonically decreasing to the empty set $\emptyset$, then $Q_{A_n} \to 0$ weakly as $n \to \infty$.

A well-known result, oftentimes called Kolmogorov's Extension Theorem, says that also the converse holds, i.e. a collection Q satisfying $(P1)$ – $(P4)$ defines a random probability measure. We state the result below for definiteness.

Theorem 6. *If $Q = \{Q_{A_1,\dots,A_k} \,:\, A_1,\dots,A_k \in \mathcal{X}, k \geq 1\}$ satisfies $(P1) - P4)$ then there exists a unique probability measure Q on $(P(\mathbb{X}), \mathcal{P}(\mathbb{X}))$ whose finite dimensional projections are in Q. Moreover, there exists a random probability measures P with probability distribution Q.*

Thus, we can use Theorem 6 to define a prior over $\mathbb{P}(\mathbb{X})$ working just on the finite dimensional distributions. This is the strategy exploited in (Ferguson, 1973) to construct a nonparametric prior, i.e. the Dirichlet process, for the first time. However, we need first of all to state some useful properties of the Dirichlet distribution, which will play the role of finite dimensional projections.

1.5.2 Dirichlet distribution and basic properties

Let $(Y_1,\dots,Y_n)$ be independent random variables such that $Y_j \sim \mathrm{Ga}(\alpha_j, 1)$, with $\alpha_j \geq 0$. Defining $W_j = Y_j / \sum_{i=1}^n Y_i$, it is possible to show that the vector $(W_1,\dots,W_{n-1})$ has density

$$f(w_1,\dots,w_{n-1}) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_n)}{\prod_{i=1}^n \Gamma(\alpha_i)} \prod_{i=1}^{n-1} w_i^{\alpha_i - 1} \left(1 - \sum_{i=1}^{n-1} w_i\right)^{\alpha_n - 1} \mathbb{1}_{\Delta_{n-1}}(w_1,\dots,w_{n-1}),$$

where $\Delta_{n-1} = \left\{(w_1,\dots,w_{n-1}) \,:\, w_i \geq 0, \sum_{i=1}^{n-1} w_i \leq 1\right\}$. We say that $(W_1,\dots,W_{n-1})$ has the law of a Dirichlet distribution and we write $(W_1,\dots,W_{n-1}) \sim D_{n-1}(\alpha_1,\dots,\alpha_n)$, with density $d_{n-1}(\underline{w}\,;\, \alpha_1,\dots,\alpha_n)$.

It is easy to show that $W_1 \sim \mathrm{Beta}(\alpha_1, \alpha_2)$ and, if $0 < r_1 < \cdots < r_l = n$, with $l < n$, it holds

$$\left(\sum_{i=1}^{r_1} W_i, \dots, \sum_{i=r_{l-1}+1}^{r_l} W_i\right) \sim D_{l-1}\left(\sum_{i=1}^{r_1} \alpha_i, \dots, \sum_{i=r_{l-2}+1}^{r_{l-1}} \alpha_i\right). \tag{1.17}$$

Moreover the mean can be easily computed as

$$\mathbb{E}\left[W_j\right] = \frac{\alpha_j}{\sum_{i=1}^n \alpha_i}, \quad j \in \{1,\dots,n\}. \tag{1.18}$$

Assume we have observations $(X_1,\dots,X_N)$ taking values in $\{1,\dots,n\}$. A typical Bayesian exchangeable model is given by

$$\mathbb{P}\left(X_i = j \mid (W_1,\dots,W_n)\right) = W_j, \quad (W_1,\dots,W_{n-1}) \sim D_{n-1}(\alpha_1,\dots,\alpha_n),$$

where $W_n = 1 - \sum_{i=1}^{n-1} W_i$. Using Bayes's formula it is not difficult to show that the model above is *conjugate*, i.e. the posterior distribution is again Dirichlet distributed. More precisely we have

$$(W_1,\dots,W_{n-1}) \mid X_{1:N} \sim D_{n-1}\left(\alpha_1 + N_1, \dots, \alpha_n + N_n\right), \quad N_j = \mathrm{Card}\left(\{i \,:\, X_i = j\}\right). \tag{1.19}$$

Therefore the relevance of group j is reinforced according to the number of collected observations equal to j. Indeed, if $\alpha = \sum_{i=1}^n \alpha_i$, thanks to (1.18) the posterior mean is equal to $\mathbb{E}\left[W_j \mid X_1,\dots,X_N\right] = (\alpha_j + N_j)/(\alpha + N)$. Thus we can obtain the predictive distribution of a new observation as

$$\begin{aligned}\mathbb{P}\left(X_{N+1}=j \mid X_1,\dots,X_N\right) &= \mathbb{E}\Big[\mathbb{P}\left(X_{N+1}=j \mid W_1,\dots,W_n\right) \mid X_1,\dots,X_N\Big] \\ &= \mathbb{E}\left[W_j \mid X_1,\dots,X_N\right] = \frac{\alpha_j+N_j}{\alpha+N} \\ &= \frac{\alpha}{\alpha+N}\alpha_j + \frac{N}{\alpha+N}\frac{N_j}{N}.\end{aligned} \tag{1.20}$$

Interestingly, the predictive distribution is a convex linear combination of the prior guess and the empirical frequency: moreover the weight assigned to the prior guess vanishes, as $N \to \infty$. In the next Section we show how to use the Dirichlet distribution to define a prior over an infinite dimensional space, through Theorem 6.

1.5.3 Dirichlet process: definition

Let α be a non null and finite measure on $(\mathbb{X}, \mathcal{X})$. Call $\theta = \alpha(\mathbb{X})$ the *concentration parameter* and $Q_0 = \alpha/\theta \in P(\mathbb{X})$ the *baseline* distribution. Setting $\alpha_i = \alpha(A_i)$, with $i = 1,\dots,k$, we write

$$Q_{A_1,\dots,A_k}(C) = \frac{\Gamma(\theta)}{\prod_{j=1}^k \Gamma(\alpha_j)} \int_{C\cap\Delta_{k-1}} w_1^{\alpha_1-1}\cdots w_{k-1}^{\alpha_{k-1}-1}(1-w_1-\dots-w_{n-1})^{\alpha_k-1}\,\mathrm{d}\underline{w},$$

for every $(A_1,\dots,A_k)$ partition of $\mathbb{X}$ and for every $C \in \mathcal{B}\left([0,1]^k\right)$. Thus, we assign a Dirichlet distribution to every partition of the space, with weights given by $\alpha(A_j)$. Consider now a generic ordered collection of sets $(A_1,\dots,A_k)$ and denote with $(C_1,\dots,C_{k'})$ the induced partition, i.e. such that

$$A_j = \cup_{(j)} C_i, \quad (j) = \left\{i \in \{1,\dots,k'\} \,:\, C_i \subset A_j\right\}.$$

Thus we define

$$Q_{A_1,\dots,A_k}(C) = Q_{C_1,\dots,C_{k'}}\left(\left\{(w_1,\dots,w_{k'-1}) \in \Delta_{k'-1} \,:\, \left(\sum_{(1)} w_i,\dots,\sum_{(k)} w_i\right) \in C\right\}\right). \tag{1.21}$$

We have now defined $Q = \{Q_{A_1,\dots,A_k} : A_1,\dots,A_k \in \mathcal{X}, k \geq 1\}$ and the next theorem shows it defines a proper random probability measure.

Theorem 7. *Let Q defined as in* (1.21)*. Then requirements* $(P1)-(P4)$ *are satisfied, so there exists a random probability measure P, which has finite dimensional distributions as in Q.*

Theorem 7 has been originally proven in (Ferguson, 1973). For a more detailed proof, see Chapter 3 of Regazzini (1996). In the following we will use the notation $P \sim DP(\theta, Q_0)$ to say that P is a random probability measure endowed with the law of a Dirichlet process (DP), with concentration parameter θ and baseline distribution Q_0.

This strategy relies mostly on Kolmogorov's Extension Theorem and the nice properties of the Dirichlet distribution (especially (1.17)). It is therefore difficult to use the same reasoning beyond this case, with the notable exception of the Normalized Inverse Gaussian process (Lijoi et al., 2005). In the next Chapters we will see different construction of the Dirichlet process, which will allow for various generalizations. In the next Section, instead, we study the statistical model associated to the DP and the associated posterior distribution.

1.5.4 Dirichlet process: posterior distribution

Consider the following Bayesian model for exchangeable data

$$X_i \mid P \overset{\text{iid}}{\sim} P, \quad P \sim DP(\theta, Q_0). \tag{1.22}$$

Thus, now the random parameter is the entire distribution of the observations, whose prior is given by the law of a Dirichlet process. Notice that

$$\mathbb{P}(X \in A) = \mathbb{E}\left[\mathbb{P}(X \in A \mid P)\right] = \mathbb{E}[P(A)] = Q_0(A),$$

since $P(A) \sim \text{Beta}\left(\theta Q_0(A), \theta Q_0(A^c)\right)$, by definition of the Dirichlet process. Thus the baseline distribution is the marginal of X according to model (2.1) and plays the role of the prior guess for the law of the observations. It is clear that the resulting statistical model is not dominated, so that formula (1.15) can not be applied. The idea is to rely again on the finite dimensional distributions, following the same lines of Chapter 3 in Regazzini (1996).

Let $(A_1, \dots, A_k)$ be an ordered partition of $\mathbb{X}$. Then the posterior distribution $\mathbb{P}\left(P(A_1) \in B_1, \dots, P(A_{k-1}) \in B_{k-1} \mid X_1\right)$ needs to satisfy the following integral equation

$$\begin{aligned}\mathbb{P}\Big(P(A_1) \in B_1, \dots,& P(A_{k-1}) \in B_{k-1}, X_1 \in C\Big)\\ &= \int_C \mathbb{P}\left(P(A_1) \in B_1, \dots, P(A_{k-1}) \in B_{k-1} \mid x_1\right) Q_0(\mathrm{d}x_1),\end{aligned}$$

for every $C \in \mathcal{X}$. Denote $C_i = A_i \cap C$ and $C_i' = A_i \cap C^c$, so that by definition

$$\Big(P(C_1), P(C_1'), \dots, P(C_k)\Big) \sim D_{2k-1}\left(\alpha_1, \alpha_1', \dots, \alpha_k, \alpha_k'\right),$$

where $\alpha_i = \alpha(C_i)$ and $\alpha_i' = \alpha(C_i')$. Therefore we obtain

$$\begin{aligned}&\mathbb{P}\Big(P(A_1) \in B_1, \dots, P(A_{k-1}) \in B_{k-1}, X_1 \in C\Big)\\ &= \int_{\Delta_{2k-1}} \mathbb{1}_{B_1 \times \dots \times B_{k-1}}(\underline{x} + \underline{y}) \left(\sum_{i=1}^k x_i\right) \frac{\Gamma(\theta)}{\prod_{j=1}^k \Gamma(\alpha_j)\Gamma(\alpha_j')} \prod_{j=1}^k x_j^{\alpha_j - 1} y_j^{\alpha_j' - 1} \,\mathrm{d}\underline{x}\mathrm{d}\underline{y}\\ &= \sum_{i=1}^k \int_{\Delta_{2k-1}} \mathbb{1}_{B_1 \times \dots \times B_{k-1}}(\underline{x} + \underline{y}) \frac{\Gamma(\theta)}{\prod_{j=1}^k \Gamma(\alpha_j)\Gamma(\alpha_j')} x_i^{\alpha_i} y_i^{\alpha_i' - 1} \prod_{j \neq i} x_j^{\alpha_j - 1} y_j^{\alpha_j' - 1} \,\mathrm{d}\underline{x}\mathrm{d}\underline{y}\\ &= \sum_{i=1}^k \frac{\alpha_i}{\theta} \int_{\Delta_{2k-1}} \mathbb{1}_{B_1 \times \dots \times B_{k-1}}(\underline{x} + \underline{y}) d_{2k-1}(\underline{x}, \underline{y}\,;\, \alpha_1, \alpha_1', \dots, \alpha_i + 1, \dots, \alpha_k') \,\mathrm{d}\underline{x}\mathrm{d}\underline{y},\end{aligned}$$

with $y_k = 1 - \sum_{j=1}^k x_j - \sum_{j=1}^{k-1} y_j$. Noticing that $Q_0(C_i) = \alpha_i/\theta$ and denoting $\tilde{\alpha}_j = \alpha(A_j)$, by applying (1.17) we have

$$\begin{aligned}
&\mathbb{P}\Big(P(A_1) \in B_1, \ldots, P(A_{k-1}) \in B_{k-1}, X_1 \in C\Big) \\
&\quad = \sum_{i=1}^{k} Q_0(C_i) \int_{\Delta_{k-1} \cap (B_1, \ldots, B_{k-1}} d_{k-1}(\underline{w}\,;\, \tilde{\alpha}_1, \ldots, \tilde{\alpha}_i + 1, \ldots, \tilde{\alpha}_k)\, \mathrm{d}\underline{w} \\
&\quad = \int_C \int_{\Delta_{k-1} \cap (B_1, \ldots, B_{k-1}} d_{k-1}\left(\underline{w}\,;\, \tilde{\alpha}_1 + \mathbb{1}_{A_1}(x), \ldots, \tilde{\alpha}_k + \mathbb{1}_{A_k}(x)\right) \mathrm{d}\underline{w} Q_0(\mathrm{d}x).
\end{aligned}$$

Therefore, by definition of conditional probability, we have

$$(P(A_1), \ldots, P(A_{k-1})) \mid X_1 \sim D_{k-1}\left(\alpha(A_1) + \mathbb{1}_{A_1}(x), \ldots, \alpha(A_k) + \mathbb{1}_{A_k}(x)\right.$$

for every ordered partition. Considering instead an arbitrary ordered collection of sets $(A_1, \ldots, A_k)$, with the same reasoning it is possible to prove that the finite dimensional distributions of P, conditional to X, are as in (1.21) with a new measure $\alpha' = \alpha + \delta_X$. Therefore we proved the following theorem for the posterior distribution of model (2.1).

Theorem 8. *Consider a random variable X generated according to model* (2.1)*. Then it holds*

$$P \mid X \sim DP\left(\theta + 1, \frac{\theta}{\theta+1} Q_0 + \frac{1}{\theta+1} \delta_X\right)$$

In order to compute the posterior distribution $P \mid X_{1:n}$, with $X_{1:n} = (X_1, \ldots, X_n)$, Theorem 8 can be applied sequentially to get

$$P \mid X_{1:n} \sim DP\left(\theta + n, \frac{\theta}{\theta+n} Q_0 + \frac{n}{\theta+n} \hat{P}_n\right), \tag{1.23}$$

where $\hat{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$. In the next section we explore some basic properties of model (2.1).

1.5.5 Dirichlet process: basic properties

Thanks to the availability of the posterior distribution given in Theorem 8, we can extract useful quantity for model (2.1). In particular we have that

$$P(A) \mid X_{1:n} \sim \text{Beta}\left(\theta Q_0(A) + n\hat{P}_n(A), \theta Q_0(A^c + n\hat{P}_n(A^c)\right),$$

so that the predictive distribution reads

$$\begin{aligned}
\mathbb{P}\left(X_{n+1} \in A \mid X_{1:n}\right) &= \mathbb{E}\Big[\mathbb{P}\left(X_{n+1} \in A \mid P\right) \mid X_{1:n}\Big] = \mathbb{E}\left[P(A) \mid X_{1:n}\right] \\
&= \frac{Q_0(A) + n\hat{P}_n(A)}{\theta + n} = \frac{\theta}{\theta+n} Q_0(A) + \frac{n}{\theta+n} \hat{P}_n(A).
\end{aligned} \tag{1.24}$$

Therefore the predictive is a convex linear combination of the prior guess and the empirical distribution of the observed datapoints. Interestingly, the fact that the prediction rule is a linear combination of Q_0 and the empirical measure is a characterization of the Dirichlet process (Regazzini, 1978; Lo, 1991). Moreover, the higher θ the higher the weight associated to the baseline distribution: therefore θ measures the confidence on the prior guess. From (1.24) we can devise a simple scheme to sample n observations from model (2.1):

1. Sample $X_1 \sim Q_0$.
2. For every $i \geq 1$ sample
$$X_{i+1} \sim \begin{cases} Q_0 & \text{w.p. } \frac{\theta}{\theta+i} \\ \hat{P}_i & \text{w.p. } \frac{i}{\theta+i} \end{cases}$$

This is often called a Pólya urn scheme, since it behaves as sampling with reinforcement from an urn with infinitely many colors. It is then clear that a sample $X_{1:n}$ from model (2.1) yields ties with positive probability and moreover

$$\mathbb{P}(X_{i+1} = \text{new} \mid X_{1:i}) = \frac{\theta}{\theta + i},$$

assuming that Q_0 is diffuse. Notice that the probability of a new value depends only on θ and n: in the following Chapters we will see suitable generalizations, to obtain dependence also on the number of distinct values observed in $X_{1:i}$ (De Blasi et al., 2013). Calling $W_i \in \{0,1\}$ the variable equal to 1 if X_i is new, we can denote with $K_n = \sum_{i=1}^n W_i$ the number of distinct values out of a sample of n elements. By the above formula, we have that W_i are independent Bernoulli random variables with parameter $\theta/(\theta + i - 1)$. Therefore

$$\mathbb{E}\left[K_n\right] = \sum_{i=1}^n \mathbb{E}\left[W_i\right] = \theta \sum_{i=1}^n \frac{1}{\theta + i - 1},$$

which behaves approximately as $\log(1 + n/\theta)$, with n large. It is possible to say more, that is $K_n/\log(n) \to \theta$ almost surely, as $n \to \infty$ (see Korwar and Hollander (1973) for a proof). Thus, the number of clusters (i.e. distinct values) grows logarithmically with n, regardless of the choice of θ and Q_0: in the next Chapters we will introduce other processes for which it is possible to tune the growth rate with suitable hyperparameters.

A sample $X_{1:n}$ from model (2.1) can be equivalently described by the $k \leq n$ unique values and the associated partition in k clusters. By exchangeability, partitions with the same multiplicities $(n_1, \ldots, n_k)$, with $n_i > 0$ and $\sum_{i=1}^k n_i = n$, yield the same probability. Therefore an object of great interest is given by the Exchangeable Partition Probability Function (EPPF) $\Pi_k^{(n)}(n_1, \ldots, n_k)$, i.e. the distribution over partitions of $\{1, \ldots, n\}$ with multiplicities $(n_1, \ldots, n_k)$ induced by model (2.1). By exchangeability and (1.24) we have

$$\begin{aligned} \Pi_k^{(n)}(n_1, \ldots, n_k) &= \mathbb{P}\left(X_1 = 1, \ldots, X_{n_1} = 1, X_{n_1+1} = 2, \ldots, X_n = k\right) \\ &= \frac{1}{\theta+1} \cdots \frac{n_1 - 1}{\theta + n_1 - 1} \frac{\theta}{\theta + n_1} \cdots \frac{n_2 - 1}{\theta + n_1 + n_2 - 1} \cdots \frac{n_k - 1}{\theta + n - 1} \\ &= \frac{\theta^k}{(\theta)_{(n)}} \prod_{j=1}^k \Gamma(n_j), \end{aligned} \tag{1.25}$$

where $\theta_{(n)} = \theta(\theta+1) \cdots (\theta + n - 1)$ is the Pochammer symbol.

Finally, again thanks by the conjugacy given by Theorem 8, it is possible to prove that realizations of the Dirichlet process are almost surely discrete probability measures. This is formalized in the next lemma.

Lemma 1. *Let Q be the law of a Dirichlet process with parameters θ and Q_0. Then*

$$Q\left(\textit{discrete probability measures over } \mathbb{X}\right) = 1.$$

Proof. For every $P \in P(\mathbb{X})$ denote with $E_P = \{x \in \mathbb{X} : P(\{x\}) > 0\}$ the set of atoms of P. Then the set $\Gamma_d \subset P(\mathbb{X})$ of discrete probability measures over $\mathbb{X}$ can be defined as $\Gamma_d = \{P \in P(\mathbb{X}) : P(E_P) = 1\}$ Thus, the statement is equivalent to $Q(\Gamma_d) = 1$.

Denoting $E_x = \{P \in P(\mathbb{X}) : P(\{x\}) > 0\}$, we have that $Q(\Gamma_d) = 1$ if and only if $\mathbb{P}\left(Q_{X_1}\left(E_{X_1}\right) = 1\right) = 1$, where Q_x is the posterior law of the Dirichlet process given $X_1 = x$, with X_1 from model (2.1). By Theorem 8 we have that

$$P\left(\{X_1\} \mid X_1\right) \sim \text{Beta}\left(\theta Q_0(\{X_1\}) + 1, \theta - \theta Q_0(\{X_1\})\right),$$

where P is a random probability measure with law Q. Therefore

$$Q_{X_1}\left(E_{X_1}\right) = Q_{X_1}\left(\{P \in P(\mathbb{X})\,;\, P(\{X_1\}) > 0\}\right) = 0,$$

as desired. □

1.6 Beyond exchangeability

As we saw in the previous Sections, the most common assumption underlying Bayesian models is exchangeability, which corresponds to invariance of the joint distribution of the observations with respect to finite permutations. However, most often the data present features that make exchangeability unrealistic, e.g. presence of covariates, temporal dependence, different experimental conditions. Therefore de Finetti's Theorem has been generalized over the decades to different probabilitic models: we mention for example exchangeability for Markov chains (Diaconis and Freedman, 1980a), arrays (Aldous, 1981) and networks (Caron and Fox, 2017). See Aldous (1985) for a detailed representation. Moreover, de Finetti's Theorem has been proven to be robust under small deviations from exchangeability, see e.g. Campbell et al. (2023).

In this thesis we focus on the setting where collected data may refer to different features, populations, or, in general, may be collected under different experimental conditions. Such situations entail a significant level of heterogeneity and opportunities for borrowing information, that can be exploited through the notion of *partial exchangeability* (de Finetti, 1938), which implies exchangeability within each experimental condition, but not across. Two sequences of observations $X = (X_i)_{i\geq 1}$ and $Y = (Y_j)_{j\geq 1}$, taking values in a space $\mathbb{X}$, are partially exchangeable if and only if, for all sample sizes (n, m) and all permutations (π_1, π_2),

$$\left((X_i)_{i=1}^n, (Y_j)_{j=1}^m\right) \overset{d}{=} \left((X_{\pi_1(i)})_{i=1}^n, (Y_{\pi_2(j)})_{j=1}^m\right).$$

From an inferential point of view, partial exchangeability entails that the order of the observations within each sample is non-informative, while the fact of belonging to a specific sample is relevant and has to be taken into account. Moreover, there exists a generalized version of de Finetti's Theorem (de Finetti, 1938) which states that X and Y are partially exchangeable if and only if there exist random probability measures (P_1, P_2) such that for every $i, j = 1, \dots, n$

$$(X_i, Y_j) \mid P_1, P_2 \overset{\text{iid}}{\sim} P_1 \times P_2, \quad (P_1,\, P_2) \sim Q, \tag{1.26}$$

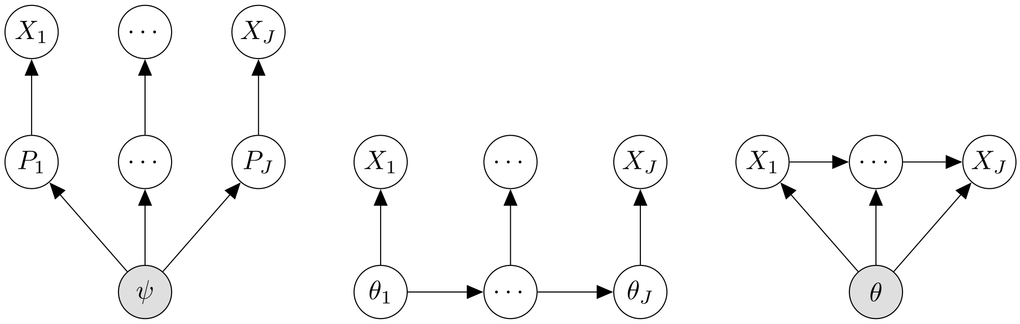


Figure 1.1: Graphical models of different hierarchical structures. Left: inducing dependence across groups. Center: defining a more flexible likelihood. Right: defining a more flexible prior.

with $Q \in P\left(P^2(\mathbb{X})\right)$ playing the role of the prior. The analogy with model (1.1) is clear: conditional on the parameters (P_1, P_2), observations are independent, with law given by the corresponding group. The joint distribution of the pair (P_1, P_2) models the dependence, i.e. the borrowing of information, across groups: in particular, if $P_1 = P_2$ almost surely exchangeability is recovered. Representation (26) can be easily extended to d groups, through a vector of d random probability measures $(P_1, \dots, P_d)$ with prior distribution $Q \in P\left(P^d(\mathbb{X})\right)$. The literature has thus developed a plethora of models specifying such law Q: in the next Chapters we will focus on Bayesian nonparametric models for partially exchangeable data, starting from the the early works of Cifarelli and Regazzini (1978); MacEachern (1999, 2000). We terminate the Chapter, instead, on the role played by hierarchical structures in exchangeable and partially exchangeable models.

1.7 Hierarchical structures for Bayesian modelling

Hierarchies play a key role in Bayesian modelling, since they provide a simple and effective way to define the joint distribution of random quantities. Thanks to the well-known Chain Rule, we say that the pair (X, Y) is defined *hierarchically* if the marginal distribution of X, namely $\mathbb{P}(X \in A)$, and the conditional distribution $\mathbb{P}(Y \in A \mid X)$ are specified. This is usually represented through Directed Acyclic Graphs (DAG), as in Figure 1.1, with "$X \to Y$". Therefore we can use such graphs to easily describe the probabilistic structure among the objects of interest.

In particular, we will use hierarchies for three distinct, yet related, task. We describe them through the Dirichlet process model defined in (2.1).

1. **Inducing dependence across groups**: in the setting of partial exchangeability, described in the previous Section, hierarchies can be used to induce dependence across distinct groups that share some common features. As shown in the left part of Figure 1.1, group j is modelled through the prior P_j, depending on a common hyperparameter ψ: the latter, being endowed with a suitable prior distribution, induces dependence across P_j and therefore across datapoints X_j. An example is given by the Hierarchical Dirichlet process (Teh et al., 2006) defined as

$$P_j \mid \psi \sim DP(\theta_j, \psi), \quad \psi \sim DP(\theta, Q_0).$$

Therefore ψ plays the role of the common baseline distribution for all the groups. The resulting simple probabilistic structure, as in the left of Figure 1.1, allows to greatly simplify the theoretical and computational analysis.

2. **Defining a more flexible likelihood**: the discreteness of the Dirichlet process, as shown in Lemma 1, may be a weakness of model (2.1). Therefore it is customary to convolve this discrete structure with a suitable kernel $k(x\,\theta)$, depending on a parameter θ and dominated by a σ-finite measure μ, so that the resulting likelihood is also dominated by μ: simple examples for k are the normal and Poisson kernels, depending on the nature of the data. The resulting model, introduced in Lo (1984) and often termed *Dirichlet process mixtures*, can be defined as

   $$X_i \mid \theta_i \sim k(x \mid \theta_i), \quad \theta_i \mid P \overset{\text{iid}}{\sim} P, \quad P \sim DP(\theta, Q_0).$$

   Therefore the discreteness of the Dirichlet process implies a latent clustering structure, which can be used to automatically partition the observations in groups. The dependence structure is defined by the center of Figure 1.1.

3. **Defining a more flexible prior**: as seen in the last Section, the predictive and asymptotic properties of the Dirichlet process model (2.1) crucially depend on the concentration parameter θ. Thus, since its role is fundamental, it is common to place a suitable hyperprior to learn it from the data. The corresponding model, called *mixture of Dirichlet processes* and first defined in Antoniak (1974), is given by

   $$X_i \mid P \overset{\text{iid}}{\sim} P, \quad P \mid \theta \sim DP(\theta, Q_0), \quad \theta \sim P_0,$$

   where $P_0 \in P(\mathbb{R}_+)$. The graph is illustrated in the right part of Figure 1.1.

Notice the *modular* property illustrated in the above three tasks: already defined objects (as the Dirichlet process) can be used as *building blocks* for more complex models, which need to adapt to specific features of the phenomenon of interest. In the next three Chapters we will discuss modelling, theoretical and computational aspects of hierarchical models.

References

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, 11(4):581–598.

Aldous, D. J. (1985). *Exchangeability and related topics*. Springer.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

Berti, P., Dreassi, E., Leisen, F., Rigo, P., and Pratelli, L. (2023). Bayesian predictive inference without a prior. *Statistica Sinica*, forthcoming.

Campbell, T., Syed, S., Yang, C.-Y., Jordan, M. I., and Broderick, T. (2023). Local exchangeability. *Bernoulli*, 29(3):2084–2100.

Caron, F. and Fox, E. B. (2017). Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 79(5):1295.

Cifarelli, D. M. and Regazzini, E. (1978). Nonparametric statistical problems under partial exchangeability: The role of associative means. *Quaderni Istituto Matematica Finanziaria dell'Università di Torino Serie III*, 12:1–36.

Cifarelli, D. M. and Regazzini, E. (1996). De finetti's contribution to probability and statistics. *Statistical Science*, 11(4):253–282.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2013). Are gibbs-type priors the most natural generalization of the dirichlet process? *IEEE transactions on pattern analysis and machine intelligence*, 37(2):212–229.

de Finetti, B. (1929). Funzione caratteristica di un fenomeno aleatorio. In *Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928*, pages 179–190.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré*, volume 7, pages 1–68.

de Finetti, B. (1938). Sur la condition d'equivalence partielle. *Actualités Scientifiques et Industrielles*, 739:5–18, Translated In: Studies in Inductive and Probability, II. Jeffrey, R. (ed.) University of California Press: Berkeley 1980.

Diaconis, P. and Freedman, D. (1980a). de finetti's theorem for markov chains. *The Annals of Probability*, pages 115–130.

Diaconis, P. and Freedman, D. (1980b). Finite exchangeable sequences. *The Annals of Probability*, pages 745–764.

Diaconis, P. and Skyrms, B. (2018). *Ten great ideas about chance*. Princeton University Press.

Dubins, L. E. and Freedman, D. A. (1979). Exchangeable processes need not be mixtures of independent, identically distributed random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 48(2):115–132.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Fong, E., Holmes, C., and Walker, S. G. (2021). Martingale posterior distributions. *arXiv preprint arXiv:2103.15671*.

Fortini, S., Ladelli, L., and Regazzini, E. (2000). Exchangeability, predictive distributions and parametric models. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 86–109.

Fortini, S. and Petrone, S. (2023). Prediction-based uncertainty quantification for exchangeable sequences. *Philosophical Transactions of the Royal Society A*, 381(2247):20220142.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press.

Heath, D. and Sudderth, W. (1976). De finetti's theorem on exchangeable variables. *The American Statistician*, 30(4):188–189.

Hewitt, E. and Savage, L. J. (1955). Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, 80(2):470–501.

Holmes, C. C. and Walker, S. G. (2023). Statistical inference with exchangeability and martingales. *Philosophical Transactions of the Royal Society A*, 381(2247):20220143.

Kingman, J. F. (1978). Uses of exchangeability. *The Annals of Probability*, 6(2):183–197.

Korwar, R. M. and Hollander, M. (1973). Contribution to the theory of dirichlet processes. *The Annals of Probability*, 1:705–711.

Lijoi, A., Mena, R. H., and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-gaussian priors. *Journal of the American Statistical Association*, 100(472):1278–1291.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1):351–357.

Lo, A. Y. (1991). A characterization of the dirichlet process. *Statistics and Probability Letters*, 12:185–187.

MacEachern, S. N. (1999). Dependent nonparametric processes. pages Alexandria, VA: American Statistical Association.

MacEachern, S. N. (2000). Dependent dirichlet processes. *Technical Report,*, page The Ohio State University.

Regazzini, E. (1978). Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilitá. *Giornale dell'Istituto Italiano degli Attuari*, 41:77–89.

Regazzini, E. (1996). *Impostazione non parametrica di problemi d'inferenza statistica bayesiana.* Consiglio nazionale delle ricerche.

Robert, C. P. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*, volume 2. Springer.

Szabó, B. and van der Vaart, A. (2023). *Bayesian Statistics.* Lecture notes available at *https://fa.ewi.tudelft.nl/ vaart/books.html.*

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Chapter 2

Hierarchies based on the Dirichlet process

2.1 Introduction

The most well known Bayesian nonparametric model for exchangeable data is likely the one induced by the Dirichlet process (DP, Ferguson (1973)), i.e.

$$X_i \mid P \overset{\text{iid}}{\sim} P, \quad P \sim DP(\theta, Q_0), \tag{2.1}$$

where $\theta > 0$ is the concentration parameter and Q_0 is the baseline distribution. For brevity in the following we often write $P \sim DP(\alpha)$, where $\alpha = \theta Q_0$ is a finite measure. As discussed in the previous chapter, the DP has nice analytical properties which allow to perform posterior inference and prediction. In the next Sections we will study two distinct problems where hierarchies help to generalize and robustify model (2.1).

First, we show how to model time series data in a nonparametric way using the Fleming-Viot process: the latter is a suitable stochastic process, used to model the evolution of the law describing the phenomenon, whose invariant measure is exactly the Dirichlet process. The Section is based on the works of Ascolani et al. (2021, 2023b). Secondly we study Dirichlet process mixtures, discussed in the last chapter, which convolve the DP with a suitable kernel: the discreteness of the process induces a latent clustering of the datapoints, which is often of interest. It has been shown (Miller and Harrison, 2013, 2014) that such clustering is often inconsistent in terms of the number of cluster, while we prove that this issue may be resolved placing an hyperprior on θ in (2.1). This is based on Ascolani et al. (2023a).

2.2 Time series modelling with the Fleming-Viot process

2.2.1 Hidden Markov models

We assume to observe datapoints collected at p times $0 = t_0 < \cdots < t_{p-1} = T$, possibly in different amount at different times. In this setting exchangeability is clearly not appropriate, so we consider the general framework of Hidden Markov models (Cappé et al., 2009), i.e.

$$X^i_{t_n} \mid P_{t_n} \overset{\text{iid}}{\sim} P_{t_n}, \quad \{P_{t_n} : n = 0, \ldots, p-1\}. \tag{2.2}$$

Therefore observations collected at the same time are exchangeable, so that the data are assumed to be partially exchangeable (de Finetti, 1938) as defined in the last Chapter. In the following, we will denote for brevity $\mathbf{X}_i := \mathbf{X}_{t_i}$ and $\mathbf{X}_{0:T} := (\mathbf{X}_0, \dots, \mathbf{X}_T)$, where $\mathbf{X}_i$ is the set of n_i observations collected at time t_i. Similarly, we sometimes denote $P_i := P_{t_i}$.

From representation (2.2), specifying a BNP model for temporally dependent observations requires to define a family of random probability measures $\{P_{t_n} : n = 0, \dots, p-1\}$, indexed by time. Previous contributions in this framework include Canale and Ruggiero (2016); Caron et al. (2007, 2017); Caron and Teh (2012); Dunson (2006); Griffin and Steel (2011); Gutiérrez et al. (2016); Kon Kam King et al. (2020); Mena and Ruggiero (2016); Rodriguez and Ter Horst (2008). Many proposals start from the celebrated stick-breaking representation of the Dirichlet process Sethuraman (1994), whereby P in (2.1) is such that

$$P \overset{d}{=} \sum_{i\geq 0} V_i \prod_{j=1}^{i-1} (1 - V_j)\, \delta_{X_i}, \qquad V_i \overset{\text{iid}}{\sim} \text{Beta}(1, \theta), \qquad X_i \overset{\text{iid}}{\sim} Q_0, \tag{2.3}$$

and the temporal dependence is induced by letting each V_i and/or X_i depend on time in a way that preserves the marginal distributions. Those are all examples of Dependent Dirichlet processes (MacEachern, 1999, 2000), such that the marginal distribution of X_i is given by the law of a Dirichlet process. This approach has many advantages, among which: simplicity and versatility, since inducing dynamics on V_i or X_i allows for a variety of solutions; flexibility, since under mild conditions the resulting processes have large support (cf. Barrientos et al. (2012)); ease of implementation, since strategies for posterior computation based on MCMC sampling are readily available. However, the stick-breaking structure makes the analytical derivation of further posterior information, like for example characterizing the predictive distribution of the observations, often a daunting task. This typically holds for other approaches to temporal Bayesian nonparametric modelling as well. Determining explicitly such quantities would not only give a deeper insight into the model posterior properties, which otherwise remain obscure to a large extent, but also provide a further tool for direct application or as a building block in more involved dependent models, whose computational efficiency would benefit from an explicit computation. In the next Sections we consider a different approach, based on the Fleming-Viot process.

2.2.2 Fleming-Viot process

We consider a class of dependent Dirichlet processes with continuous temporal covariate. Instead of inducing the temporal dependence through the building blocks of the stick-breaking representation (2.3), we let the dynamics of the dependent process be driven by a Fleming–Viot (FV) diffusion. FV processes have been extensively studied in relation to population genetics (see Ethier and Kurtz (1993) for a review), while their role in Bayesian nonparametrics was first pointed out in Walker et al. (2007) (see also Favaro et al. (2009)). A loose but intuitive way of thinking a FV diffusion is of being composed by infinitely-many probability masses, associated to different locations in the sampling space $\mathbb{X}$, each behaving like a diffusion in the interval $[0, 1]$, under the overall constraint that the masses sum up to 1. In addition, locations whose masses touch 0 are removed, while new locations are inserted at a rate which depends on a parameter $\theta > 0$. As a consequence, the random measures P_t and P_s, with $t \neq s$, will share some, though not all, of their support points.

The transition function that characterizes a FV process admits the following natural interpretation in Bayesian nonparametrics (cf. Walker et al. (2007)). Initiate the process at the

Random Probability Measure (RPM) $P_0 \sim DP(\alpha)$, and denote by D_t a time-indexed latent variable taking values in $\mathbb{Z}_+$. Conditional on $D_t = m \in \mathbb{Z}_+$, the value of the process at time t is a posterior DP P_t with law

$$P_t \mid (D_t = m, Y_1, \dots, Y_m) \sim DP\left(\alpha + \sum_{i=1}^{m} \delta_{X_i}\right) \qquad X_i \mid P_0 \overset{\text{iid}}{\sim} P_0. \tag{2.4}$$

Here, the realisation of the latent variable D_t determines how many atoms m are drawn from the initial state P_0, to become atoms of the posterior Dirichlet from which the arrival state is drawn. Such D_t is a pure-death process, which starts at infinity with probability one and jumps from state m to state $m-1$ after an exponentially distributed waiting time with inhomogenous parameter $\lambda_m = m(\theta + m - 1)/2$. The transition probabilities of D_t have been computed by Griffiths (1980); Tavaré (1984), and in particular

$$\mathbb{P}(D_t = m \mid D_0 = \infty) = d_m(t) \tag{2.5}$$

where

$$d_m(t) = \sum_{k=m}^{\infty} e^{-\lambda_k t} (-1)^{k-m} \frac{(\theta + 2k - 1)(\theta + m)_{(k)}}{m!(k-m)!},$$

$\lambda_k = k(\theta + k - 1)/2$ and where $\theta_{(k)} = \theta(\theta - 1) \cdots (\theta - k + 1)$ is the Pochhammer symbol. Here the fact that $D_0 = \infty$ almost surely should be understood as an entrance boundary, i.e., the process decreases from infinity at infinite speed so that at each $t > 0$ the value of D_t is finite. The unconditional transition of the FV process is thus obtained by integrating $D_t, X_1, \dots, X_{D_t}$ out of (2.4), leading to

$$Q_t(x, \mathrm{d}x') = \sum_{m=0}^{\infty} d_m(t) \int_{\mathbb{X}^m} DP\left(\alpha + \sum_{i=1}^{m} \delta_{y_i}\right)(\mathrm{d}x') x(\mathrm{d}y_1) \cdots x(\mathrm{d}y_m). \tag{2.6}$$

This was first found by Ethier and Griffiths (1993). It is known that $DP(\alpha)$ is the invariant measure of Q_t, i.e. if $P_0 \sim DP(\alpha)$ all the marginal RPMs P_t are Dirichlet processes with the same parameter. In particular, the death process D_t determines the correlation between RPMs at different times. Indeed, a larger t implies a lower m with higher probability, hence a decreasing (on average) number of support points will be shared by the random measures P_0 and P_t when t increases. On the contrary, as $t \to 0$ we have $D_t \to \infty$, which in turn implies infinitely-many atoms shared by P_0 and P_t, until the two RPMs eventually coincide.

For definiteness, we formalise the following definition.

Definition 1. *A Markov process $\{P_t\}_{t \geq 0}$ taking values in the space of atomic measures on $\mathbb{X}$ is a* Fleming–Viot dependent Dirichlet process *with parameter α, denoted $P_t \sim$ FV-DDP(α), if $P_0 \sim DP(\alpha)$ and its transition function is* (2.6).

Seeing a FV-DDP as a collection of RPMs, one is immediately led to wonder about the support properties of the induced prior. The weak support of a DDP indexed by an $\mathbb{R}_+$-valued covariate is the smallest closed set in $\mathcal{B}\{P(\mathbb{X})^{\mathbb{R}_+}\}$ with probability one, where $P(\mathbb{X})$ is the set of probability measures on $\mathbb{X}$ and $\mathcal{B}\{P(\mathbb{X})^{\mathbb{R}_+}\}$ is the Borel σ-field generated by the product topology of weak convergence. Barrientos et al. (2012) investigated these aspects for a large class of DDPs based on the stick-breaking representation of the Dirichlet process. Since no such representation is known for the FV process, our case falls outside that class. The following proposition states that a FV-DDP has full weak support, relative to the support of Q_0.

Proposition 1. *Let $\alpha = \theta Q_0$ and $\mathbb{X}$ be the support of Q_0. Then the weak support of a* FV-DDP(α) *is given by $P(\mathbb{X})^{\mathbb{R}_+}$.*

In order to formalise the statistical setup, we cast the FV-DDP into a hidden Markov model framework. A hidden Markov model is a double sequence $\{(P_{t_n}, X_{t_n}), n \geq 0\}$ where P_{t_n} is an unobserved Markov chain, called hidden or *latent signal*, and X_{t_n} are conditionally independent observations given the signal. The signal can be thought of as the discrete-time sampling of a continuous time process, and is assumed to completely specify the distributions of the observations, called *emission distributions*. While the literature on hidden Markov models has mainly focused on finite-dimensional signals, infinite-dimensional cases have been previously considered in Beal et al. (2001); Van Gael et al. (2008); Stepleton et al. (2009); Yau et al. (2011); Zhang et al. (2014); Papaspiliopoulos et al. (2016).

Here we take P_{t_n} to be a FV-DDP as in Definition 1, evaluated at p times $0 = t_0 < \cdots < t_{p-1} = T$. The sampling model is thus

$$X_{t_n}^i \mid P_{t_n} \overset{\text{iid}}{\sim} P_{t_n}, \qquad P_t \sim \text{FV-DDP}(\alpha). \tag{2.7}$$

It follows that any two variables $X_{t_n}^i, X_{t_m}^j$ are conditionally independent given P_{t_n} and P_{t_m}, with product distribution $P_{t_n} \times P_{t_m}$.

In addition, similarly to mixing a DP with respect to its parameter measure as in Antoniak (1974), one could also consider randomizing the parameter α in (2.7), e.g. by letting $\alpha = \alpha_\gamma$ and $\gamma \sim \pi$ on an appropriate space.

We will sometimes refer to $\mathbf{X}_{0:T}$ as the *past values*, since the inferential interest will be set at time $T + t$. We will also denote by $(x_1^*, \ldots, x_K^*)$ the K distinct values in $\mathbf{X}_{0:T}$, where $K \leq \sum_{i=0}^T n_i$. In this framework, Papaspiliopoulos et al. (2016) showed that the conditional distribution of the RPM P_T, given $\mathbf{X}_{0:T}$, can be written as

$$\mathcal{L}(P_T|\mathbf{X}_{0:T}) = \sum_{\mathbf{m}\in\mathbf{M}} w_{\mathbf{m}} DP\left(\alpha + \sum_{j=1}^K m_j \delta_{x_j^*}\right). \tag{2.8}$$

The weights $w_{\mathbf{m}}$ can be computed recursively as detailed in Papaspiliopoulos et al. (2016). In particular, $\mathbf{M}$ is a finite convex set of vector multiplicities $\mathbf{m} = (m_1, \ldots, m_K) \in \mathbb{Z}_+^K$ determined by $\mathbf{X}_{0:T}$, which identify the mixture components in (2.8) with strictly positive weight. We will call $\mathbf{M}$ the set of *currently active indices*. In particular, $\mathbf{M}$ is given by the points that lie between the counts of $(x_1^*, \ldots, x_K^*)$ in $\mathbf{X}_T$, which is the bottom node, and the counts of $(x_1^*, \ldots, x_K^*)$ in $\mathbf{X}_{0:T}$, which is the top node. For example, if $T = 1$ suppose we observe $\mathbf{X}_0 = (x_1^*, x_2^*)$ for some values $x_1^* \neq x_2^*$ and $\mathbf{X}_1 = \mathbf{X}_0$, hence $K = 2$. Then the top node is $(2, 2)$ since in $\mathbf{X}_{0:1}$ there are 2 of each of (x_1^*, x_2^*) and the bottom node is $(1, 1)$ which is the counts of (x_1^*, x_2^*) in $\mathbf{X}_1$. Cf. Figure 2.1. Note that observations with $K = 3$ distinct values would generate a 3-dimensional graph, with the origin $(0, 0, 0)$ linked to 3 level-1 nodes $(1, 0, 0), (0, 1, 0), (0, 0, 1)$, and so on. In general, each upper level node is obtained by adding 1 to one of the lower node coordinates.

We note here that the presence of $d_m(t)$ in (2.6) makes the computations with FV processes in principle intractable, yielding in general infinite mixtures difficult to simulate from (cf. Jenkins and Spano (2017)). It is then remarkable that conditioning on past data one is able to obtain conditional distribution for the signal given by finite mixtures as in (2.8).

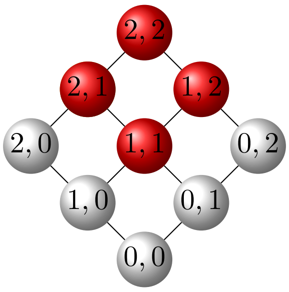


Figure 2.1: Red indices in the graph identify active mixture components at time T, i.e. the set $\mathbf{M}$ in (2.8), corresponding to points $\mathbf{m} \in \mathbb{Z}_+^K$ with positive weight. In this example $K = 2$, and the graph refers to $\mathbf{M}$ at time $T = 1$ if we observe $\mathbf{X}_0 = (x_1^*, x_2^*) = \mathbf{X}_1$.

2.2.3 Predictive inference

Predictive distribution

In the above framework, we are now interested in predictive inference, which requires obtaining the predictive distribution of $X_{T+t}^1, \dots, X_{T+t}^k | \mathbf{X}_{0:T}$, that is the marginal distribution of a k-sized sample drawn at time $T+t$, given data collected up to time T, when the random measures involved are integrated out. See Figure 2.2. Note that by virtue of the stationarity of the FV process, if $P_0 \sim DP(\alpha)$, then $\mathbb{P}(X_t \in A) = Q_0(A)$ for any $t \geq 0$. Note also that if one mixes model (2.7) by randomizing the parameter measure $\alpha = \alpha_\gamma$ as mentioned above, the evaluation the predictive distributions is of paramount importance for posterior computation. Indeed, one needs the distribution of $\gamma | \mathbf{X}_{0:T}$, and if for example γ has discrete support on $\mathbb{Z}_+$ with probabilities $\{p_j, j \in \mathbb{Z}_+\}$, then

$$\mathbb{P}(\gamma = j | \mathbf{X}_{0:T}) \propto p_j \mathbb{P}(\mathbf{X}_{0:T} | j) \propto p_j \mathbb{P}(\mathbf{X}_0 | j) \mathbb{P}(\mathbf{X}_1 | \mathbf{X}_0, j) \cdots \mathbb{P}(\mathbf{X}_T | \mathbf{X}_{0:T-1}, j).$$

Denote for brevity $X_{T+t}^{1:k} := (X_{T+t}^1, \dots, X_{T+t}^k)$ the k values drawn at time $T+t$. For $\mathbf{m} \in \mathbb{Z}_+^K$, let $\{\mathbf{n} \in \mathbb{Z}_+^K : \mathbf{n} \leq \mathbf{m}\}$ be the set of nonnegative vectors such that $n_i \leq m_i$ for all i. Define also $|\mathbf{n}| := \sum_{j=1}^K n_i$, and

$$L(\mathbf{M}) := \{\mathbf{n} \in \mathbb{Z}_+^K : \mathbf{n} \leq \mathbf{m}, \mathbf{m} \in \mathbf{M}\} \tag{2.9}$$

to be all the points in $\mathbb{Z}_+^K$ lying below the top node of $\mathbf{M}$. E.g., if $\mathbf{M}$ is given by the red nodes in Figure 2.1, then $L(\mathbf{M})$ is given by all nodes shown in the figure.

Proposition 2. *Assume (2.7), and let the law of P_T given data $\boldsymbol{X}_{0:T}$ be as in (2.8), where the weights $w_\mathbf{m}$ have been computed recursively. Then, for any Borel set A of $\mathbb{X}$, the first observation at time $T+t$ has distribution*

$$\mathbb{P}\left(X_{T+t} \in A | \boldsymbol{X}_{0:T}\right) = \sum_{\mathbf{n} \in L(\mathbf{M})} p_t(\mathbf{M}, \mathbf{n}) \left(\frac{\theta}{\theta + |\mathbf{n}|} Q_0(A) + \frac{|\mathbf{n}|}{\theta + |\mathbf{n}|} P_\mathbf{n}(A)\right) \tag{2.10}$$

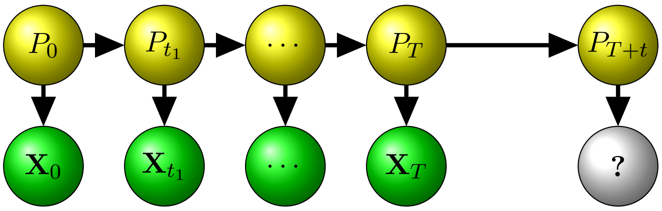


Figure 2.2: The predictive problem depicted as a graphical model. The upper yellow nodes are nonobserved states of the infinite-dimensional signal, the lower green nodes are conditionally independent observed data whose distribution is determined by the signal, the light gray node is the object of interest.

and the $(k+1)$st observation at time $T+t$, given the first k, has distribution

$$\begin{aligned}\mathbb{P}\left(X_{T+t}^{k+1} \in A | \boldsymbol{X}_{0:T}, X_{T+t}^{1:k}\right) = & \sum_{\mathbf{n}\in L(\mathbf{M})} p_t^{(k)}(\mathbf{M},\mathbf{n}) \\ & \times \left(\frac{\theta}{\theta+|\mathbf{n}|+k} Q_0(A) + \frac{|\mathbf{n}|}{\theta+|\mathbf{n}|+k} P_{\mathbf{n}}(A) + \frac{k}{\theta+|\mathbf{n}|+k} P_k(A)\right)\end{aligned} \tag{2.11}$$

where

$$P_{\mathbf{n}} = \frac{1}{|\mathbf{n}|}\sum_{i=1}^{K} n_i \delta_{x_i^*}, \qquad P_k = \frac{1}{k}\sum_{j=1}^{k} \delta_{X_{T+t}^j} \tag{2.12}$$

and $(x_1^*, \dots, x_K^*)$ are the distinct values in $\boldsymbol{X}_{0:T}$.

Before discussing the details of the above statement, a heuristic read of (2.10) is that the first observation at time $T+t$ is either a draw from the baseline distribution Q_0, or a draw from a random subset of the past data points $\mathbf{X}_{0:T}$, identified by the latent variable $\mathbf{n} \in L(\mathbf{M})$. Given how $L(\mathbf{M})$ is defined, X_{T+t} can therefore be thought of as being drawn from a mixture of Pólya urns, each conditional on a different subset of the data, ranging from the full dataset to the empty set. Indeed, recall that the top node of $\mathbf{M}$, hence of $L(\mathbf{M})$ in (2.9), is the vector of multiplicities of the distinct values $(x_1^*, \dots, y_K^*)$ contained in the entire dataset $\mathbf{X}_{0:T}$. The probability weights associated to each lower node $\mathbf{n} \in L(\mathbf{M})$ are determined by a death process on $L(\mathbf{M})$, that differs from D_t in (2.5). In particular this is a Markov process that jumps from node $\mathbf{m}$ to node $\mathbf{m} - \mathbf{e}_i$ after an Exponential amount of time with parameter $m_i(\theta + |\mathbf{m}| - 1)/2$, with $\mathbf{e}_i$ being the canonical vector in the ith direction. The weight associated with node $\mathbf{n} \in L(\mathbf{M})$ is then given by the probability that such death process is in $\mathbf{n}$ after time t, if started from any node in $\mathbf{M}$. For example, if $\mathbf{M}$ is as in Figure 2.1, than the weight of the node $(0,2)$ is given by the probability that the death process is in $(0,2)$ after time t if started from any other node of $\mathbf{M}$. Being a non increasing process, the admissible starting nodes are $(2,2)$ and $(1,2)$. Figure 2.3 highlights these two admissible paths of the death process which land at node $(0,2)$.

The transition probabilities of this death process are

$$p_{\mathbf{m},\mathbf{n}}(t) = p_{|\mathbf{m}|,|\mathbf{n}|}(t)\mathrm{HG}(\mathbf{m}-\mathbf{n}; \mathbf{m}, |\mathbf{m}-\mathbf{n}|), \qquad \mathbf{0} \le \mathbf{n} \le \mathbf{m}, \tag{2.13}$$

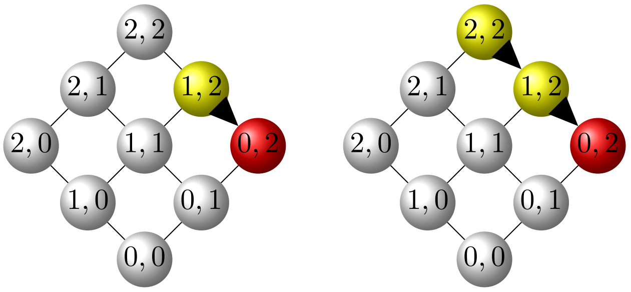


Figure 2.3: The weight associated to an index $\mathbf{n} \in L(\mathbf{M})$ at time $T+t$ is determined by the probability that the death process reaches $\mathbf{n}$ from any active index $\mathbf{m} \in \mathbf{M}$ at time T. For $\mathbf{M}$ as in Figure 2.1, the weight of the mixture component with index $\mathbf{n} = (0,2)$, i.e., no atoms x_1^* and 2 atoms x_2^*, is the sum of the probabilities of reaching node $(0,2)$ via the path starting from $(1,2)$ (left) and from $(2,2)$ (right).

where $\mathrm{HG}(\mathbf{i};\mathbf{m},|\mathbf{i}|)$ is the multivariate hypergeometric probability function evaluated at $\mathbf{i}$, namely

$$\mathrm{HG}(\mathbf{i};\mathbf{m},|\mathbf{i}|) = \frac{\binom{\mathbf{m}_1}{\mathbf{i}_1}\cdots\binom{\mathbf{m}_l}{\mathbf{i}_l}}{\binom{|\mathbf{m}|}{|\mathbf{i}|}}, \qquad l = \dim(\mathbf{m})$$

with $\dim(\mathbf{m})$ denoting the dimension of vector $\mathbf{m}$, while $p_{|\mathbf{m}|,|\mathbf{n}|}(t)$ is the probability of descending from level $|\mathbf{m}|$ to $|\mathbf{n}|$ (see Lemma 7 in the Supplementary Material). Hence, in general, the probability of reaching node $\mathbf{n} \in L(\mathbf{M})$ from any node in $\mathbf{M}$ is

$$p_t(\mathbf{M},\mathbf{n}) = \sum_{\mathbf{m}\in\mathbf{M},\mathbf{m}\geq\mathbf{n}} w_{\mathbf{m}} p_{\mathbf{m},\mathbf{n}}(t). \tag{2.14}$$

In conclusion, with probability $p_t(\mathbf{M},\mathbf{n})$ the first draw at time $T+t$ will be either from Q_0, with probability $\theta/(\theta+|\mathbf{n}|)$, or a uniform sample from the subset of data identified by the multiplicity vector $\mathbf{n}$.

Concerning the general case for the $(k+1)$st observation at time $T+t$, trivial manipulations of (2.11) provide different interpretative angles. Rearranging the term in brackets one obtains

$$\frac{\theta_{\mathbf{n}}}{\theta_{\mathbf{n}}+k}Q_{0,\mathbf{n}} + \frac{k}{\theta_{\mathbf{n}}+k}P_k, \tag{2.15}$$

which bears a clear structural resemblance to the predictive distribution of the DP. Here

$$\theta_{\mathbf{n}} = \theta + |\mathbf{n}|, \qquad P_{0,\mathbf{n}} := \frac{\theta}{\theta+|\mathbf{n}|}Q_0 + \frac{|\mathbf{n}|}{\theta+|\mathbf{n}|}P_{\mathbf{n}}$$

play the role of concentration parameter and baseline probability measure (i.e, the initial urn configuration), respectively. Thus (2.11) can be seen as a mixture of Pólya urns where the base measure has a randomised discrete component $P_{\mathbf{n}}$. Unlike the exchangeable case, observations not drawn from empirical measure P_k of the current sample can therefore be drawn either from Q_0 or from the empirical measure $P_{\mathbf{n}}$, where past observations are assigned multiplicities $\mathbf{n}$ with probability $p_t^{(k)}(\mathbf{M},\mathbf{n})$.

An alternative interpretation is obtained by developing the sum in (2.11) to obtain a single generalised Pólya urn, written in compact form as

$$\mathbb{P}\Big(X^{k+1}_{T+t} \in \cdot \mid \mathbf{X}_{0:T}, X^{1:k}_{T+t}\Big) = A_k Q_0(\cdot) + \sum_{i=1}^{K} C_{i,k}\delta_{y_i^*}(\cdot) + B_k P_k(\cdot) \tag{2.16}$$

where A is a Borel set of $\mathbb{X}$. In this case, the first observation is either from Q_0 or a copy of a past value $\mathbf{X}_{0:T}$, namely

$$X^1_{T+1} \sim \begin{cases} Q_0 & \text{w.p. } A_0 \\ \delta_{x_i^*} & \text{w.p. } C_{i,0}, \end{cases}$$

while the $(k+1)$st can also be a copy of one of the first k current observations $X^{1:k}_{T+t}$, namely

$$X^{k+1}_{T+1} \sim \begin{cases} Q_0 & \text{w.p. } A_k \\ \delta_{y_i^*} & \text{w.p. } C_{i,k} \\ P_k & \text{w.p. } B_k. \end{cases}$$

The pool of values to be copied is therefore given by past values $\mathbf{X}_{0:T}$ and current, already sampled observations $X^{1:k}_{T+t}$.

After each draw, the weights associated to each node need to be updated according to the likelihood that the observation was generated by the associated mixture component, similarly to what is done for mixtures of Dirichlet processes. Specifically,

$$p_t^{(k+1)}(\mathbf{M}, \mathbf{n}) \propto p_t^{(k)}(\mathbf{M}, \mathbf{n}) p(x^{k+1}_{T+t} \mid x^{1:k}_{T+t}, \mathbf{n}) \tag{2.17}$$

where

$$p(x^{k+1}_{T+t} \mid x^{1:k}_{T+t}, \mathbf{n}) := \frac{\theta q_0(x^{k+1}_{T+t}) + \sum_{i=1}^{K} n_i \delta_{x_i^*}(\{x^{k+1}_{T+t}\}) + \sum_{j=1}^{k} \delta_{x^j_{T+t}}(\{x^{k+1}_{T+t}\})}{\theta + |\mathbf{n}| + k} \tag{2.18}$$

is the predictive distribution of the $(k+1)$st observation given the previous k and conditional on $\mathbf{n}$, and q_0 is the density of Q_0 with respect to the Lebsegue or the counting measure.

As a byproduct of Proposition 2, we can evaluate the correlation between observations at different time points.

Proposition 3. *For $t, s > 0$, let X_t, X_{t+s} be from* (2.7)*. Then*

$$\mathrm{Corr}(X_t, X_{t+s}) = \frac{e^{-\frac{\theta}{2}s}}{\theta + 1}.$$

Unsurprisingly, the correlation decays to 0 as the lag s goes to infinity. Moreover,

$$\mathrm{Corr}(X_t, X_{t+s}) \to \frac{1}{\theta + 1}, \quad \text{as } s \to 0$$

which is the correlation of two observations from a DP as in (2.1).

Sampling from the predictive distribution

In order to make Proposition 2 useful in practice, we provide an explicit algorithm to sample from the predictive distribution (2.11), which can be useful *per se* or for approximating posterior quantities of interest. Exploiting (2.15) and the fact that (2.11) can be seen as a mixture of Pólya urns, we can see $\mathbf{n} \in \mathbb{Z}_+^K$ as a latent variable whereby, given $\mathbf{n}$, sampling proceeds very similarly to a usual Pólya urn.

Recalling that $|\mathbf{n}| = \sum_{j=1}^K n_i$, a simple algorithm for the $(k+1)$st observation would therefore be:

- sample $\mathbf{n} \in L(\mathbf{M})$ w.p. $p_t^{(k)}(\mathbf{M}, \mathbf{n})$;
- sample from $Q_0, P_\mathbf{n}$ or P_k with probabilities proportional to $\theta, |\mathbf{n}|, k$ respectively;
- update weights $p_t^{(k)}(\mathbf{M}, \mathbf{n})$ to $p_t^{(k+1)}(\mathbf{M}, \mathbf{n})$ for each $\mathbf{n} \in L(\mathbf{M})$.

A detailed pseudo-code description is provided in Algorithm 1.

Algorithm 1 Exact sampling from (2.11)

```
Input: - active nodes at time $T$: $\mathbf{M}$
       - precision parameter: $\theta$
       - last mixture weights $p_t^{(k)}(\mathbf{M}, \mathbf{n})$, $\mathbf{n} \in L(M)$
       - past unique observations: $x_1^*, \dots, x_K^*$
       - current observations: $x_{T+t}^1, \dots, x_{T+t}^k$

Sample n w.p. $p_t^{(k)}(\mathbf{M}, \mathbf{n})$, $\mathbf{n} \in L(\mathbf{M})$
Sample $X$ from $Q_0, P_\mathbf{n}$ or $P_k$ w.p. $\frac{\theta}{\theta+|\mathbf{n}|+k}, \frac{|\mathbf{n}|}{\theta+|\mathbf{n}|+k}, \frac{k}{\theta+|\mathbf{n}|+k}$ respectively
Set $x_{T+t}^{k+1} = X$
Update parameters:
for $\mathbf{n} \in L(\mathbf{M})$ and $p(x_{T+t}^{k+1} \mid x_{T+t}^{1:k})$ as in (2.18) do
    $p_t^{(k+1)}(\mathbf{M}, \mathbf{n}) = p_t^{(k)}(\mathbf{M}, \mathbf{n}) p(x_{T+t}^{k+1} \mid x_{T+t}^{1:k})$
    Normalize $p_t^{(k+1)}(\mathbf{M}, \mathbf{n})$
```

A possible downside of the above sampling strategy is that when the set $L(\mathbf{M})$ is large, updating all weights may be computationally demanding. Indeed, the size of the set $L(\mathbf{M})$ is $|L(\mathbf{M})| = \prod_{j=1}^K (1 + m_j)$, where m_j is the multiplicity of x_j^* in the data, which can grow considerably with the number of observations (cf. also Proposition 2.5 in Papaspiliopoulos and Ruggiero (2014)). It is however to be noted that, due to the properties of the death process that ultimately governs the time-dependent mixture weights, typically only a small portion of these will be significantly different from zero. Figure 2.4 illustrates this point by showing the nodes in $\{0, \dots, 50\}$ with weight larger than 0.05 at different times, if at time 0 there is a unit mass at the node 50, when $\theta = 1$. A deeper investigation of these aspects in a similar, but parametric, framework, can be found in Kon Kam King et al. (2021).

Hence an approximate version of the above algorithm can be particularly useful to exploit this aspect. We can therefore target a set $\tilde{\mathbf{M}} \subset L(\mathbf{M})$ such that $|\tilde{\mathbf{M}}|\mathbf{l}|L(\mathbf{M})|$ and $\sum_{\mathbf{n} \in \tilde{\mathbf{M}}} p_t(\mathbf{M}, \mathbf{n}) \approx 1$ by inserting a Monte Carlo step in the algorithm and simulate the death process with a large number of particles. The empirical frequencies of the particles landing nodes will then provide an estimate of the weights $p_t(\mathbf{M}, \mathbf{n})$ in (2.10). Furthermore, the simulation of the multidimensional death process can be factorised into simulating a one-dimensional death process, which simply tracks the number of steps down the graph, and hypergeometric sampling for choosing the landing node within the reached level. A simple algorithm for simulating the death process is as follows: for $i = 1, \dots, N$,

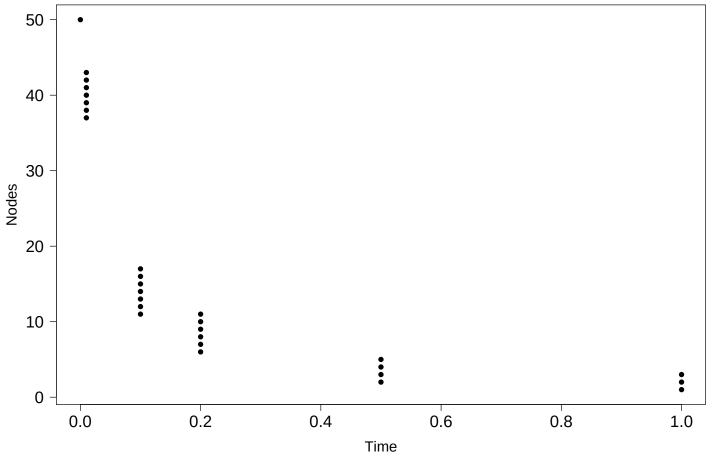


Figure 2.4: Nodes in $\{0,\dots,50\}$ (black dots) with probability of being reached by the death process bigger than .05 after lags .01, .1, .2, .5 and 1 (horizontal axis). Starting with mass 1 at the point 50, only a handful of nodes have significant mass after these lags.

- draw $\mathbf{m}$ with probability $w_{\mathbf{m}}$ and set $m = |\mathbf{m}|$;
- run a one-dimensional death process from m, and let n be the landing point after time t;
- draw $\mathbf{n}^{(i)} \sim \mathrm{HG}(n, \mathbf{m}/|\mathbf{m}|)$;

and return $\{\mathbf{n}^{(i)}, i = 1,\dots,N\}$. Note, in turn, that the simulation of the death process trajectories does not require to evaluate its transition probabilities (2.13), which are prone to numerical instability, and can instead be straightforwardly set up in terms of successive exponential draws by repeating the following cycle: for $i \geq 1$,

- draw $Z_i \sim \mathrm{Exp}(m(\theta + m - 1)/2)$
- if $\sum_{j\leq i} Z_j < t$ set $m = m - 1$ else return $n = m - i + 1$ and exit cycle.

Algorithm 2 outlines the pseudocode for sampling approximately from (2.11) according to this strategy.

Asymptotics

We investigate two asymptotic regimes for (2.11). The following Proposition shows that when $t \to \infty$, the FV-DDP predictive distribution converges to the usual Pólya urn.

Proposition 4. *Under the hypotheses of Proposition 2, we have*

$$\mathcal{L}\Big(X_{T+t}^{k+1} | \boldsymbol{X}_{0:T}, X_{T+t}^{1:k}\Big) \longrightarrow \frac{\theta}{\theta + k} Q_0 + \frac{k}{\theta + k} P_k, \qquad a.s., \text{ as } t \to \infty,$$

in total variation distance, with P_k as in (2.12).

Here the statement is almost sure with respect to the probability measure induced by the FV model on the space of measure-valued temporal trajectories. A heuristic interpretation of

Algorithm 2 Approximate sampling from (2.11)

```
Input: - active nodes at time $T$: $\mathbf{M}$
       - time to propagate: $t$
       - precision parameter: $\theta$
       - mixture weights at time $T$: $w_{\mathbf{m}}$
       - past unique observations: $x_1^*, \dots, x_K^*$
       - number of Monte Carlo iterates: $N$
$\tilde{\mathbf{M}} = \emptyset$; $w = \emptyset$
for $i \in 1 : N$ do
    Sample $\mathbf{m}$ w.p. $w_{\mathbf{m}}$, $\mathbf{m} \in \mathbf{M}$
    $n = |\mathbf{m}|$; $s = t$
    for $j \geq 1$ do
        Sample $Z$ from $\text{Exp}(n(\theta + n - 1)/2)$ and set $s = s - Z$
        if $s > 0$ and $n > 0$ then
            Set $n = n - 1$
        else
            Return $n$ and exit cycle.
    Sample $\mathbf{n} \sim \text{HG}(n, \mathbf{m}/|\mathbf{m}|)$
    if $\mathbf{n} \notin \tilde{M}$ then
        Add $\mathbf{n}$ to $\tilde{\mathbf{M}}$ and add 1 to $w$
    else
        Add 1 to the corresponding element of $w$
Normalize $w$.
Apply algorithm 1 with $\mathbf{M} = \tilde{\mathbf{M}}$ and $p_t(\mathbf{M}, \mathbf{n}) = w$
```

the above result is that, when the lag between the last and the current data collection point diverges, the information given by past observations $\mathbf{X}_{0:T}$ becomes obsolete, and sampling from (2.11) approximates sampling from the prior Pòlya urn. This should be intuitive, as very old information, relative to the current inferential goals, should have a negligible effect.

The following proposition shows that when $k \to \infty$ in (2.11), we recover the law of P_{T+t} given $\mathbf{X}_{0:T}$ as de Finetti measure.

Proposition 5. *Under the hypotheses of Proposition 2, we have*

$$\mathcal{L}\Big(X_{T+t}^{k+1} | \boldsymbol{X}_{0:T}, X_{T+t}^{1:k}\Big) \longrightarrow P^*, \qquad a.s., \text{ as } k \to \infty,$$

weakly, where $P^* \sim \mathcal{L}(P_{T+t} | \boldsymbol{X}_{0:T})$.

Here P^* is a random measure with the same distribution as the FV-DDP at time $T + t$ given only the past information $\mathbf{X}_{0:T}$. Recall for comparison that the same type of limit for the exchangeable case yields

$$\mathcal{L}(X_{k+1} | Y_1, \dots, X_k) \longrightarrow P^*, \qquad P^* \sim \Pi_\alpha, \quad \text{as } k \to \infty,$$

where $DP(\alpha)$ is the de Finetti measure of the sequence and P^* is sometimes called the directing random measure.

2.2.4 Illustration

We illustrate predictive inference using FV-DDPs, based on Proposition 2. Besides the usual prior specification regarding models based on the Dirichlet process, that concern the choice of the total mass θ and of the baseline distribution P_0, here we can also introduce a parameter $\sigma > 0$ that controls the speed of the DDP. This acts as a time rescaling, whereby the data collection times t_i are rescaled to σt_i. This additional parameter provides extra flexibility for estimation, as it can be used to adapt the prior to the correct time scale of the underlying data generating process.

Synthetic data

We consider data generated by the model

$$\begin{aligned} X_t &\sim \frac{1}{2}\text{Po}(\mu_t^{-1}, 0) + \frac{1}{2}\text{Po}(\nu_t^{-1}, 5), \\ \mu_t &= \mu_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{Exp}(1), \\ \nu_t &= \nu_{t-1} + \eta_t, \quad \eta_t \sim \text{Exp}(1), \eta_t \perp\!\!\!\perp \epsilon_t \end{aligned}$$

where $\text{Po}(\lambda, b)$ denotes a b-translated Poisson distribution with parameter λ (i.e. if $Y \sim \text{Po}(\lambda, b)$ then $Y - b \sim \text{Po}(\lambda)$), and where $\mu_0^{-1} = \nu_0^{-1} = 5$, for $t = 0, 1, 2, \ldots$. We collect 15 observations at each $t \in \{0, \ldots, 15\}$ and consider one-step-ahead predictions based on the first 5 and 15 data collection times.

We fit the data by using a FV-DDP model as specified in (2.7), with the following prior specification. We consider two choices for P_0, a Negative Binomial with parameters $(2, 0.5)$ and a Binomial with parameters $(99, 0.3)$, which respectively concentrate most of their mass around small values and around the value 30. We consider a uniform prior on θ concentrated on the points $\{.5, 1, 1.5, \ldots, 15\}$. A continuous prior could also be envisaged, at the cost of adding a Metropolis–Hastings step in the posterior simulation, which we avoid here for the sake of computational efficiency. Similarly, for σ we consider a uniform prior on the values $\{0.01, 0.1, 0.3, 0.5, 0.7, 0.9, 1.5\}$. The estimates are obtained by means of 500 replicates of (2.11) of 1000 observations each, using the approximate method outlined in Algorithm 2 with 10000 Monte Carlo iterates. We also compare the FV-DDP estimate with that obtained using the DDP proposed in Mena and Ruggiero (2016). This is constructed from the stick-breaking representation (2.3) by letting

$$V_i(t_n) \sim c\delta_{V'} + (1-c)\delta_{V_i(t_{n-1})}, \qquad V^{'} \sim \text{Beta}(1, \theta).$$

in (2.3) and keeping the locations fixed. We let the resulting DDP be the mixing measure in a time-dependent mixture of Poisson kernels, which provides additional flexibility to this model with respect to our proposal. Furthermore, we give the competitor model a considerable advantage by training it also with the data points collected at times 6 and 7, which provide information on the prediction targets, and by centering it on the Negative Binomial with parameters $(2, 0.5)$, rather than on the above mentioned mixture, which puts mass closer to where most mass of the true pmf lies.

Figure 2.5 shows the results on one-step-ahead prediction with 15 collection times: the pointwise credible intervals, computed with the empirical quantiles, are also plotted. The posterior of σ (not shown) concentrates most of the mass on points 0.7 and 0.9, which leads to learning the correct time scale for prediction, resulting in an accurate estimate of the true pmf. The credible intervals are quite wide, and a better precision may be achieved by increasing the number of time points at which the data are recorded.

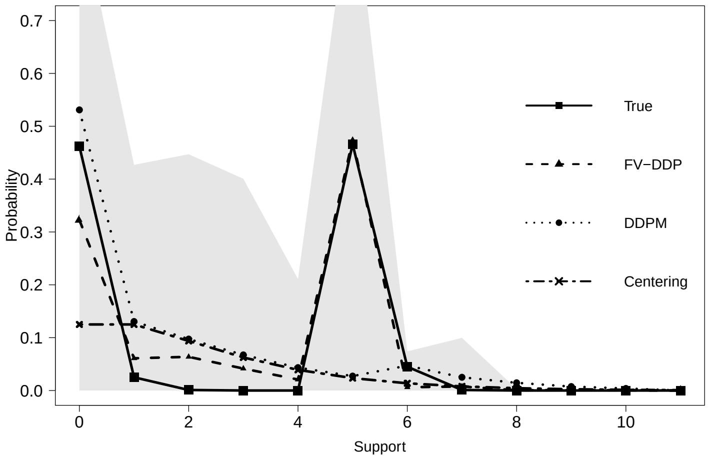


Figure 2.5: One-step-ahead prediction and 95% pointwise credible intervals, based on 15 data collection times.

We compare the previous results with those obtained by choosing σ via out-of-sample validation. This is done here using times 0 to 4 as training and time 5 as test, whereby for each $\sigma \in \{.0001, .001, .01, .1, 0.5, 1, 1.5\}$ we compute the sum of absolute errors (SAE) between the FV-DDP posterior predictive mean and the true pmf. These are shown in Table 3.1, leading to choose $\sigma = .01$.

σ	.0001	.001	.01	.1	.5	1	1.5
SAE	.1410	.1345	**.1064**	.1301	.1261	.1595	.1847

Table 2.1: Sum of the absolute error between predicted and true pmf at time 5 for different values of σ.

Table 3.2 shows the posterior weights of relevant values of θ among those with positive prior mass, for the above mentioned choices of Q_0 and using the chosen value of σ. The model correctly assigns all posterior probability to the Negative Binomial centering (Binomial not reported in the table), which moves mass towards smaller values as time increases.

θ	1	1.5	2	3
NegBinom	.5644	.001694	.04702	0.3868

Table 2.2: Relevant posterior weights of θ

Figure 2.6 shows the results in this case for the one- and two-step-ahead predictions given only 5 data collection times. The true pmf is correctly predicted by the FV-DDP estimate even in this short horizon scenario, and the associated 95% pointwise credible intervals are significantly sharper if compared to Figure 2.5, obtained with a longer horizon. The prediction based on the alternative DDP mixture does not infer correctly the target, leading to an associated normalised ℓ_1 distance from the true pmf of 12.72% and 12.84%, compared to 4.95% and 4.90% for the FV-DDP prediction.

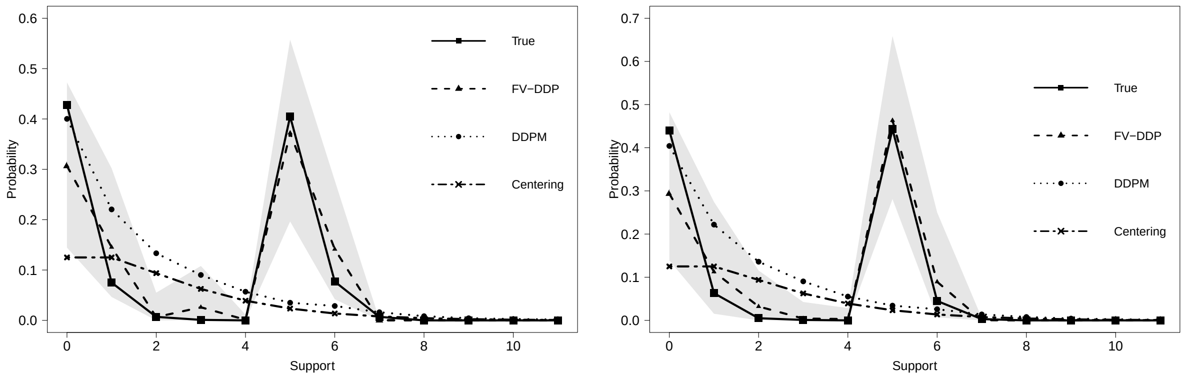


Figure 2.6: One- (left) and two-step-ahead prediction (right) based on 5 data collection times, with 95% pointwise credible intervals.

Karnofsky score data

We consider the dataset *hodg* used in Klein and Moeschberger (2003), which contains records on the time to death or relapse and the Karnofsky score for 43 patients with a lymphoma disease. The Karnofsky score (KS) is an index attributed to individual patients, with higher values indicating a better prognosis.

In the framework of model (2.7), we take the times of death or relapse as collection times and let the KS of the survivors at each time be the data. We aim at predicting the future distribution of the KS among the patients who are still in the experiment at that time, which would be an indirect assessment of the effectiveness of the score in describing the patients' prognosis. We also include censored observations (patients leaving the experiment for reasons different from death or relapse), without having them trigger a collection time. The FV-DDP appears as the ideal modeling tool in this framework since it includes a probabilistic mechanism that accounts for the reduced number of observations through different time points.

We train the model up to 42, 108 and 406 days after the start of the experiment, and we make predictions 28, 112 and 144 days ahead, respectively. As regards the prior, we put a uniform distribution on the observed scores (note that new score values cannot appear along the experiment) and we uniformly randomize θ over $\{.5, 1, 1.5, \dots, 15\}$, analogously to Section 2.2.4. Given the results of the previous subsection for different approaches to selecting σ, here, after transforming the lags in annual, we proceed by selecting σ for each value of θ by maximizing the probability that the death process makes the right number of transitions in the desired laps of time. Some of the selected values for $\sigma_1, \sigma_2, \sigma_3$ for the three different trainings, depending on θ, are shown in Table 2.3.

Figure 2.7 shows the three predictions of the scores distribution. Coherently with the intuition, as the experiment goes by, individuals with higher KS become predominant: from 70 to 230 days the predicted weight associated to a score of 90 increases of more than 10%, and

θ	.5	1	1.5	$\cdots$	29	29.5	30
σ_1	0.4947	0.4913	0.4885	$\cdots$	0.3235	0.3266	0.3228
σ_2	0.6059	0.6014	0.5696	$\cdots$	0.3684	0.3130	0.3361
σ_3	0.6149	0.6150	0.5789	$\cdots$	0.3063	0.3018	0.2901

Table 2.3: Choice of σ for some values of θ for the three trainings.

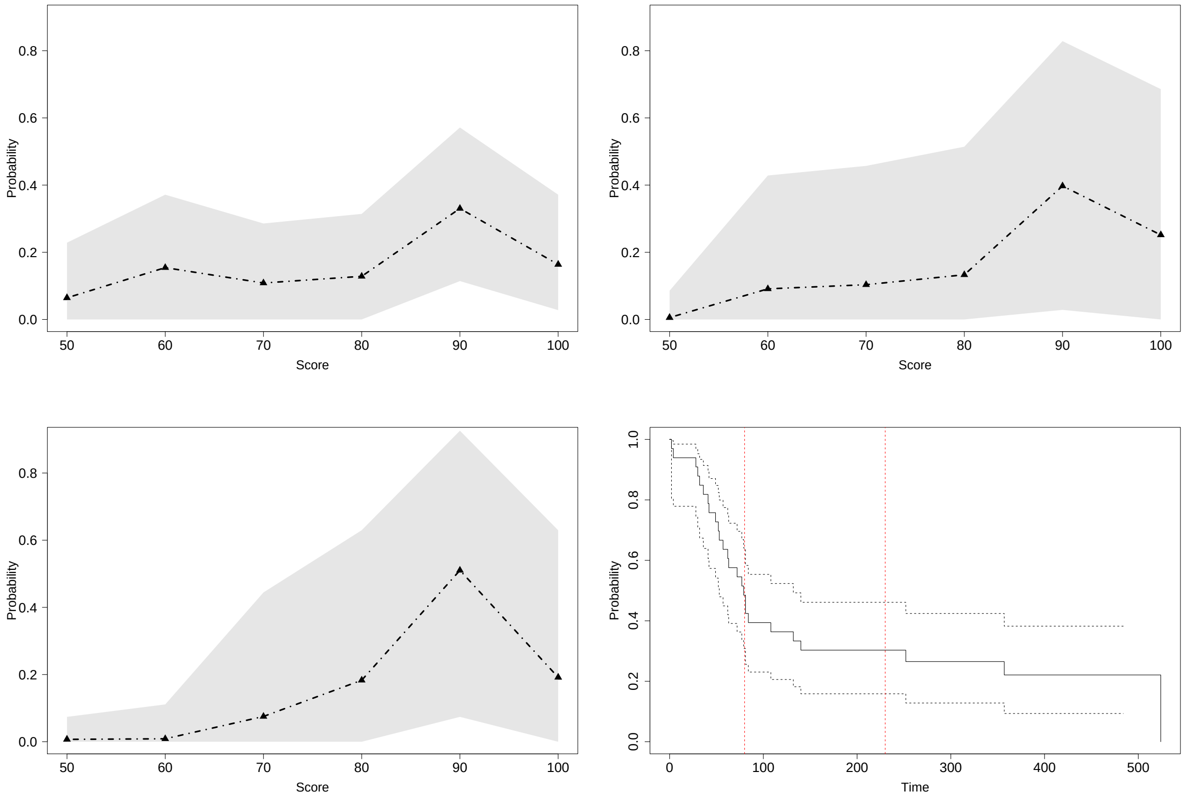


Figure 2.7: From top left: pmf prediction at 70, 230 and 550 days after the experiment. Bottom right: Kaplan-Meyer estimate of the survival times up to time 550.

similarly for 100. However the distribution of the scores remains pretty stable, apart from the lowest values, meaning that the highest scoring patients actually had much better prognoses, as showed by the third prediction.

These findings are consistent with the Kaplan-Meyer estimate (Kaplan and Meier, 1958) of the survival function, shown in the bottom right panel, which decreases rapidly between 70 and 230 and flattens after that point, implying that the FV-DDP prediction adapted to the periods of quick change in the underlying distribution and periods of relative steady behaviour.

2.2.5 Smoothing distribution

In this Section we are interested in determining the so-called *smoothing distributions* of the marginal states P_{t_i} of an unobserved Markov process, often called the hidden or latent *signal*, evaluated at time t_i given samples $X_{t_0}, \dots, X_{t_{p-1}}$ from the observation model, which is parameterised by the signal state, collected before and after t_i. Here $0 = t_0 < \cdots < t_{p-1} = T$ and $0 < i < p-1$. These conditional distributions are typically used to improve previous estimates obtained at a certain time once additional observations become available at later times, often resulting in a smoother estimated trajectory for the unobserved signal, and they also constitute the starting point for performing Bayesian inference on the model parameters (see, e.g., Kon Kam King et al. (2021)).

In particular we characterize the laws of the marginal states of FV model (2.7), given samples from the respective underlying populations collected before and after the state temporal index, thus solving the smoothing problem. We show that these distributions can be written as finite mixtures of laws of Dirichlet random measures respectively, whose time-dependent mixture weights are fully described and can account for different time intervals between data collection times. As a byproduct of the above results, we describe the predictive distribution for further samples from the population given the entire dataset, which are shown to be mixtures of generalized Pólya urns. Our results prove that computable smoothing and conditional sampling from the population are feasible with signals given by the FV process, bringing forward this model as possible canonical choice in a nonparametric framework for hidden Markov models.

Some operators on measures

To obtain the smoothing distributions we are going to exploit the projective properties of the FV process. To this end, we need to set a few tools that ease notation and computation for the respective finite-dimensional counterparts.

Consider a Markov process $\mathbf{P}$ on $\mathbb{R}^K$, $K < \infty$, with transition function Q_t and initial distribution ν. We assume Q_t is reversible with reversible measure ν_0. In this section we regard $\mathbf{P}$ as generic, but as anticipated above this setting will be used to model finite-dimensional projections of the measure-valued diffusion P_t. The dimension K can therefore be thought of as representing the number of cells in which types in the population have been grouped or binned. Accordingly, *iid* observations collected at time t given $\mathbf{P}_t = \mathbf{p}_t$ generate multiplicities associated to the K groups which can be encoded into a vector $\mathbf{n} \in \mathbb{Z}_+^K$, whose associated density is $p(\mathbf{n}|\mathbf{p}_t)$. We can then define the following operators acting on measures ξ:

- *Update*:
$$\mathcal{U}_{\mathbf{n}}(\xi)(\mathrm{d}\mathbf{p}) := \frac{p(\mathbf{n}|\mathbf{p})\xi(\mathrm{d}\mathbf{p})}{p_\xi(\mathbf{n})}, \qquad p_\xi(\mathbf{n}) := \int p(\mathbf{n}|\mathbf{p})\xi(\mathrm{d}\mathbf{p}). \tag{2.19}$$
This provides the conditional measure ξ given observations with associated multiplicities $\mathbf{n}$. It is analogous to Bayes' Theorem with $\mathbf{p}$ acting as the random parameter and ξ as its prior: hence, $\mathcal{U}_{\mathbf{n}}(\xi)$ yields the posterior, whereas the denominator $p_\xi(\mathbf{n})$ is the marginal likelihood of $\mathbf{n}$ obtained by integrating out the random parameter $\mathbf{x}$. Here the multiplicities $\mathbf{n}$ are observed at the same time $\mathbf{p}$ refers to, as in (2.7).

- *Forward propagation*:
$$\mathcal{F}_t(\xi)(\mathrm{d}\mathbf{p}') := \xi Q_t(\mathrm{d}\mathbf{p}') = \int \xi(\mathrm{d}\mathbf{p})Q_t(\mathbf{p}, \mathrm{d}\mathbf{p}'). \tag{2.20}$$

This yields the unconditional measure of P_{s+t} if ξ is that of P_s, once the initial state is integrated out.

- *Backward propagation*:

$$\mathcal{B}_t(\xi)(\mathrm{d}\mathbf{p}') := \xi Q'_t(\mathrm{d}\mathbf{p}') = \int \xi(\mathrm{d}\mathbf{x})Q_t(\mathbf{x}, \mathrm{d}\mathbf{x}'). \tag{2.21}$$

It is the forward propagation obtained by using the transition function of the time reversal of the signal, denoted here Q'_t.

With a slight abuse of notation, when ξ is dominated by a sigma-finite measure μ on $\mathbb{R}^K$, we specialize the previous operators as acting on densities, e.g., if $\xi(\mathrm{d}\mathbf{p}) = f(\mathbf{p})\mu(\mathrm{d}\mathbf{p})$, then $\mathcal{U}_\mathbf{n}(f)(\mathbf{p}) := p(\mathbf{n}|\mathbf{p})f(\mathbf{p})/p_f(\mathbf{n})$ and $p_f(\mathbf{n}) := \int p(\mathbf{n}|\mathbf{p})f(\mathbf{p})\mu(\mathrm{d}\mathbf{p})$, and similarly for (2.20) and (2.21). Note that the specific form of the backward transition is not necessary for our treatment, as we will leverage on Bayes' Theorem. See, e.g., Lemma 3 below.

Remark 1. *Expanding on the above, and assuming all probability distributions of interest are dominated by μ, one could define a* smoothing operator *acting on two densities f_s, f_u, indexed by $s < u$, by letting*

$$\mathcal{S}^{\mathbf{n}}_{s,t,u}(f_s, f_u)(\mathbf{p}) := C\ \mathcal{F}_{t-s}(f_s)(\mathbf{x})\mathcal{B}_{u-t}(f_u)(\mathbf{p})\mathcal{U}_\mathbf{n}(f_0)(\mathbf{p})/f_0(\mathbf{p})^2, \tag{2.22}$$

for every $\mathbf{p}$ such that $f_0(\mathbf{p}) > 0$, where $s < t < u$, f_0 is the density of ν_0 with respect to μ and C is a normalising constant that makes the left hand side a density. This yields the distribution of $\mathbf{P}_t$, given observations at time t, if $\mathbf{P}_s \sim f_s$ and $\mathbf{P}_u \sim f_u$, obtained by jointly propagating $\mathbf{P}_s$ forward of a $t-s$ interval, $\mathbf{P}_u$ backward of a $u-t$ interval, and then conditioning on observations collected at time t. The rationale of this operator can be outlined by considering that, for $t_0 < t_1 < t_2$, if $x_{t_1}|\mathbf{p}_{t_1} \sim p(x_{t_1}|\mathbf{p}_{t_1})$, we have

$$\begin{aligned} p(\mathbf{p}_{t_1}|\mathbf{p}_{t_0}, x_{t_1}, \mathbf{p}_{t_2}) &\propto p(\mathbf{p}_{t_1}|\mathbf{p}_{t_0})p(x_{t_1}|\mathbf{p}_{t_1}, \mathbf{p}_{t_0})p(\mathbf{p}_{t_2}|\mathbf{p}_{t_1}, \mathbf{p}_{t_0}, x_{t_1}) \\ &= p(\mathbf{p}_{t_1}|\mathbf{p}_{t_0})p(x_{t_1}|\mathbf{p}_{t_1})p(\mathbf{p}_{t_2}|\mathbf{p}_{t_1}) \end{aligned}$$

where we have used the fact that conditionally on $\mathbf{p}_{t_1}$, y_{t_1} is independent on everything else, together with the Markov property. By virtue of Bayes' Theorem, we now have $p(\mathbf{p}_{t_2}|\mathbf{p}_{t_1}) = p(\mathbf{p}_{t_2})p(\mathbf{p}_{t_1}|\mathbf{p}_{t_2})/p(\mathbf{p}_{t_1})$ and $p(\mathbf{p}_{t_1}|x_{t_1}) = p(\mathbf{p}_{t_1})p(x_{t_1}|\mathbf{p}_{t_1})/p(x_{t_1})$ whereby the previous expression is proportional to $p(\mathbf{p}_{t_1}|\mathbf{p}_{t_0})p(\mathbf{p}_{t_1}|\mathbf{p}_{t_2})p(\mathbf{p}_{t_1}|x_{t_1})/p(\mathbf{p}_{t_1})^2$. This operator will not be essential for our calculations of the next sections, but it can provide a unified treatment of the previous operators applied at stationarity. In fact, we have the following special cases:

$$\mathcal{S}^{\mathbf{0}}_{s,t,u}(f_s, f_0) = \mathcal{F}_{t-s}(f_s), \qquad \mathcal{S}^{\mathbf{0}}_{s,t,u}(f_0, f_u) = \mathcal{B}_{u-t}(f_u), \qquad \mathcal{S}^{\mathbf{n}}_{s,t,u}(f_0, f_0) = \mathcal{U}_\mathbf{n}(f_0).$$

Appropriate compositions of the above operators allow to represent all quantities of interest in this framework. For example $p(\mathbf{p}_{t_i}|\mathbf{n}_{i-1}, \mathbf{n}_i, \mathbf{n}_{i+1}) = \mathcal{S}^{\mathbf{n}_i}_{t_{i-1},t_i,t_{i+1}}(\mathcal{U}_{\mathbf{n}_{i-1}}(f_0), \mathcal{U}_{\mathbf{n}_{i+1}}(f_0))$ identifies the distribution of $\mathbf{P}_{t_i}$ given observations at times t_{i-1}, t_i, t_{i+1}, obtained by first updating the stationary measure at times t_{i-1}, t_{i+1} given observations with multiplicities $\mathbf{n}_{t-1}, \mathbf{n}_{t_{i+1}}$ respectively, then propagating both distributions to the intermediate time t_i, and finally updating the output of the last operation given the multiplicities observed at time t_i.

Preliminary results on projections

A projection of the Dirichlet process law onto a measurable partition $(A_1,\dots,A_K)$ of the sampling space $\mathbb{X}$ yields a Dirichlet distribution with parameters $(\alpha(A_1),\dots,\alpha(A_K))$, whose density with respect to the Lebesgue measure on the $(K-1)$-dimensional simplex, denoted $\pi_{\boldsymbol{\alpha}}(\mathbf{x})$, $\boldsymbol{\alpha}=(\alpha(A_1),\dots,\alpha(A_K))$, is proportional to $\mathbf{p}^{\boldsymbol{\alpha}-\mathbf{1}}:=p_1^{\alpha(A_1)-1}\dots p_K^{\alpha(A_K)-1}$. With a little abuse of notation, we will use the symbol $\pi_{\boldsymbol{\alpha}}$ to denote both the Dirichlet density and the corresponding measure. Similarly, a projection of a FV process with transition (2.6) onto the same partition yields a Wright–Fisher diffusion, denoted $\mathrm{WF}_{\boldsymbol{\alpha}}$, with transition function

$$Q_t(\mathbf{p},\mathrm{d}\mathbf{p}')=\sum_{m=0}^{\infty}d_m(t)\sum_{\mathbf{m}\in\mathbb{Z}_+^K:\,|\mathbf{m}|=m}\mathbf{p}^{\mathbf{m}}\binom{m}{\mathbf{m}}\pi_{\boldsymbol{\alpha}+\mathbf{m}}(\mathrm{d}\mathbf{p}'), \tag{2.23}$$

and $d_m(t)$ as in (2.6), which has reversible distribution $\pi_{\boldsymbol{\alpha}}$. We will denote by $q_t(\cdot\,|\mathbf{p})$ the corresponding density function. In this scenario, (2.7) reduces to

$$X_t^i|P_t=\mathbf{P}\overset{\text{iid}}{\sim}\text{Categorical}(\mathbf{p}),\quad \mathbf{P}\sim\mathrm{WF}_{\boldsymbol{\alpha}}, \tag{2.24}$$

whereby for each i, $X_t^i=j$ with probability p_j, for $j=1,\dots,K$, and the update operator (2.19) yields the familiar Bayesian update for Dirichlet distributions $\mathcal{U}_{\mathbf{n}}(\pi_{\boldsymbol{\alpha}})=\pi_{\boldsymbol{\alpha}+\mathbf{n}}$. It is useful to note for later reference that in (2.19) the marginal likelihood is

$$m(\mathbf{n}):=p_{\pi_\alpha}(\mathbf{n})=\frac{B(\boldsymbol{\alpha}+\mathbf{n})}{B(\boldsymbol{\alpha})},\qquad B(\boldsymbol{\alpha}):=\frac{\prod_{j=1}^K\Gamma(\alpha_j)}{\Gamma(|\boldsymbol{\alpha}|)} \tag{2.25}$$

often called Dirichlet-Categorical distribution. Define now

$$h(\mathbf{p},\mathbf{n}):=\frac{p(\mathbf{n}|\mathbf{p})}{m(\mathbf{n})}=\frac{B(\boldsymbol{\alpha})}{B(\boldsymbol{\alpha}+\mathbf{n})}\mathbf{p}^{\mathbf{n}},\qquad \mathbf{n}\in\mathbb{Z}_+^K, \tag{2.26}$$

where $p(\mathbf{n}|\mathbf{p})$ is the categorical likelihood in (2.24) expressed in terms of multiplicities of types. It will also be useful to note that for $\mathbf{n},\mathbf{m}\in\mathbb{Z}_+^K$, we have

$$h(\mathbf{p},\mathbf{n})h(\mathbf{p},\mathbf{m})=c(\mathbf{n},\mathbf{m})h(\mathbf{p},\mathbf{n}+\mathbf{m}) \tag{2.27}$$

where

$$c(\mathbf{n},\mathbf{m})=\frac{m(\mathbf{n}+\mathbf{m})}{m(\mathbf{n})m(\mathbf{m})}=\frac{B(\boldsymbol{\alpha})B(\boldsymbol{\alpha}+\mathbf{m}+\mathbf{n})}{B(\boldsymbol{\alpha}+\mathbf{n})B(\boldsymbol{\alpha}+\mathbf{m})}. \tag{2.28}$$

Recall now that the WF diffusion is known to have moment-dual given by Kingman's *typed* coalescent. More specifically, let $\mathbf{M}_t$ be a death process on $\mathbb{Z}_+^K$ with rates $\lambda_{\mathbf{m}}=m_j(\theta+|\mathbf{m}|-1)/2$ from $\mathbf{m}$ to $\mathbf{m}-\mathbf{e}_j$, where $\mathbf{e}_j$ is the canonical vector in direction j. Then the following duality identity

$$\mathbb{E}[h(\mathbf{P}_t,\mathbf{m})|\mathbf{P}_0=\mathbf{p}]=\mathbb{E}[h(\mathbf{p},\mathbf{M}_t)|\mathbf{P}_0=\mathbf{m}] \tag{2.29}$$

holds with h as in (2.26). We will denote by $p_{\mathbf{n},\mathbf{m}}(t)$ the transition probabilities of $\mathbf{M}_t$ (cf. Papaspiliopoulos et al. (2016), Section 4.2). The above duality is used to prove the following Lemma, needed later.

Lemma 2. *Let* $\mathbf{n}_{i+1}$ *be the multiplicities observed at time* t_{i+1}*. Then*

$$p(\mathbf{n}_{i+1}|\mathbf{p}_{t_i}) = m(\mathbf{n}_{i+1}) \sum_{\mathbf{0}\leq\mathbf{m}\leq\mathbf{n}_{i+1}} p_{\mathbf{n}_{i+1},\mathbf{m}}(t_{i+1}-t_i)h(\mathbf{p}_{t_i},\mathbf{m}). \tag{2.30}$$

The next lemma formalizes the fact that a backward propagation, after a change of measure with respect to the stationary distribution, yields an analogous distributional result to a forward propagation, somehow carrying over the reversibility of FV processes to their conditional versions. In the following statement, the forward and backward operators $\mathcal{F}_t, \mathcal{B}_t$ are applied to laws of FV states by extension of (2.20)-(2.21), with Q_t as in (2.6).

Lemma 3. *Assume* (2.7)*, let* $x_1^*,\dots,x_K^*$ *be distinct values and* $\mathbf{n}\in\mathbb{Z}_+^K$*. Then* $\mathcal{F}_t(DP\left(\alpha+\sum_{j=1}^K n_j\delta_{x_j^*}\right)) = \mathcal{B}_t(DP\left(\alpha+\sum_{j=1}^K n_j\delta_{x_j^*}\right))$*, with* $\mathcal{F}_t, \mathcal{B}_t$ *as in* (2.20)-(2.21)*, and in particular*

$$\mathcal{B}_t\Big(DP\left(\alpha+\sum_{j=1}^K n_j\delta_{x_j^*}\right)\Big) = \sum_{\mathbf{0}\leq\mathbf{k}\leq\mathbf{n}} p_{\mathbf{n},\mathbf{k}}(t)DP\left(\alpha+\sum_{j=1}^K k_j\delta_{x_j^*}\right). \tag{2.31}$$

Thanks to this equivalence between backward and forward propagation, we have that the same result of Lemma 2 holds with $\mathbf{n}_{i+1}$ replaced by $\mathbf{n}_{i-1}$, i.e., referred to time t_{i-1}, leading to the expression obtained by replacing $t_{i+1}-t_i$ with t_i-t_{i-1} in the right hand side of (2.30).

Main result

Using the results of the previous section, the characterization of the smoothing distribution for conditional FV processes will be provided in three steps. First, in Theorem 9, we show that conditioning on observations collected at adjacent times yields a finite mixture of laws of Dirichlet random measures; then, in Proposition 6, we give a full description of the mixture weights for different choices of the offspring distribution Q_0; finally, in Proposition 7, we show how the general expression can be obtained by recursive computation based on the previous results.

We denote by $\mathbf{X}_{i-1},\mathbf{X}_i,\mathbf{X}_{i+1}$ vectors of observations collected as in (2.7) at times t_{i-1}, t_i, t_{i+1} respectively, with associated multiplicities $\mathbf{n}_{i-1},\mathbf{n}_i,\mathbf{n}_{i+1}$ for the distinct values $(x_1^*,\dots,x_K^*)$ observed overall.

Theorem 9. *Under model* (2.7)*, let* $\boldsymbol{X}_{i-1},\boldsymbol{X}_i,\boldsymbol{X}_{i+1}$ *be as above. Then there exist weights summing up to one, denoted* $w_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}(\Delta_i,\Delta_{i+1})$*, for* $\mathbf{k}_{i-1}\leq\mathbf{n}_{i-1},\mathbf{k}_{i+1}\leq\mathbf{n}_{i+1}$*, such that*

$$\begin{aligned}\mathcal{L}(P_{t_i}\,|\boldsymbol{X}_{i-1},\boldsymbol{X}_i,\boldsymbol{X}_{i+1}) =\\ = \sum_{\mathbf{0}\leq\mathbf{k}_{i-1}\leq\mathbf{n}_{i-1}}\sum_{\mathbf{0}\leq\mathbf{k}_{i+1}\leq\mathbf{n}_{i+1}} w_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}(\Delta_i,\Delta_{i+1})DP\left(\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}\right),\end{aligned} \tag{2.32}$$

where $\Delta_i = t_i - t_{i-1}$*,* $\Delta_{i+1} = t_{i+1}-t_i$*, and*

$$\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}} = \alpha + \sum_{j=1}^K (k_{i-1,j}+n_{i,j}+k_{i+1,j})\delta_{x_j^*}. \tag{2.33}$$

Remark 2. *In the previous result, we have used notation* $\mathbf{k}_{i-1}$ *and* $\mathbf{k}_{i+1}$ *for the integrating variables, whose indices should help the intuition by indicating the time point they refer to.*

Note however that in principle these quantities are determined at time t_i, being the number of lineages in a time-reversed genealogy. Instead of using generic integrating variables $\mathbf{i}, \mathbf{j}$, we choose to adopt this notational convention here and later for the sake of readability.

The previous result provides an explicit representation of the conditional law of a FV state given observations at adjacent times, but does not investigate in full detail the mixture weights, denoted generically in the Theorem statement, which we do next. To pursue this task, by looking at the proof of Theorem 9 we need to compute

$$\lim_{n\to\infty} \frac{m^{(n)}(\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1})}{C_n m^{(n)}(\mathbf{k}_{i-1}) m^{(n)}(\mathbf{n}_i) m^{(n)}(\mathbf{k}_{i+1})}, \tag{2.34}$$

where $m^{(n)}$ denotes the marginal distribution in (2.25) relative to the model induced by the partition $\mathcal{B}_n$ and C_n is the normalizing constant. In the setting of Theorem 9, denote now by

$$\begin{aligned} D_{i-1} &= \left\{ j \in \{1,\dots,K\} \,:\, n_{i-1,j} > 0 \text{ and either } n_{i,j} > 0 \text{ or } n_{i+1,j} > 0 \right\}, \\ D_{i+1} &= \left\{ j \in \{1,\dots,K\} \,:\, n_{i+1,j} > 0 \text{ and either } n_{i,j} > 0 \text{ or } n_{i-1,j} > 0 \right\}, \end{aligned}$$

the set of distinct values in $\mathbf{X}_{i-1}$ shared with $\mathbf{X}_i$ or $\mathbf{X}_{i+1}$, and those in $\mathbf{X}_{i+1}$ shared with $\mathbf{X}_{i-1}$ or $\mathbf{X}_i$, respectively. Then

$$\mathcal{D} = \Big\{ (\mathbf{k}, \mathbf{k}') \leq (\mathbf{n}_{i-1}, \mathbf{n}_{i+1}) \,:\, k_j > 0,\, k'_{j'} > 0, \forall j \in D_{i-1} \text{ and } j' \in D_{i+1} \Big\}$$

is the set of multiplicities $(\mathbf{k}, \mathbf{k}')$ not greater than $(\mathbf{n}_{i-1}, \mathbf{n}_{i+1})$ such that the frequency of distinct values shared between different collection times is strictly positive. For example, if t_i is the current time index, suppose we have $\mathbf{n}_{i-1} = (1,3,0)$, $\mathbf{n}_i = (0,0,1)$ and $\mathbf{n}_{i+1} = (0,2,1)$, whereby of the three types observed overall, at time t_{i-1} we observed multiplicities 1 and 3 for the first two, at time t_i an instance of a third type, and so on. Then

$$\mathcal{D} = \big\{ (\mathbf{i}, \mathbf{j}) \,:\, \mathbf{i} \leq (1,3,0),\, \mathbf{j} \leq (0,2,1),\, i_2 > 0, j_2 > 0, j_3 > 0 \big\}$$

is given by vectors of multiplicities not greater than $(\mathbf{n}_{i-1}, \mathbf{n}_{i+1})$, with positive entries for type two, which is shared by times t_{i-1}, t_{i+1}, and for type three, limited to the second coordinate, since it is shared by time t_{i+1} and the current time. In other words, multiplicities not greater than those observed, with positive entries for types: (i) observed at both times different from the current, or (ii) observed at the current time and at least another time. Notice that $\mathcal{D} = \emptyset$ corresponds to the case in which no values are shared between the three collection times t_{i-1}, t_i and t_{i+1}, which holds, for example, when all the observations are distinct.

Before stating the result, note that when Q_0 is supported by a countably infinite set, $\mathbf{m}(\mathbf{n})$ can be defined by extension of (2.25), where all but a finite number of terms simplify in the ratio. Let also $a^{(b)} = a(a+1)\dots(a+b-1)$ denote the Pochhammer symbol.

Proposition 6. *In the setting of Theorem 9, let $\tilde{p} = p_{\mathbf{n}_{i-1},\mathbf{k}_{i-1}}(\Delta_i) p_{\mathbf{n}_{i+1},\mathbf{k}_{i+1}}(\Delta_{i+1})$. Then*

A. if Q_0 is discrete,

$$w^{\mathbf{n}_{i-1},\mathbf{n}_{i+1}}_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}(\Delta_i, \Delta_{i+1}) \propto \tilde{p}\, \frac{m(\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1})}{m(\mathbf{k}_{i-1}) m(\mathbf{n}_i) m(\mathbf{k}_{i+1})}; \tag{2.35}$$

B. if Q_0 is nonatomic and $\mathcal{D} = \emptyset$,

$$w^{\mathbf{n}_{i-1},\mathbf{n}_{i+1}}_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}(\Delta_i, \Delta_{i+1}) \propto \tilde{p}\, \frac{\theta^{(|\mathbf{k}_{i-1}|)}\theta^{(|\mathbf{k}_{i+1}|)}}{(\theta + |\mathbf{n}_i|)^{(|\mathbf{k}_{i-1}|+|\mathbf{k}_{i+1}|)}};$$

C. if Q_0 is nonatomic and $\mathcal{D} \neq \emptyset$,

$$w^{\mathbf{n}_{i-1},\mathbf{n}_{i+1}}_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}(\Delta_i, \Delta_{i+1}) \propto \tilde{p}\, \frac{\theta^{(|\mathbf{k}_{i-1}|)}\theta^{(|\mathbf{k}_{i+1}|)}}{(\theta + |\mathbf{n}_i|)^{(|\mathbf{k}_{i-1}|+|\mathbf{k}_{i+1}|)}} \prod_{j=1}^{K} \frac{(k_{i-1,j} + n_{i,j} + k_{i+1,j} - 1)!}{(k_{i-1,j} - 1)!\,(n_{i,j} - 1)!\,(k_{i+1,j} - 1)!}$$

if $(\mathbf{k}_{i-1}, \mathbf{k}_{i+1}) \in \mathcal{D}$, and zero otherwise.

We now have a full description of (2.32), which is a finite mixture of laws of Dirichlet random measures whose parameter measure $\alpha + \sum_{j=1}^{K}(k_{i-1,j} + n_{i,j} + k_{i+1,j})\delta_{x_j^*}$ contains, besides the unnormalised offspring measure α, the current observations $\mathbf{X}_i$ and a subset of the observations $(\mathbf{X}_{i-1}, \mathbf{X}_{i+1})$ collected at adjacent times. The mixture weights are in turn determined by the following two elements. The first is given by the transition probabilities of the death process associated to the typed coalescent, which determines the probability that past and future data are atoms in the respective random measures as a function of the distance between time t_i and the adjacent times. As the lags Δ_i and Δ_{i+1} grow, the number of survived lineages is lower with higher probability and the random measures in the mixture will carry, on average, less information in terms of types observed at different times. The second element is the joint marginal likelihood of past, present and future data. For instance, when the offspring distribution is discrete, the ratio in (2.35) is higher when $m(\mathbf{k}_{i-1} + \mathbf{n}_i + \mathbf{k}_{i+1}) > m(\mathbf{k}_{i-1})m(\mathbf{n}_i)m(\mathbf{k}_{i+1})$, i.e. when sampling jointly $\mathbf{k}_{i-1}$, $\mathbf{n}_i$ and $\mathbf{k}_{i+1}$ has high probability relative to sampling them separately: this provides a smoothing effect by favouring the nodes $(\mathbf{k}_{i-1}, \mathbf{n}_i, \mathbf{k}_{i+1})$ with the same types collected at different times. Such mechanism comes to an extreme when the offspring distribution is nonatomic. In this case, the weights of the mixture components that do not carry atoms observed at multiple times, in the sense of the set $\mathcal{D}$, vanish in the limit.

The above results do not include the case of α having both a continuous and discrete component. The same tools used for proving Proposition 6 can in principle be used to deal with this case as well, where we expect parts A, B and C of the statement to hold for the respective parts of the parameters measure. In essence, values drawn by the discrete part of α are subjected to the probability that lineages survive as controlled by the term $\tilde{p}$ (hence ultimately by the death process), whereas values drawn from the continuous part of α are also, in addition, subjected to whether they are shared across collection times. A full description of such results would require a cumbersome notation and would not add further valuable insight, hence, we will not pursue this task here.

Let now $\mathbf{X}_{0:T}$ be the entire dataset sampled in model (2.7), and let K be the number of distinct values in $\mathbf{X}_{0:T}$. Denoting by $\overleftarrow{\mathbf{n}}_{i-1}$ the total multiplicities of the vector $\mathbf{X}_{t_{0:i-1}}$, we know there exist weights $\{v_{1,\mathbf{m}}\}$ such that

$$\mathcal{L}(P_{t_i}|\mathbf{X}_{t_{0:i-1}}) = \sum_{\mathbf{k}_{i-1} \leq \overleftarrow{\mathbf{n}}_{i-1}} v_{1,\mathbf{k}_{i-1}} DP\left(\alpha + \sum_{j=1}^{K} k_{i-1,j}\delta_{x_j^*}\right). \tag{2.36}$$

This can be obtained recursively starting from $P_{t_0} \sim DP(\alpha)$, where the reversible measure $DP(\alpha)$ acts as prior (or unconditional distribution) for the marginal state of the signal. Upon observing $\mathbf{X}_{t_0}$, the update yields $\mathcal{L}(P_{t_0}|\mathbf{X}_{t_0}) = DP\left(\alpha + \sum_{j=1}^{K_0} n_{0,j}\delta_{x_j^*}\right)$,where $\mathbf{n}_0$ is the vector of multiplicities of the K_0 distinct values in $\mathbf{X}_{t_0}$. Propagating forward the previous through $\mathcal{F}_t$, one obtains

$$\mathcal{L}(P_{t_1}|\mathbf{X}_{t_0}) = \sum_{\mathbf{k}_0 \leq \mathbf{n}_0} p_{\mathbf{n}_0,\mathbf{k}_0}(t_1 - t_0) DP\left(\alpha + \sum_{j=1}^{K_0} k_{0,j}\delta_{x_j^*}\right),$$

which can then be updated once data in t_1 become available by observing that (2.19) satisfies

$$\mathcal{U}_{\mathbf{n}}\left(\sum_{i=1}^{H} w_i \xi_i\right) = \sum_{i=1}^{H} \frac{w_i p_{\xi_i}(\mathbf{n})}{\sum_{h=1}^{H} w_h \xi_h(\mathbf{n})} \mathcal{U}_{\mathbf{n}}(\xi_i).$$

Proceeding in this way, alternating updates and forward propagations, leads to (2.36); see Papaspiliopoulos et al. (2016), Section 3.1, for further details. Denoting now by $\overrightarrow{\mathbf{n}}_{i+1}$ the total multiplicities of the vector $\mathbf{X}_{t_{i+1:T}}$, by virtue of Lemma 3 and of the linearity of (2.21), there exist weights $\{v_{2,\mathbf{n}}\}$ such that

$$\mathcal{L}(P_{t_i}|\mathbf{X}_{t_{i+1:T}}) = \sum_{\mathbf{k}_{i+1} \leq \overrightarrow{\mathbf{n}}_{i+1}} v_{2,\mathbf{k}_{i+1}} DP\left(\alpha + \sum_{j=1}^{K} k_{i+1,j}\delta_{x_j^*}\right). \tag{2.37}$$

This can also be obtained by working backwards from $P_T \sim DP(\alpha)$, then updating given the multiplicities $\mathbf{n}_{p-1}$ of the K_{p-1} distinct values in $\mathbf{X}_T$, which yields $\mathcal{L}(P_T|\mathbf{X}_T) = DP\left(\alpha + \sum_{j=1}^{K_{p-1}} n_{p-1,j}\delta_{x_j^*}\right)$, then, using Lemma 3, propagating backwards to get

$$\mathcal{L}(P_T|\mathbf{X}_T) = \sum_{\mathbf{k}_{p-1} \leq \mathbf{n}_{p-1}} p_{\mathbf{n}_{p-1},\mathbf{k}_{p-1}}(t_{p-1} - t_{p-2}) DP\left(\alpha + \sum_{j=1}^{K_{p-1}} k_{p-1,j}\delta_{x_j^*}\right),$$

and so on. The following proposition connects the two above distributions to yield the general representation.

Proposition 7. *Assume* (2.36) *and* (2.37) *hold, and let* $\mathbf{n}_i$ *be the vector of multiplicities of* $\boldsymbol{X}_{t_i}$ *relative to the* K *distinct values in the whole dataset* $\boldsymbol{X}_{0:T}$. *Then*

$$\mathcal{L}(P_{t_i}|\boldsymbol{X}_{0:T}) = \sum_{\mathbf{k}_{i-1} \leq \overleftarrow{n}_{i-1}} \sum_{\mathbf{k}_{i+1} \leq \overrightarrow{n}_{i+1}} p_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}} DP\left(\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}\right), \tag{2.38}$$

where $\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}$ *is as in* (2.33) *and the weights*

$$p_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}} = \sum_{\mathbf{h} \leq \overleftarrow{n}_{i-1} : \, \mathbf{h} \geq \mathbf{k}_{i-1}} \sum_{\mathbf{l} \leq \overrightarrow{n}_{i+1} : \, \mathbf{l} \geq \mathbf{k}_{i+1}} v_{1,\mathbf{h}} v_{2,\mathbf{l}} w^{\mathbf{h},\mathbf{l}}_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}, \quad \mathbf{k}_{i-1} \leq \overleftarrow{n}_{i-1}, \mathbf{k}_{i+1} \leq \overrightarrow{n}_{i+1}, \tag{2.39}$$

with $w^{\mathbf{h},\mathbf{l}}_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}$ *as in Proposition 6, sum up to one.*

The proof of Proposition 7 clarifies that the smoothing mixture is computed in two steps: first $\mathcal{L}(P_{t_{i-1}}|\mathbf{X}_{t_{0:i-1}})$ and $\mathcal{L}(P_{t_{i+1}}|\mathbf{X}_{t_{i+1:N}})$ are computed through backward and forward filtering respectively, then the smoothing operator is applied, as in Theorem 9. The first step leads to two mixtures whose number of components is $\prod_{k=1}^{K}(1 + \sum_{j=0}^{i-1} n_{t_j,k})$ and $\prod_{k=1}^{K}(1 + \sum_{j=i+1}^{N} n_{t_j,k})$ respectively, as shown in Section 4 of Kon Kam King et al. (2021). Recall that here K is the number of distinct values observed in the entire dataset, which is considered as given. Since each distinct element of the smoothing distribution is now given by a distinct choice of $\mathbf{k}_{i-1}$ and $\mathbf{k}_{i+1}$, the total number of components in the smoothing distribution is therefore

$$\prod_{k=1}^{K}\left(1+\sum_{j=0}^{i-1} n_{t_j,k}\right)\left(1+\sum_{j=i+1}^{N} n_{t_j,k}\right). \tag{2.40}$$

As expected, smoothing comes at a greater nominal computational cost than filtering, since, roughly speaking, it combines information from both past and future. However, the actual cost of smoothing is expected to be much lower than the nominal, due to two factors. The first, specific to the current modelling assumptions, is that in the scenario of Statement C in Proposition 6, with a continuous baseline distribution, the number of components is automatically pruned by the smoothing operator, which discards values that are not shared across times. Hence (2.40) represents a crude upper bound. The second factor is that some mixture component weights are typically negligible. This aspect, which had already been noted in Chaleyat-Maurel and Genon-Catalot (2006) and was investigated in detail for Wright–Fisher and Cox–Ingersoll–Ross models in Kon Kam King et al. (2021), suggests various possible pruning strategies that allow to approximate the smoothing distribution, lowering the actual computational cost by some order of magnitudes while keeping a high precision in the approximation.

Predictive distributions

As a corollary to Proposition 7, we can derive the predictive distribution of further samples collected at time t_i, given the original data set $\mathbf{X}_{0:T}$. This extends Proposition 2.

Corollary 2. *In the setting of Proposition 7, let* (2.38) *be the conditional law of* P_{t_i} *given* $\boldsymbol{X}_{0:T}$. *Then the law of the* $(k+1)$*th further sample* X^{k+1} *from* P_{t_i} *is*

$$\begin{aligned}&\mathbb{P}(\, X^{k+1} \in A|\boldsymbol{X}_{0:T}, X^{1:k}) = \\ &= \sum_{\mathbf{k}_{i-1}\leq \overleftarrow{\boldsymbol{n}}_{i-1}} \sum_{\mathbf{k}_{i+1}\leq \overrightarrow{\boldsymbol{n}}_{i+1}} p_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}} \frac{\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}(A)+\sum_{j=1}^{k}\delta_{Y^j}(A)}{\theta+|\mathbf{k}_{i-1}|+|\mathbf{n}_i|+|\mathbf{k}_{i+1}|+k}\end{aligned} \tag{2.41}$$

for every Borel set A *of* $\mathbb{X}$, *with* $\alpha_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}$ *is as in* (2.33) *and* $p_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}$ *as in Proposition 7.*

Here for brevity we have used the notation X^{k+1} for the additional $(k+1)$-st sample instead of the correct notation $X^{|\mathbf{n}_i|+k+1}$, given the original dataset already contains $|\mathbf{n}_i|$ observations sampled at t_i. Recall that the predictive distribution for observations sampled from a Dirichlet random measure is described by the Blackwell–MacQueen Pólya urn scheme, whereby for $\alpha = \theta Q_0$,

$$X_1 \sim Q_0, \qquad X_{k+1}|X_1,\dots,X_k \sim \frac{\alpha+\sum_{j=1}^{k}\delta_{X_j}}{\theta+k}.$$

It is then clear that (2.41) is a finite mixture of generalized Pólya urn schemes, whose sampling mechanism can be described as follows. For each $k \geq 1$, given we already observed the further sample $Y^{1:k}$,

- choose a pair $(\mathbf{k}_{i-1}, \mathbf{k}_{i+1})$ with probability $p_{\mathbf{k}_{i-1},\mathbf{n}_i,\mathbf{k}_{i+1}}$
- draw a categorical random variable $J \in \{1, 2, 3\}$ with probabilities proportional to θ, $|\mathbf{k}_{i-1}|+|\mathbf{n}_i| + |\mathbf{k}_{i+1}|$ and k respectively

- given $X^{1:k}$, draw

$$X^{k+1} \sim \begin{cases} Q_0, & \text{if } J=1, \\ \frac{\sum_{j=1}^{K}(k_{i-1,j}+n_{i,j}+k_{i+1,j})\delta_{x_j^*}}{|\mathbf{k}_{i-1}|+|\mathbf{n}_i|+|\mathbf{k}_{i+1}|}, & \text{if } J=2, \\ \frac{1}{k}\sum_{j=1}^{k}\delta_{X^j}, & \text{if } J=3, \end{cases}$$

where $(x_1^*,\dots,x_K^*)$ in the second expression are the distinct values in $\mathbf{X}_{0:T}$.

We conclude the section with the observation that (2.38) is the de Finetti measure of the sequence $\{X^{|\mathbf{n}_i|+k}, k \geq 1\}$, i.e.

$$\mathbb{P}(X^{|\mathbf{n}_i|+k+1} \in \cdot\,|\mathbf{X}_{t_{0:N}}, X^{|\mathbf{n}_i|+1:|\mathbf{n}_i|+k}) \to P^* \qquad \text{a.s.}$$

weakly as $k \to \infty$ and P^* is the law in (2.38). This can be proved along the same lines of Proposition 5.

2.3 Clustering consistency with Dirichlet process mixtures

2.3.1 Introduction

As we discussed in the last chapter, Bayesian nonparametric methods have experienced a huge development in the last two decades, often standing out for their flexibility and coherent probabilistic foundations; see the monographs by Müller et al. (2017) and Ghosal and Van Der Vaart (2017) for recent stimulating accounts. The success of the Dirichlet process in actual implementations of the Bayesian approach to nonparametric problems is mostly due to its mathematical tractability, which is highlighted by conjugacy and flexibility, assessed in terms of its large topological support.

Let $P \sim DP(\alpha, Q_0)$ be random probability measure, where now $\alpha > 0$ denotes the concentration parameter: in this Section θ will be used to denote parameters of the likelihood. Since P is almost surely discrete, if one wishes to model continuous data one may convolve it with a density kernel k parametrized by a latent variable θ that is drawn from a Dirichlet process. This yields the popular Dirichlet process mixture (Lo, 1984), which exhibits appealing asymptotic properties in the context of density estimation: in several relevant cases, the posterior distribution concentrates at the true data-generating density at the minimax-optimal rate, up to a logarithmic factor, as the sample size increases (Ghosal et al., 1999; Ghosal and Van der Vaart, 2007). Such a model and many of its variants are widely used across scientific areas, thanks also to the availability of a wide variety of efficient computational methods to perform inference, see for instance Escobar and West (1995, 1998); MacEachern and Müller (1998); Neal (2000); Blei and Jordan (2006).

Since they are draws from the Dirichlet process, which is almost surely discrete, the latent parameters θ_i's exhibit ties with positive probability. Hence, the Dirichlet process mixture model is also routinely used to perform clustering since it partitions observations into groups based on whether their corresponding latent parameters θ_i coincide or not. The ubiquitous use of Dirichlet process mixtures for clustering motivates the interest in the asymptotic behaviour of the posterior distribution of the underlying partition, and in particular in the inferred number of clusters (i.e. subpopulations), as the number of observations increases. Nguyen (2013) showed posterior consistency of the mixing distribution P under general conditions. However, this does not imply consistency for the number of clusters, due to the use of the Wasserstein distance. Indeed, Miller and Harrison (2013) proved that Dirichlet process mixtures are not consistent for the number of components when data are generated from a mixture with a single standard normal component. See also Miller and Harrison (2014) for extensions. These results, however, are derived under the assumption that the concentration parameter α is known and fixed. This is crucial because the clustering behaviour of Dirichlet process mixtures is governed by the choice of α. Indeed, under the Dirichlet process mixture model, the prior probability of observing ties is a function solely of α, since $\mathbb{P}(\theta_i = \theta_j) = 1/(\alpha+1)$.

In order to have a more flexible distribution on the clustering of the data, in most implementations of the Dirichlet process mixture a prior π for α is specified, leading to a mixing measure that is itself a mixture in the sense of Antoniak (1974). Here we show that introducing such a prior has a major impact on the asymptotic behaviour of the number of clusters, as Dirichlet process mixtures can be consistent for the number of clusters.

We provide consistency results under fairly general conditions on π and for a moderately large class of kernels k, including uniform and truncated normal distributions. Following Miller and Harrison (2013), we focus on data-generating mixtures with a single component. Our results also extend to the more general case of finite mixtures with multiple components, when a suitable separation assumption between the elements of the mixtures is fulfilled. Crucially, we prove consistency for cases where using a non-random α yields inconsistency, thus suggesting that a hyperprior may be beneficial even beyond the cases considered here. We stress that the framework we study is arguably closer to the way Dirichlet process mixtures are used in practice, compared to holding α fixed.

Studying an asymptotic regime where the data-generating truth is a mixture with a finite and fixed number of components entails some degree of model misspecification. Indeed, Dirichlet process mixtures are nonparametric models with an infinite number of components or, in other words, a number of clusters growing with the size of the dataset. Thus, our results can be interpreted as a form of robustness of the prior: if the number of components of the data-generating is finite, it can still be recovered by adapting appropriately the value of α, despite the prior is concentrated on mixtures with infinitely many components. In particular we show that, under the data generation mechanisms we consider, the posterior distribution of α converges to a point mass at 0 at a specific rate, which is crucial to ensure consistency.

2.3.2 Dirichlet process mixtures and random partitions

Henceforth, we will be focusing on Dirichlet process mixture models with a prior on the concentration parameter, namely

$$X_i|\theta_i \overset{\text{ind.}}{\sim} k(\cdot|\theta_i), \quad \theta_i \mid P \overset{\text{iid}}{\sim} P, \quad P \mid \alpha \sim \mathrm{DP}(\alpha, Q_0), \quad \alpha \sim \pi, \tag{2.42}$$

where $k(\,\cdot\,|\theta)$ is some density function, for every θ. Since we are interested in the distribution of the number of clusters, it is reasonable to rewrite (2.42) in terms of the distribution on partitions, related to the so-called Chinese restaurant process. For every pair of natural numbers (n, s) such that $s \leq n$, denote with $\tau_s(n)$ the set of partitions of $\{1, \ldots, n\}$ into s non empty subsets. Conditionally on α, the sequence $(\theta_i)_{i\geq 1}$ induces a prior distribution on the space of partitions of $\mathbb{N}$ that, for every $n \geq 2$, is characterized by

$$\mathbb{P}(A \mid \alpha) = \frac{\alpha^s}{\alpha^{(n)}} \prod_{j=1}^{s} (a_j - 1)!, \quad (A = \{A_1, \ldots, A_s\} \in \tau_s(n), s \leq n), \tag{2.43}$$

where $\alpha^{(n)} = \alpha \cdots (\alpha + n - 1)$ is the ascending factorial and $a_j = |A_j|$ stands for the cardinality of set A_j. Conditionally on the partition A, the probability distributions of the data $X_{1:n} = (X_1, \dots, X_n)$ and of the cluster-specific parameters $\hat{\theta}_{1:s} = (\hat{\theta}_1, \dots, \hat{\theta}_s)$ are

$$\mathbb{P}(X_{1:n} \mid \hat{\theta}_{1:s}, A) = \prod_{j=1}^{s} \prod_{i \in A_j} k(X_i \mid \hat{\theta}_j), \quad \mathbb{P}(\hat{\theta}_{1:s} \mid A, \alpha) = \mathbb{P}(\hat{\theta}_{1:s} \mid A) = \prod_{j=1}^{s} q_0(\hat{\theta}_j), \tag{2.44}$$

where q_0 is the density of Q_0 with respect to the Lebsegue or the counting measure. The number of clusters in a sample of size n is denoted by K_n and under (2.42) its distribution is given by $\mathbb{P}(K_n = s) = \int \sum_{A \in \tau_s(n)} \mathbb{P}(A \mid \alpha) \pi(\mathrm{d}\alpha)$. Since we are concerned with the large sample properties of $\mathbb{P}(K_n = s \mid X_{1:n})$, we focus on the joint distribution of the vector $(X_{1:n}, K_n)$ which, for every $x_{1:n} = (x_1, \dots, x_n) \in \mathbb{X}^n$, is given by

$$\mathbb{P}(X_{1:n} = x_{1:n}, K_n = s) = \sum_{A \in \tau_s(n)} \mathbb{P}(A) \prod_{j=1}^{s} m(x_{A_j}), \tag{2.45}$$

where $\mathbb{P}(A) = \int \mathrm{pr}(A|\alpha)\, \pi(\mathrm{d}\alpha)$ and $m(x_{A_j}) = \int \prod_{i \in A_j} k(x_i \mid \theta) q_0(\theta) \mathrm{d}\theta$ is the marginal likelihood for the subset of observations identified by A_j, given that they are clustered together. We study the asymptotic behaviour of the posterior induced by model (2.42) when the observations are independent and identically distributed samples from a finite mixture, that is we assume the following data generation mechanism

$$X_i \overset{\text{iid}}{\sim} P_* = \sum_{j=1}^{t} p_j R_j, \quad (i = 1, 2, \dots), \tag{2.46}$$

where, for every $t \geq 1$, the R_j's are distinct probability measures on $\mathbb{X}$ and the p_j's are probability weights, i.e. $p_j \in (0,1)$ for every j and $\sum_j p_j = 1$. We will let $P_*^{(n)}$ and $P_*^{(\infty)}$ be the product probability measures induced on $\mathbb{X}^n$ and $\mathbb{X}^\infty$ respectively, and denote (2.46) by $X_{1:\infty} \sim P_*^{(\infty)}$. In the following, we will consider each R_j to be dominated by a suitable measure and denote the resulting density by $f_j(\cdot) := f(\cdot \mid \theta_j^*)$. We say that model in (2.42) is *well-specified* for P_* if $k(\cdot|\theta) = f(\cdot \mid \theta)$, that is if the data-generating distribution is a mixture of kernels belonging to the same parametric family that defines (2.42).

We say that posterior consistency for the number of clusters holds if $\mathbb{P}(K_n = t \mid X_{1:n}) \to 1$ as $n \to \infty$ in $P_*^{(\infty)}$-probability. The conditional probability $\mathbb{P}(K_n = t \mid X_{1:n})$ is defined with respect to the model in (2.42), while the convergence in probability is with respect to the data-generating process $X_{1:\infty} \sim P_*^{(\infty)}$.

2.3.3 Main consistency results

The investigation of the asymptotics of the number of clusters K_n, induced by the model in (2.42), will rely on the following assumptions on the prior π of α

*A*1. *Absolute continuity*: π is absolutely continuous with respect to the Lebesgue measure and its density is still denoted as π;

*A*2. *Polynomial behaviour around the origin*: $\exists\, \epsilon,\ \delta,\ \beta$ such that $\forall \alpha \in (0, \epsilon)$ it holds $\frac{1}{\delta}\alpha^\beta \leq \pi(\alpha) \leq \delta \alpha^\beta$;

$A3$. *Subfactorial moments*: $\exists\, D, \nu, \rho > 0$ such that $\int \alpha^s \pi(\alpha)\, d\alpha < D\rho^{-s}\Gamma(\nu + s + 1)$ for every $s \geq 1$.

The first two assumptions are sufficient to study the posterior moments of α, conditional to the number of groups K_n, as will be clarified in Proposition 10. Assumption $A3$, instead, will be useful specifically for consistency purposes: the minimum value of ρ required to achieve consistency depends on the problem at hand, that is on the specific choice of P in (2.46) and k in (2.42), as will be stated in Theorems 11 and 12. Assumptions $A1$-$A3$ are satisfied by common families of distributions, as displayed in the next lemma.

Lemma 4. *The following choices of π satisfy assumptions $A1$, $A2$ and $A3$ (for a fixed $\rho > 0$)*

(1) *every distribution with bounded support that satisfies assumptions $A1$ and $A2$, such as the uniform distribution over $(0, c)$, with $c > 0$;*

(2) *The Generalized Gamma distribution with density proportional to $\alpha^{d-1}e^{-\left(\frac{\alpha}{a}\right)^p}$, provided that $p > 1$;*

(3) *The Gamma distribution with shape ν and rate ρ.*

The rate parameter of the Gamma distribution corresponds to the quantity ρ in assumption $A3$.

2.3.4 General consistency result for location families with bounded support

For our general result we consider kernels of the form

$$k(x \mid \theta) = g(x - \theta) \quad (x \in \mathbb{R}), \tag{2.47}$$

where $\theta \in \mathbb{R}$ is a location parameter. Here g is a density function on the real line satisfying the following assumptions

$B1$. g is strictly positive on some interval $[a, b]$ and 0 elsewhere;

$B2$. g is differentiable with bounded derivative in (a, b);

$B3$. The base measure Q_0 is absolutely continuous with respect to the Lebesgue measure, and its density q_0 is bounded.

The above assumptions essentially require that the kernel is a location-family distribution with positive density on a bounded support. The class is fairly general and it includes, as relevant special cases, the uniform distribution and the truncated Gaussian distribution, among others.

When considering a mixture of the kernels in (2.47) as data generation mechanism satisfying $B1$–$B3$, with true parameters $\theta^* = (\theta_1^*, \ldots, \theta_t^*)$, we say that θ^* is *completely separated* if $|\theta_j^* - \theta_k^*| > b - a$, for every $j \neq k$. This assumption is somewhat restrictive, but sufficient to prove that the addition of a prior on α may solve the inconsistency issue. Indeed, we have the following general consistency result.

Theorem 10. *Suppose k and q_0 satisfy assumptions $B1$–$B3$. If π satisfies assumptions $A1$–$A3$ with ρ high enough then, for every P_* as in (2.46) with $t \in \{1, 2, \ldots\}$, $f_j = k(\cdot|\theta_j^*)$, θ^* completely separated and θ_j^* belonging to the interior support of Q_0 for every j, we have*

$$\mathbb{P}(K_n = t \mid X_{1:n}) \to 1$$

as $n \to \infty$ in $P_*^{(\infty)}$-probability. On the contrary, if $\pi(\alpha) = \delta_{\alpha^*}(\alpha)$, with $\alpha^* > 0$, then

$$\limsup \mathbb{P}(K_n = t \mid X_{1:n}) < 1$$

as $n \to \infty$ in $P_*^{(\infty)}$-probability.

As discussed above, the minimum value of ρ needed depends on the specific function g and prior distribution Q_0. Therefore, a prior on the concentration parameter yields consistency when the true data generating distribution meets a condition of complete separability, that informally amounts to having cluster locations sufficiently distinct. This condition is automatically satisfied when $t = 1$. We additionally show that, even under such an assumption, the Dirichlet process mixture model with fixed α still fails to be consistent at the number of clusters. Hence, a prior on α is crucial to overcome issues with learning the true number of clusters as the sample size increases.

Moreover, the posterior mass on a smaller number of clusters than the truth vanishes under mild conditions, as explained in the next proposition.

Proposition 8. *Let P_* be as in (2.46), with true parameters $\theta_1^*, \dots, \theta_t^*$. Let θ_j^* belong to the support of Q_0 for every $j = 1, \dots, t$ and let k satisfy assumptions $B1$–$B3$ above or $H1$–$H4$ in the supplementary material. Then*

$$\mathbb{P}(K_n < t \mid X_{1:n}) \to 0 \tag{2.48}$$

in $P_*^{(\infty)}$-probability as $n \to \infty$.

Consistency on specific examples

Theorem 10 requires ρ in assumption $A3$ to be high enough, depending on the specific formulation of the model. In order to provide an example, we focus on the case of uniform kernel and $t = 1$, that is

$$f = \mathrm{Unif}(\theta^* - c, \theta^* + c), \quad k(\cdot|\theta) = \mathrm{Unif}(\theta - c, \theta + c), \quad q_0 = \mathrm{Unif}(\theta^* - c, \theta^* + c), \tag{2.49}$$

where $\theta^* \in \mathbb{R}$ is a fixed location parameter and $c > 0$.

Theorem 11. *Consider f, k and q_0 as in (2.49), and assume π satisfies $A1$–$A3$ (with $\rho \geq 38$). Then*

$$\mathbb{P}(K_n = 1 \mid X_{1:n}) \to 1$$

as $n \to \infty$ in $P_*^{(\infty)}$-probability.

As a second example, we move beyond bounded kernels and consider a simple, yet interesting, case. More precisely, we specialize model (2.42) to Gaussian kernels and assume constant data, equal to some fixed real number θ^*, setting

$$f = \delta_{\theta^*}, \quad k(\cdot|\theta) = \mathrm{N}(\theta, 1), \quad q_0 = \mathrm{N}(0, 1). \tag{2.50}$$

Unlike the other examples, this case is not well-specified, as $k(\cdot|\theta) \neq f(\cdot)$ for every θ. This makes the definition of true or data-generating number of clusters more delicate. Nonetheless, being an example with constant data, one would hope the posterior of the number of clusters to concentrate on one cluster. However, even in such a limiting case, Miller and Harrison (2013) show that under (2.42) with fixed concentration parameter $\mathbb{P}(K_n = 1|X_{1:n})$ does not converge to 1 as n diverges.

Theorem 12. *Consider* (f, k, q_0) *as in* (2.50) *and assume* π *satisfies* $A1$–$A3$ *(with* $\rho > 16$*). Then*

$$\mathbb{P}(K_n = 1 \mid X_{1:n}) \to 1$$

$P_*^{(\infty)}$*-almost surely as* $n \to \infty$*.*

Finally, the previous consistency results are related to another property of general interest, namely the posterior distribution of the concentration parameter converges to a point mass at 0, if posterior consistency for the number of clusters holds.

Proposition 9. *Let the data be generated as in* (2.46) *with* $t \in \mathbb{N}$ *and assume* π *satisfies* $A1$ *and* $A2$*. Then if* $\mathbb{P}(K_n = t \mid X_{1:n}) \to 1$ *we have*

$$\pi(\alpha \mid X_{1:n}) \to \delta_0$$

weakly, as $n \to \infty$*, in* $P_*^{(\infty)}$*-probability.*

This is not surprising since the Dirichlet process mixture model is concentrated on mixtures with infinitely many components and one way to achieve consistency is to let α tend to zero, which entails that the prior is swamped by the data.

2.3.5 Methodology and proof technique

The role of the prior on the concentration parameter

Our proofs of consistency in Theorems 10, 11 and 12 rely on the following lemma.

Lemma 5. *The convergence* $\mathbb{P}(K_n = t \mid X_{1:n}) \to 1$ *as* $n \to \infty$ *in* $P_*^{(\infty)}$*-probability holds true if and only if one has, in* $P_*^{(\infty)}$*-probability,*

$$\sum_{s \neq t} \frac{\mathbb{P}(K_n = s \mid X_{1:n})}{\mathbb{P}(K_n = t \mid X_{1:n})} \to 0 \quad \text{as } n \to \infty\,. \tag{2.51}$$

Working with the ratios of conditional probabilities in (2.51) is beneficial, as the marginal distribution of $X_{1:n}$ involved in the definition of $\mathbb{P}(K_n = t \mid X_{1:n})$ cancels. Also, it is convenient to write such ratios of probabilities as follows: first, recall from (2.43) and (2.45) that

$$\mathbb{P}(X_{1:n} = x_{1:n}, K_n = s) = \int \frac{\alpha^s}{\alpha^{(n)}} \pi(\alpha) \mathrm{d}\alpha \sum_{A \in \tau_s(n)} \prod_{j=1}^{s} (a_j - 1)! m(x_{A_j})$$

for every $s \geq 1$, which implies that

$$\frac{\mathbb{P}(K_n = s \mid X_{1:n})}{\mathbb{P}(K_n = t \mid X_{1:n})} = \underbrace{\frac{\int \frac{\alpha^s}{\alpha^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha}{\int \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha}}_{C(n,t,s)} \underbrace{\frac{\sum_{A \in \tau_s(n)} \prod_{j=1}^{s} (a_j - 1)! \prod_{j=1}^{s} m(X_{A_j})}{\sum_{B \in \tau_t(n)} \prod_{j=1}^{t} (b_j - 1)! \prod_{j=1}^{t} m(X_{B_j})}}_{R(n,t,s)}\,. \tag{2.52}$$

The decomposition of (2.52) into the factors $C(n,t,s)$ and $R(n,t,s)$ is useful to understand the role of the prior distribution over α, and to compare our results with the one of Miller and Harrison (2013, 2014). In particular, the term $R(n,t,s)$ does not depend on α and, hence, on the choice of π. This is indeed the key term studied in Miller and Harrison (2014), where it is shown that, under some assumptions, $\liminf R(n,t,s) > 0$ as $n \to \infty$ in $P_*^{(\infty)}$-probability, for $t < s$. On the contrary, $C(n,t,s)$ incorporates information about α and its prior distribution. In the fixed α case, which can be thought of as having a degenerate prior $\pi = \delta_\alpha$ for some $\alpha > 0$, the term $C(n,t,s)$ boils down to α^{s-t} which is constant with respect to n. This is sufficient for Miller and Harrison (2014) to deduce lack of consistency for fixed α, which means that $\limsup \mathbb{P}(K_n = t \mid X_{1:n}, \alpha) < 1$ as $n \to \infty$ in $P_*^{(\infty)}$-probability for every $\alpha > 0$.

However, once a non-degenerate prior π is employed, $C(n,t,s)$ depends on n and, as we show in the next section, converges to 0 as $n \to \infty$ under mild assumptions on π. Thus, $\liminf R(n,t,s) > 0$ is not anymore sufficient to establish whether consistency holds true or not. Instead, one needs to compare the rate at which $C(n,t,s)$ converges to 0 with the behaviour of $R(n,t,s)$, as done in the following sections. Further lower bounds for $R(n,t,s)$ for general values of s are given in Miller and Harrison (2014); Yang et al. (2019). However, once combined with $C(n,t,s)$, these are too loose to deduce either consistency or lack thereof. Therefore, we need to exploit different techniques to determine the rate of $R(n,t,s)$. Since $\mathbb{P}(K_n = t \mid X_{1:n}) = \int \mathbb{P}(K_n = t \mid X_{1:n}, \alpha)\pi(\alpha \mid X_{1:n})\,\mathrm{d}\alpha$, we deduce $\limsup \mathbb{P}(K_n = t \mid X_{1:n}, \alpha) < 1$ for every $\alpha > 0$. This, however, does not imply that $\limsup \mathrm{pr}(K_n = t \mid X_{1:n}) < 1$, as one first needs to ascertain whether limit and integral can be interchanged. The main reason is that, in the asymptotic regime we are considering, the posterior distribution $\pi(\alpha \mid X_{1:n})$ concentrates around 0 as $n \to \infty$, see Proposition 9 above.

Asymptotic behaviour of the concentration parameter

We are now concerned with studying $C(n,t,s)$ in (2.52). We prove that for priors π satisfying assumptions $A1$–$A3$ $C(n,t,s)$ converges to 0 at a logarithmic rate in n. The asymptotic behaviour of $C(n,t,s)$ is not specific to some kernel k and data generating distribution f and thus can be useful to prove consistency, or lack thereof, for arbitrary Dirichlet process mixture models with random concentration parameter. In order to facilitate the intuition, the term $C(n,t,s)$ can be interpreted as a moment of α, conditional on the n observations being clustered in t groups. Indeed, under (2.42) it holds $\pi(\alpha \mid K_n = t) \propto \frac{\alpha^t}{\alpha^{(n)}}\pi(\alpha)$ and thus $C(n,t,t+s) = \int \alpha^s \pi(\alpha \mid K_n = t)\,d\alpha = E(\alpha^s \mid K_n = t)$. The next proposition shows its asymptotic behaviour.

Proposition 10. *Suppose π satisfies $A1$–$A2$. Then there exist $F, G > 0$ such that for every $0 < s \leq n - t$*

$$F\,\frac{\gamma\{t+s+\beta, \epsilon\log(n)\}}{\{\log(n)+1\}^s} \leq C(n,t,t+s) \leq \frac{Gs}{\epsilon^s}\mathbb{E}[\alpha^{t+s-1}]\frac{\gamma\{t+s+\beta, \epsilon\log(n)\}}{\{\log n/(1+\epsilon)\}^s},$$

where $\gamma(x,y)$ is the lower incomplete Gamma function and $\mathbb{E}[\alpha^s] = \int \alpha^s \pi(\alpha)\,d\alpha$.

Thus, for a fixed s that does not depend on n, $C(n,t,t+s)$ decreases logarithmically as a function of n since $\gamma(x,y) \leq \gamma(x)$ for every x and y. Thus, by looking at the ratios in (2.52), the addition of a prior favours a smaller number of clusters when $n \to \infty$, with s fixed.

The consistency results of the previous section are established by combining Proposition 10 with suitable upper bounds on $R(n,t,s)$ to prove the convergence in (2.51), so that

$$\mathbb{E}\left[\sum_{s=1}^{n-t} \frac{\mathrm{pr}(K_n = t+s \mid X_{1:n})}{\mathrm{pr}(K_n = t \mid X_{1:n})}\right] \leq \frac{1}{\log n} \sum_{s=1}^{n-t} h(s),$$

where $h(s)$ is a function that depends on the specific kernel k and is such that $\limsup \sum_{s=1}^{n} h(s) < \infty$ for every s. The following lemma shows how the problem simplifies in this case, when $t = 1$.

Lemma 6. *Assume $(X_1, X_2, \dots)$ is an exchangeable sequence. Then for every n*

$$\mathbb{E}\left[\sum_{A \in \tau_s(n)} \frac{\prod_{j=1}^{s}(a_j - 1)!}{(n-1)!} \frac{\prod_{j=1}^{s} m(X_{A_j})}{m(X_{1:n})}\right] = \sum_{\boldsymbol{a} \in \mathcal{F}_s(n)} \frac{n}{s! \prod_{j=1}^{s} a_j} E\left\{\frac{\prod_{j=1}^{s} m(X_{A_j^{\boldsymbol{a}}})}{m(X_{1:n})}\right\},$$

where the sum runs over $\mathcal{F}_s(n) = \{\boldsymbol{a} \in \{1, \dots, n\}^s : \sum_{j=1}^{s} a_j = n\}$ and $A^{\boldsymbol{a}}$ is an arbitrary partition in $\tau_s(n)$ such that $|A_j^{\boldsymbol{a}}| = a_j$ for $j = 1, \dots, s$.

2.3.6 Discussion

There are many avenues to extend our results and some of the tools we introduced here may prove useful to accomplish such tasks. First of all, the separability assumption given in Theorem 10 could be relaxed to prove consistency in the setting with a general number of components. The main issue is that $R(n,t,s)$ in (2.52) is harder to study, since it becomes the ratio of sums over the space of partitions: in particular Lemma 6 is not easy to generalize and this explains why the case $t = 1$ is simpler to address. Different mixture kernels present similar difficulties. Summarising, the impact of the prior is fully understood, by Proposition 10, but a more general positive result would require finer bounds on the likelihood component.

Another interesting question is whether consistency can also be attained by estimating the concentration parameter through maximization of the marginal likelihood, in an empirical Bayes fashion (Liu, 1996; McAuliffe et al., 2006). In this paper we preferred to focus on the fully Bayesian approach because it is arguably the one most commonly employed by practitioners. Moreover, the empirical Bayes estimator of α may not be well defined on $(0, \infty)$, thus raising theoretical and practical issues.

It is also worth noticing that our consistency results require the kernel to be perfectly specified: even a small amount of misspecification will probably lead the number of clusters to diverge. Indeed, recovering the true density will require an increasing number of components. This phenomenon has been formally studied in Cai et al. (2021) for finite mixture models, when a prior on the number of components is placed.

The asymptotic analysis of the posterior distribution of the number of clusters for Dirichlet process mixtures has recently attracted considerable theoretical interest (Yang et al., 2019; Ohn and Lin, 2023; Cai et al., 2021), and has motivated various methodological developments (Miller and Harrison, 2018; Zeng and Duan, 2020). Ohn and Lin (2023) showed that, if α is sent deterministically to 0 at appropriate rates as $n \to \infty$, the posterior distribution of the number of clusters concentrates on finite values when data are generated from a finite mixture, which is a necessary condition for consistency. Such results are similar in spirit to ours, although our setting is arguably more natural in a Bayesian framework. Moreover, another interesting extension would be the the case with a growing number of components, rather than fixed: indeed, in this setting a Dirichlet-based model would be a natural choice. We do not pursue this task here, but see Ohn and Lin (2023) for a discussion on asymptotic properties of Bayesian models for mixtures of this type.

A1 Proofs of Section 2.2

Lemma 7. *The transition probabilities $p_{|\mathbf{m}|,|\mathbf{n}|}(t)$ in (2.13) equal $e^{-\lambda_{|\mathbf{m}|}t}$ when $\mathbf{n} = \mathbf{m}$ and*

$$\left(\prod_{h=0}^{|\mathbf{m}-\mathbf{n}|-1} \lambda_{|\mathbf{m}|-h}\right)(-1)^{|\mathbf{m}-\mathbf{n}|} \sum_{k=0}^{|\mathbf{m}-\mathbf{n}|} \frac{e^{-\lambda_{|\mathbf{m}|-k}t}}{\prod_{0\le h\le|\mathbf{m}-\mathbf{n}|, h\neq k}(\lambda_{|\mathbf{m}|-k} - \lambda_{|\mathbf{m}|-h})},$$

when $\mathbf{0} < \mathbf{n} \le \mathbf{m}$, where $\lambda_n = n(\theta + n - 1)/2$.

Proof. See Papaspiliopoulos et al. (2016), Lemma 4.1. □

Proof of Proposition 1

Proof. In this proof we use the same notation of Barrientos et al. (2012) and denote by $G(t)$ the FV-DDP, i.e. $G(t) = X_t$. We also emphasise the elementary event $\omega \in \Omega$ by writing $G(t, \omega)$. By Eq. 3 in Barrientos et al. (2012), it suffices to show that for $\epsilon > 0$, $N \in \mathbb{N}$ and $(t_1, \dots, t_N) \in \mathbb{R}_+^N$ we have

$$\mathbb{P}\Big\{\omega \in \Omega \,:\, [G(t_i, w)(A_0), \dots, G(t_i, w)(A_k)] \in B(\mathbf{s}_{t_i}, \epsilon), i = 1, \dots, N\Big\} > 0. \tag{53}$$

Here:

- $A_0, \dots, A_k$ is a partition of $\mathbb{X}$, with A_i a measurable set with P_0-null boundary;
- $B(\mathbf{s}_{t_i}, \epsilon) = \{(w_0, \dots, w_k) \in \Delta_k \,:\, w_{(t_i,j)} - \epsilon < w_j < w_{(t_i,j)} + \epsilon, j = 0, \dots, k\}$, with $\Delta_k = \{(w_0, \dots, w_k) : w_i \ge 0, i = 0, \dots, k, \sum_{i=0}^k w_i = 1\}$ the k-simplex.
- $\mathbf{s}_{t_i} = (w_{(t_i,0)}, \dots, w_{(t_i,k)}) = (Q_{t_i}(A_0), \dots, Q_{t_i}(A_k)) \in \Delta_k$.
- $Q_{t_i}, , i = 1, \dots, N$ is a probability measure absolutely continuous with respect to Q_0.

As is well known, projecting a Dirichlet process $DP(\alpha)$ on a partition $A_0, \dots, A_k$ yields a k-dimensional Dirichlet density π_α with parameters $(\alpha(A_0), \dots, \alpha(A_k))$. Similarly, projecting a FV process yields a a k-dimensional Wright-Fisher (WF) diffusion, which is reversible and stationary with respect to π_α. Consistently with (2.6), the transition density of the WF is given by:

$$P_t(\mathbf{x}, \mathrm{d}\mathbf{x}') = \sum_{m=0}^{\infty} d_m(t) \sum_{\mathbf{m}\in\mathbb{Z}_+^{k+1}:|\mathbf{m}|=m} \binom{m}{\mathbf{m}} \mathbf{x}^{\mathbf{m}} \pi_{\alpha+\mathbf{m}}(\mathbf{x}')\mathrm{d}\mathbf{x}'.$$

Then we can rewrite (53) as:

$$\int_{B(\mathbf{s}_{t_1},\epsilon)} \cdots \int_{B(\mathbf{s}_{t_N},\epsilon)} \pi_\alpha(\mathbf{x}_1) P_{t_2-t_1}(\mathbf{x}_1, \mathbf{x}_2) \dots P_{t_N-t_{N-1}}(\mathbf{x}_{N-1}, \mathbf{x}_N)\, \mathrm{d}\mathbf{x}_1 \dots \mathrm{d}\mathbf{x}_N$$

Since $B(\mathbf{s}_{t_1}, \epsilon)$ has strictly positive Lebsegue measure, we just need to show that the integrand is strictly bigger than 0 for any $(\mathbf{x}_1, \dots, \mathbf{x}_N) \in B(\mathbf{s}_{t_1}, \epsilon) \times \cdots \times B(\mathbf{s}_{t_N}, \epsilon)$. Clearly $\pi_\alpha(\mathbf{x}_1) > 0$ for any $\mathbf{x}_1 \in B(\mathbf{s}_{t_1}, \epsilon)$. For what concerns $1 < j \le N$, we have:

$$P_{t_j-t_{j-1}}(\mathbf{x}_{j-1}, \mathbf{x}_j) \ge d_0(t_j - t_{j-1})\pi_\alpha(\mathbf{x}_j) > 0, \quad \forall \mathbf{x}_j \in B(\mathbf{s}_{t_j}, \epsilon),$$

which completes the proof. □

Proof of Proposition 2

Conditioning on the random measure P_{T+t} at time $T+t$ yields

$$\begin{aligned}
&\mathbb{P}\Big(X_{T+t}^{k+1} \in A \mid \mathbf{X}_{0:T}, X_{T+t}^{1:k}\Big) \\
&= \mathbb{E}\Big[\mathbb{P}\Big(X_{T+t}^{k+1} \in A \mid P_{T+t}, \mathbf{X}_{0:T}, X_{T+t}^{1:k}\Big) \mid \mathbf{X}_{0:T}, X_{T+t}^{1:k}\Big] \\
&= \mathbb{E}\Big[\mathbb{P}\Big(X_{T+t}^{k+1} \in A \mid P_{T+t}\Big) \mid \mathbf{X}_{0:T}, X_{T+t}^{1:k}\Big] \\
&= \mathbb{E}\Big[P_{T+t}(A) \mid \mathbf{X}_{0:T}, X_{T+t}^{1:k}\Big]
\end{aligned} \tag{54}$$

where the second equality follows from the conditional independence of the observations given the signal; cf. (2.7). From (2.8), eq. (3.7) in Papaspiliopoulos et al. (2016) implies that $P_{T+t} \mid \mathbf{X}_0, \dots, \mathbf{X}_T$ is the mixture of Dirichlet processes

$$\sum_{\mathbf{n}\in L(\mathbf{M})} p_t\left(\mathbf{M}, \mathbf{n}\right) DP\left(\alpha + \sum_{i=1}^{K} n_i \delta_{x_i^*}\right).$$

By linearity of the expectation and the predictive of the Dirichlet process, when $k=0$ the RHS of (57) reads

$$\begin{aligned}
&\sum_{\mathbf{n}\in L(\mathbf{M})} p_t\left(\mathbf{M}, \mathbf{n}\right) \mathbb{E}\left[DP\left(\alpha + \sum_{i=1}^{K} n_i \delta_{x_i^*}\right)(A)\right] = \\
&= \sum_{\mathbf{n}\in L(\mathbf{M})} p_t\left(\mathbf{M}, \mathbf{n}\right)\left[\frac{\theta}{\theta+|\mathbf{n}|} Q_0\left(A\right) + \frac{|\mathbf{n}|}{\theta+|\mathbf{n}|}\sum_{i=1}^{K} n_i \delta_{x_i^*}\left(A\right)\right] \\
&= \sum_{\mathbf{n}\in L(\mathbf{M})} p_t(\mathbf{M}, \mathbf{n}) \frac{\theta}{\theta+|\mathbf{n}|} Q_0\left(A\right) + \sum_{\mathbf{n}\in L(\mathbf{M})} p_t\left(\mathbf{M}, \mathbf{n}\right)\frac{|\mathbf{n}|}{\theta+|\mathbf{n}|} P_{\mathbf{n}},
\end{aligned}$$

which is (2.10) with $k=0$. When $k>0$, by the conjugacy property of mixture of Dirichlet processes the RHS of (57) reads

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\mathbf{n}\in L(m)} p_t(\mathbf{M}, \mathbf{n}) \Pi_{\alpha+\sum_{i=1}^{K} n_i \delta_{x_i^*}} \middle| X_{T+t}^{1:k}\right] \\
&= \sum_{\mathbf{n}\in L(\mathbf{M})} p_t^{(k)}(\mathbf{M}, \mathbf{n}) \mathbb{E}\left[\Pi_{\alpha+\sum_{i=1}^{K} n_i \delta_{x_i^*} + \sum_{j=1}^{k}\delta_{x_j}}\right]
\end{aligned}$$

yielding (2.11).

Proof of Proposition 3

Proof. Denote:

$$\mathbb{E}_0[X] = \int x\, Q_0(\mathrm{d}x)$$
$$\mathbb{E}_0[X^2] = \int x^2\, Q_0(\mathrm{d}x)$$

Then we want to compute:

$$\mathrm{Corr}(X_t, X_{t+s}) = \frac{\mathrm{Cov}(X_t, X_{t+s})}{\mathbb{E}_0[X^2] - \mathbb{E}_0^2[X]} = \frac{\mathbb{E}[X_t X_{t+s}] - \mathbb{E}_0^2[X]}{\mathbb{E}_0[X^2] - \mathbb{E}_0^2[X]}$$

The only object left to compute is

$$\mathbb{E}[X_t X_{t+s}] = \int x_t x_{t+s}\, Q(\mathrm{d}x_t, \mathrm{d}x_{t+s})$$

Note that from Proposition 2 we can write the joint distribution using the chain rule::

$$Q(\mathrm{d}x_t, \mathrm{d}x_{t+s}) = Q_0(\mathrm{d}x_t)\left[\left(1 - e^{-\frac{\theta}{2}s}\right) Q_0(\mathrm{d}x_{t+s}) + \frac{\theta e^{-\frac{\theta}{2}s}}{\theta+1} Q_0(\mathrm{d}x_{t+s}) + \frac{e^{-\frac{\theta}{2}s}}{\theta+1}\delta_{x_t}(\mathrm{d}x_{t+s})\right]$$

so we get:

$$\mathbb{E}[X_t X_{t+s}] = \left(1 - e^{-\frac{\theta}{2}s} + \frac{\theta e^{-\frac{\theta}{2}s}}{\theta+1}\right)\mathbb{E}_0^2[X] + \frac{e^{-\frac{\theta}{2}s}}{\theta+1}\mathbb{E}_0[X^2]$$

Consequently:

$$\mathrm{Cov}(X_t, X_{t+s}) = \frac{e^{-\frac{\theta}{2}s}}{\theta+1}\left(\mathbb{E}_0[X^2] - \mathbb{E}_0^2[X]\right)$$

from which the result follows. □

Proof of Proposition 4

Proof. Denote by $P_{0,k}$ the predictive distribution of the Dirichlet process. We have to prove that

$$\left|\mathbb{P}\left(X_{T+t}^{k+1} \in A \mid \mathbf{X}_{0:T}, X_{T+t}^{1:k}\right) - P_{0,k}(A)\right| \to 0, \tag{55}$$

as $t \to \infty$. Using the triangle inequality the LHS of (55) is smaller than:

$$\left|\sum_{\mathbf{n}\in L(\mathbf{M})} p_t^{(k)}(\mathbf{M},\mathbf{n})\frac{-|\mathbf{n}|}{\theta+|\mathbf{n}|+k}P_{0,k}(A)\right| + \left|\sum_{\mathbf{n}\in L(\mathbf{M})} p_t^{(k)}(\mathbf{M},\mathbf{n})\frac{|\mathbf{n}|}{\theta+|\mathbf{n}|+k}P_{\mathbf{n}}(A)\right|$$

Note now that the time-dependence of (2.11) is ultimately due to $p_{|\mathbf{m}|,|\mathbf{n}|}(t)$ in (2.13). These are the transition probabilities of a one-dimensional death process on $\mathbb{Z}_+$ which jumps from m to $m-1$ at infinitesimal rate $\lambda_m = m(\theta+m-1)/2$. It can be easily verified that, as $t \to \infty$, we have $p_{|\mathbf{m}|,0}(t) \to 1$ for any $\mathbf{m}$ and $p_{|\mathbf{m}|,|\mathbf{n}|}(t) \to 0$ for any $\mathbf{0} < \mathbf{n} \leq \mathbf{m}$, and similar statement holds for (3.6). Then, denoting B_1, B_2 the two sums in the previous display respectively, we have

$$0 \leq \max\{B_1, B_2\} \leq \sum_{\mathbf{n} \in L(\mathbf{M})} p_t^{(k)}(\mathbf{M}, \mathbf{n}) \frac{|\mathbf{n}|}{\theta + |\mathbf{n}| + k} \to 0$$

which implies (55), as desired. □

Proof of Proposition 5

Proof. By de Finetti's Representation Theorem $P_k \to P^*$ as $k \to \infty$, with P^* being the de Finetti measure of the sequence $\left(X_{T+t}^k\right)_{k \geq 1}$. Moreover, recalling that $L(\mathbf{M})$ is a finite set, we have:

$$\lim_{k \to \infty} \sum_{\mathbf{n} \in L(\mathbf{M})} p_t^{(k)}(\mathbf{M}, \mathbf{n}) \frac{k}{\theta + |\mathbf{n}| + k} = \lim_{k \to \infty} \sum_{\mathbf{n} \in L(\mathbf{M})} p_t^{(k)}(\mathbf{M}, \mathbf{n}) = 1$$

As regards the other two components of (2.11) we have

$$0 \leq \lim_{k \to \infty} \sum_{\mathbf{n} \in L(\mathbf{M})} p_t^{(k)}(\mathbf{M}, \mathbf{n}) \frac{1}{\theta + |\mathbf{n}| + k} \leq \lim_{k \to \infty} \sum_{\mathbf{n} \in L(\mathbf{M})} \frac{1}{\theta + |\mathbf{n}| + k} = 0$$

and we have the result. □

A2 Proofs of Section 2.2.5

Proof of Lemma 2

Proof. Using (2.26), we have

$$\begin{aligned} p(\mathbf{n}_{i+1} | \mathbf{p}_{t_i}) = & \int p(\mathbf{n}_{i+1} | \mathbf{p}_{t_{i+1}}) q_{t_{i+1} - t_i}(\mathbf{p}_{t_{i+1}} | \mathbf{p}_{t_i}) \mathrm{d}\mathbf{p}_{t_{i+1}} \\ = & m(\mathbf{n}_{i+1}) \int h(\mathbf{p}_{t_{i+1}}, \mathbf{n}_{i+1}) q_{t_{i+1} - t_i}(\mathbf{p}_{t_{i+1}} | \mathbf{p}_{t_i}) \mathrm{d}\mathbf{p}_{t_{i+1}} \\ = & m(\mathbf{n}_{i+1}) \mathbb{E}[h(\mathbf{P}_{t_{i+1}}, \mathbf{n}_{i+1}) | \mathbf{P}_{t_i} = \mathbf{p}_{t_i}], \end{aligned}$$

where the integral is over the $(K-1)$-dimensional simplex, from which (2.29) leads to the result. □

Proof of Lemma 3

Proof. Denote by $\mathbf{n}_{i-1}$ and $\mathbf{n}_{i+1}$ the multiplicities of types for observations sampled at times t_{i-1} and t_{i+1} respectively in the setting of (2.24). Then Papaspiliopoulos and Ruggiero (2014) showed that

$$\mathcal{F}_{t_i - t_{i-1}}(\pi_{\boldsymbol{\alpha} + \mathbf{n}_{i-1}})(\mathbf{p}_{t_i}) = \sum_{\mathbf{0} \leq \mathbf{k} \leq \mathbf{n}_{i-1}} p_{\mathbf{n}_{i-1}, \mathbf{k}}(t_{i+1} - t_i) \pi_{\boldsymbol{\alpha} + \mathbf{k}}(\mathbf{p}_{t_i}). \tag{56}$$

Furthermore, using (2.19) first and then Lemma 2 we have

$$\begin{aligned}\mathcal{B}_{t_{i+1}-t_i}(\pi_{\boldsymbol{\alpha}+\mathbf{n}_{i+1}})(\mathbf{p}_{t_i}) = p(\mathbf{p}_{t_i}|\mathbf{n}_{i+1}) &= \frac{p(\mathbf{p}_{t_i})p(\mathbf{n}_{i+1}|\mathbf{p}_{t_i})}{m(\mathbf{n}_{i+1})}\\ &= p(\mathbf{p}_{t_i}) \sum_{\mathbf{0}\leq\mathbf{k}\leq\mathbf{n}_{i+1}} p_{\mathbf{n}_{i+1},\mathbf{k}}(t_{i+1}-t_i)h(\mathbf{p}_{t_i},\mathbf{k})\\ &= \sum_{\mathbf{0}\leq\mathbf{k}\leq\mathbf{n}_{i+1}} p_{\mathbf{n}_{i+1},\mathbf{k}}(t_{i+1}-t_i)h(\mathbf{p}_{t_i},\mathbf{k})p(\mathbf{p}_{t_i})\\ &= \sum_{\mathbf{0}\leq\mathbf{k}\leq\mathbf{n}_{i+1}} p_{\mathbf{n}_{i+1},\mathbf{k}}(t_{i+1}-t_i)\pi_{\boldsymbol{\alpha}+\mathbf{k}}(\mathbf{p}_{t_i})\end{aligned}$$

where the last identity follows from from (2.26) and (2.19). By equating $\mathbf{n}_{i-1}$ with $\mathbf{n}_{i+1}$ and $t_i - t_{i-1}$ with $t_{i+1}-t_i$, one can now see that $\mathcal{B}_t(\pi_{\boldsymbol{\alpha}+\mathbf{n}}) = \mathcal{F}_t(\pi_{\boldsymbol{\alpha}+\mathbf{n}})$, with $\mathcal{B}_t$ as in (2.21). The fact that $\mathcal{F}_t(DP\left(\alpha + \sum_{j=1}^{K} n_j\delta_{x_j^*}\right))$ equals the right hand side of (2.31) now follows from Theorem 3.1 in Papaspiliopoulos et al. (2016), and the same proof can be used to show (2.31), by seeing $\mathcal{B}_t(\pi_{\boldsymbol{\alpha}+\mathbf{n}})$ as the projection of $\mathcal{B}_t(DP\left(\alpha + \sum_{j=1}^{K} n_j\delta_{x_j^*}\right))$ onto an arbitrary partition, from which the first statement also follows. □

Proof of Theorem 9

Proof. Without loss of generality, let $i = 1$ and denote $P_1 = P_{t_1}$. Given a measurable partition $\mathcal{A} = (A_1,\dots,A_m)$ of $\mathbb{X}$, let $P_1(\mathcal{A}) := (P_1(A_1),\dots,P_1(A_m))$ and denote by $\mathbf{X}(\mathcal{A})$ the list of labels derived from binning $\mathbf{X}$ into $\mathcal{A}$, i.e., whose i-th element is j if $X_i \in A_j$. Further, let $\{\mathcal{B}_n, n\geq 1\} = \{(B_1^n,\dots,B_n^n), n \geq 1\}$ be a sequence of increasingly finer partitions of $\mathbb{X}$ such that $\mathcal{B}_n$ is finer than $\mathcal{A}$, for every n, and such that $\max_j \operatorname{diam}(B_j^n) \to 0$ as n diverges. Since $\mathcal{B}_n$ is increasingly finer, we have that $\left(\mathbb{E}\left[f(P_1(\mathcal{A}))|\mathbf{X}_0(\mathcal{B}_n),\mathbf{X}_1(\mathcal{B}_n),\mathbf{X}_2(\mathcal{B}_n)\right]\right)_n$ is a martingale for every bounded and continuous function f (see Proposition V.2.7 in Cinlar (2011)). Thus, by the martingale convergence theorem we have that $P_1(\mathcal{A})|\mathbf{X}_0(\mathcal{B}_n),\mathbf{X}_1(\mathcal{B}_n),\mathbf{X}_2(\mathcal{B}_n)$ converges weakly to $P_1(\mathcal{A})|\mathbf{X}_0,\mathbf{X}_1,\mathbf{X}_2$ as $n\to\infty$. The left hand side of the previous expression can be characterized, by virtue of de Finetti's Theorem, in terms of the predictive distributions of $X_1^{1:k}(\mathcal{A})|\mathbf{X}_0(\mathcal{B}_n),\mathbf{X}_1(\mathcal{B}_n),\mathbf{X}_2(\mathcal{B}_n)$ for arbitrary k, where $X_1^{1:k}(\mathcal{A})$ denotes k samples from $P_1(\mathcal{A})$. Without loss of generality, let now n be large enough that different observations lie in different sets of $\mathcal{B}_n$, and write, for brevity, $P_{1,n} := P_1(\mathcal{B}_n)$ and $\mathbf{X}_{i,n} := \mathbf{X}_i(\mathcal{B}_n)$. Let also $p(x^{1:k}|\mathbf{X}_{0,n},\mathbf{X}_{1,n},\mathbf{X}_{2,n})$ be the density of the vector $X_1^{1:k}(\mathcal{A})$ evaluated at $x^{1:k}$, conditional on the binned observation $\mathbf{X}_{0,n},\mathbf{X}_{1,n},\mathbf{X}_{2,n}$. Then we have

$$\begin{aligned}p(x^{1:k}|\mathbf{X}_{0,n},\mathbf{X}_{1,n},\mathbf{X}_{2,n}) \propto p(x^{1:k},\mathbf{X}_{0,n},\mathbf{X}_{1,n},\mathbf{X}_{2,n}) &= \mathbb{E}\left[p(x^{1:k},\mathbf{X}_{0,n},\mathbf{X}_{1,n},\mathbf{X}_{2,n}|P_{1,n})\right]\\ &= \mathbb{E}\left[p(x^{1:k}|P_{1,n})p(\mathbf{X}_{0,n}|P_{1,n})p(\mathbf{X}_{1,n}|P_{1,n})p(\mathbf{X}_{2,n}|P_{1,n})\right]\end{aligned} \tag{57}$$

where in the last identity we have used the conditional independence of the observations given the signal state (cf. (2.7)). By Lemma 2 and the subsequent comment we get

$$
\begin{aligned}
p(\mathbf{X}_{0,n}|P_{1,n}) &\propto \sum_{\mathbf{0}\leq\mathbf{k}_{i-1}\leq\mathbf{n}_{i-1}} p_{\mathbf{n}_{i-1},\mathbf{k}_{i-1}} h(P_{1,n},\mathbf{k}_{i-1})\\
p(\mathbf{X}_{2,n}|P_{1,n}) &\propto \sum_{\mathbf{0}\leq\mathbf{k}_{i+1}\leq\mathbf{n}_{i+1}} p'_{\mathbf{n}_{i+1},\mathbf{k}_{i+1}} h(P_{1,n},\mathbf{k}_{i+1})\\
p(\mathbf{X}_{1,n}|P_{1,n}) &\propto h(P_{1,n},\mathbf{n}_{i})
\end{aligned}
$$

where $p_{\mathbf{n}_{i-1},\mathbf{k}_{i-1}} := p_{\mathbf{n}_{i-1},\mathbf{k}_{i-1}}(\Delta_i)$ and $p'_{\mathbf{n}_{i+1},\mathbf{k}_{i+1}} := p_{\mathbf{n}_{i+1},\mathbf{k}_{i+1}}(\Delta_{i+1})$. By linearity and by (2.27), (57) is thus proportional to

$$
\begin{aligned}
\sum_{\mathbf{0}\leq\mathbf{k}_{i-1}\leq\mathbf{n}_{i-1}} \sum_{\mathbf{0}\leq\mathbf{k}_{i+1}\leq\mathbf{n}_{i+1}} & p_{\mathbf{n},\mathbf{k}_{i-1}} p'_{\mathbf{m},\mathbf{k}_{i+1}} \frac{m^{(n)}(\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1})}{m^{(n)}(\mathbf{k}_{i-1})m^{(n)}(\mathbf{n}_i)m^{(n)}(\mathbf{k}_{i+1})}\\
&\times \mathbb{E}\left[p(x^{1:k}|P_{1,n})h(P_{1,n},\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1})\right],
\end{aligned}
$$

where $m^{(n)}$ denotes the marginal distribution in (2.25) relative to the model induced by the partition $\mathcal{B}_n$. Moreover, using (2.26) and (2.19) it can be seen that

$$
\mathbb{E}\Big[p(\mathbf{n}'|\mathbf{p})h(\mathbf{p},\mathbf{n})\Big] = \int p(\mathbf{n}'|\mathbf{p})\frac{p(\mathbf{n}|\mathbf{p})}{m(\mathbf{n})}p(\mathbf{p})\mathrm{d}\mathbf{p} = \frac{m(\mathbf{n},\mathbf{n}')}{m(\mathbf{n})} = m_{\mathbf{n}}(\mathbf{n}'),
$$

with $m_{\mathbf{n}}(\mathbf{n}') := p(\mathbf{n}'|\mathbf{n})$, hence we can write

$$
\mathbb{E}\left[p(x^{1:k}|P_{1,n})h(P_{1,n},\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1})\right] = m_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}(x^{1:k}).
$$

Note that the above identity holds since $\mathcal{B}_n$ is finer than $\mathcal{A}$. Hence the left hand side of (57) equals

$$
\begin{aligned}
\sum_{\mathbf{0}\leq\mathbf{k}_{i-1}\leq\mathbf{n}_{i-1}} \sum_{\mathbf{0}\leq\mathbf{k}_{i+1}\leq\mathbf{n}_{i+1}} & p_{\mathbf{n}_{i-1},\mathbf{k}_{i-1}} p'_{\mathbf{n}_{i+1},\mathbf{k}_{i+1}}\\
&\times \frac{m^{(n)}(\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1})}{C_n m^{(n)}(\mathbf{k}_{i-1})m^{(n)}(\mathbf{n}_i)m^{(n)}(\mathbf{k}_{i+1})} m_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}(\mathbf{n}')
\end{aligned}
$$

where C_n is a normalizing constant and $\mathbf{n}'$ is the vector of multiplicities associated to $x^{1:k}$. Since $m_{\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1}}(\mathbf{n}')$ is the distribution induced by the Pólya Urn scheme of the Dirichlet–multinomial model, it follows that the law of $X_1^{1:k}(\mathcal{A})|\mathbf{X}_{0,n},\mathbf{X}_{1,n},\mathbf{X}_{2,n}$ is exchangeable. Note in particular that this marginal distribution does not depend on the partition $\mathcal{B}_n$, since $\mathcal{A}$ is a coarser partition. Given the arbitrariness of k, we can appeal to de Finetti's Theorem to conclude that the law of $P_1(\mathcal{A})|\mathbf{X}_{0,n},\mathbf{X}_{1,n},\mathbf{X}_{2,n}$ is given by

$$
\begin{aligned}
\sum_{\mathbf{0}\leq\mathbf{k}_{i-1}\leq\mathbf{n}_{i-1}} \sum_{\mathbf{0}\leq\mathbf{k}_{i+1}\leq\mathbf{n}_{i+1}} & p_{\mathbf{n}_{i-1},\mathbf{k}_{i-1}} p'_{\mathbf{n}_{i+1},\mathbf{k}_{i+1}}\\
&\times \frac{m^{(n)}(\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1})}{C_n m^{(n)}(\mathbf{k}_{i-1})m^{(n)}(\mathbf{n}_i)m^{(n)}(\mathbf{k}_{i+1})} DP\left(\alpha(\mathcal{A})+\mathbf{k}_{i-1}(\mathcal{A})+\mathbf{n}_i(\mathcal{A})+\mathbf{k}_{i+1}(\mathcal{A})\right)
\end{aligned} \tag{58}
$$

where $\alpha(\mathcal{A}) = (\alpha(A_1),\dots,\alpha(A_m))$ and $\mathbf{k}_{i-1}(\mathcal{A}),\mathbf{n}_i(\mathcal{A}),\mathbf{k}_{i+1}(\mathcal{A})$ denote the multiplicities projected onto $\mathcal{A}$. The limit as $n\to\infty$ can now be computed by virtue of the martingale convergence theorem. The proof is completed by observing that the limiting weights do not depend on the partition $\mathcal{B}_n$ and the previous display coincides with the projection onto the partition $\mathcal{A}$ of a finite mixture of laws of Dirichlet processes. □

Proof of Proposition 6

Proof. Statement A follows from the fact that ultimately the limit partition sets with positive multiplicities will be those coinciding with the support points of Q_0.

Assume now, without loss of generality, that the partition $\mathcal{B}_n$ is such that the first observation lies in B_1^n, the second in B_2^n and so on. The density of a vector $\mathbf{k}$ of multiplicities is in this case determined by the Blackwell–MacQueen Pólya urn scheme to be

$$m^{(n)}(\mathbf{k}) = \frac{\prod_{j=1}^{K}\prod_{h=0}^{k_j-1}\Big(\theta Q_0(B_j^n)+h\Big)}{\theta^{(|\mathbf{k}|)}},$$

with the convention that $\prod_{h=0}^{-1}=1$. Denoting $m_{\mathbf{n}}(\mathbf{n}') := p(\mathbf{n}'|\mathbf{n})$, as in the proof of Theorem 9, it follows that

$$\begin{aligned}\frac{m^{(n)}(\mathbf{k}_{i-1}+\mathbf{n}_i+\mathbf{k}_{i+1})}{m^{(n)}(\mathbf{k}_{i-1})m^{(n)}(\mathbf{n}_i)m^{(n)}(\mathbf{k}_{i+1})} &= \frac{m^{(n)}_{\mathbf{n}_i}(\mathbf{k}_{i-1}+\mathbf{k}_{i+1})}{m^{(n)}(\mathbf{k}_{i-1})m^{(n)}(\mathbf{k}_{i+1})}\\ = \frac{\theta^{(|\mathbf{k}_{i-1}|)}\theta^{(|\mathbf{k}_{i+1}|)}}{(\theta+|\mathbf{n}_i|)^{(|\mathbf{k}_{i-1}|+|\mathbf{k}_{i+1}|)}}&\prod_{j=1}^{K}\frac{\prod_{h=0}^{k_{i-1,j}+k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+n_{i,j}+h\Big)}{\prod_{h=0}^{k_{i-1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)\prod_{h=0}^{k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)}.\end{aligned}$$

If $\mathcal{D}=\emptyset$, since no values are shared across times, we have that, for every j, at most one between $k_{i-1,j}$ and $k_{i+1,j}$ is non zero, and in such case we have $n_{i,j}=0$. Then, if $k_{i-1,j}>0$ we have $k_{i+1,j}=0$ and $n_{i,j}=0$, so

$$\frac{\prod_{h=0}^{k_{i-1,j}+k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+n_{i,j}+h\Big)}{\prod_{h=0}^{k_{i-1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)\prod_{h=0}^{k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)}=1,$$

and the same happens when $k_{i+1,j}>0$, $k_{i-1,j}=0$ and $n_{i,j}=0$. This leads to statement B.

If $\mathcal{D}\neq\emptyset$, since some values are shared across times, there exists a j such that one of the following is true: (i) $n_{i,j}>0$ and $k_{i-1,j}>0$; (ii) $n_{i,j}>0$ and $k_{i+1,j}>0$; (iii) $k_{i-1,j}>0$ and $k_{i+1,j}>0$. In case (i)

$$\frac{\prod_{h=0}^{k_{i-1,j}+k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+n_{i,j}+h\Big)}{\prod_{h=0}^{k_{i-1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)\prod_{h=0}^{k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)}\to\infty$$

as $n\to\infty$, since $Q_0(B_j^n)\to 0$ and the denominator vanishes. Case (ii) is obtained similarly. In case (iii), rewrite the weights as

$$\begin{aligned}&\frac{\prod_{h=0}^{k_{i-1,j}+k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+n_{i,j}+h\Big)}{\prod_{h=0}^{k_{i-1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)\prod_{h=0}^{k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)}\\&=\frac{\prod_{h=0}^{k_{i-1,j}-1}\Big(\theta Q_0(B_j^n)+n_{i,j}+h\Big)}{\prod_{h=0}^{k_{i-1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)}\frac{\prod_{h=k_{i-1,j}}^{k_{i-1,j}+k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+n_{i,j}+h\Big)}{\prod_{h=0}^{k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)}.\end{aligned}\tag{59}$$

Here the left factor is such that

$$\frac{\prod_{h=0}^{k_{i-1,j}-1}\Big(\theta Q_0(B_j^n)+n_{i,j}+h\Big)}{\prod_{h=0}^{k_{i-1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)}\geq 1$$

and the right factor can be written

$$\frac{\prod_{h=k_{i-1,j}}^{k_{i-1,j}+k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+n_{i,j}+h\Big)}{\prod_{h=0}^{k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)}=\frac{\prod_{h=0}^{k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+k_{i-1,j}+n_{i,j}+h\Big)}{\prod_{h=0}^{k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)}.$$

Therefore, the left hand side of (59) is greater than or equal to

$$\frac{\prod_{h=0}^{k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+k_{i-1,j}+n_{i,j}+h\Big)}{\prod_{h=0}^{k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)}$$

which diverges to infinity as $n\to\infty$ as well. Thus, nodes with shared observations have divergent unnormalized weights. Let $\mathcal{S}=D_{i-1}\cup D_{i+1}$ be the set of shared values and let $(\mathbf{k}_{i-1},\mathbf{k}_{i+1})\in\mathcal{D}$. Then, we can write the associated weight as

$$\begin{aligned}&\frac{\theta^{(|\mathbf{k}_{i-1}|)}\theta^{(|\mathbf{k}_{i+1}|)}}{(\theta+|\mathbf{n}_i|)^{(|\mathbf{k}_{i-1}|+|\mathbf{k}_{i+1}|)}}\\&\quad\times\prod_{j=1}^{K}\frac{\prod_{h=0}^{k_{i-1,j}+n_{i,j}+k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)}{\prod_{h=0}^{k_{i-1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)\prod_{h=0}^{n_{i,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)\prod_{h=0}^{k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)}\\&=\frac{\theta^{(|\mathbf{k}_{i-1}|)}\theta^{(|\mathbf{k}_{i+1}|)}}{(\theta+|\mathbf{n}_i|)^{(|\mathbf{k}_{i-1}|+|\mathbf{k}_{i+1}|)}}\\&\quad\times\prod_{j\in\mathcal{S}}\frac{\prod_{h=1}^{k_{i-1,j}+n_{i,j}+k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)}{\prod_{h=1}^{k_{i-1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)\prod_{h=1}^{n_{i,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)\prod_{h=1}^{k_{i+1,j}-1}\Big(\theta Q_0(B_j^n)+h\Big)}\\&\quad\times\frac{1}{\prod_{j\in\mathcal{S}}Q_0(B_j^n)\prod_{h=0}^{n_{i,j}-1}Q_0(B_j^n)}.\end{aligned}$$

Here the third factor on the right hand side is common to each node and is cancelled upon normalizing, while the second factor converges to the product in statement C, proving the result. □

Proof of Proposition 7

Proof. The first statement follows as in the proof of Theorem 9 by noting that conditioning on $\mathbf{X}_{i-1}, \mathbf{X}_{i+1}$ in (2.32) is qualitatively analogous to conditioning to $\mathbf{X}_{t_{0:i-1}}, \mathbf{X}_{t_{i+1,T}}$ in (2.36)-(2.37), since the main argument (57) is based on the factorization of the likelihoods of the data collected prior, concurrently and after the signal state. The second statement follows by the linearity of the expected value in (57) and by readjusting the weights. □

Proof of Corollary 2

Proof. The statement can be easily proved by noting that

$$\mathbb{P}(X_i \in A|\mathbf{X}_{0:T}) = \mathbb{E}\left[\mathbb{P}(X_i \in A|P_i, \mathbf{X}_{0:T})|\mathbf{X}_{0:T}\right] = \mathbb{E}\left[P_i(A)|\mathbf{X}_{0:T}\right]$$

and using (2.38), after recalling that if $W \sim DP(\alpha)$ then $\mathbb{E}(W) = \alpha/\alpha(\mathbb{X})$. □

A3 Proofs of Section 2.3

Proof of Lemma 5

Proof. The result immediately follows upon noting that

$$\mathbb{P}(K_n = t \mid X_{1:n}) = \left\{1 + \sum_{s \neq t} \frac{\mathbb{P}(K_n = s \mid X_{1:n})}{\mathbb{P}(K_n = t \mid X_{1:n})}\right\}^{-1}.$$

□ □

Proof of Proposition 10

By assumptions $A1$ and $A2$ there exist $\epsilon, \delta, \beta > 0$ such that

$$\frac{1}{\delta^2} \frac{\int_0^\epsilon \frac{\alpha^{t+s+\beta}}{\alpha^{(n)}} \, \mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^{t+\beta}}{\alpha^{(n)}} \, \mathrm{d}\alpha} \leq \frac{\int_0^\epsilon \frac{\alpha^{t+s}}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^{t}}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha} \leq \delta^2 \frac{\int_0^\epsilon \frac{\alpha^{t+s+\beta}}{\alpha^{(n)}} \, \mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^{t+\beta}}{\alpha^{(n)}} \, \mathrm{d}\alpha}. \tag{60}$$

Notice that, if assumption $A2$ holds for $\epsilon \geq 1$, it holds also for $\epsilon < 1$. Thus, without loss of generality, we will assume $\epsilon < 1$ and the main object of interest will be

$$E_n[\alpha^s] = \int_0^\epsilon \alpha^s p_n(\alpha) \mathrm{d}\alpha,$$

where E_n denotes the expected value with respect to the probability distribution with density

$$p_n(\alpha) = \frac{f_n(\alpha)}{\int_0^\epsilon f_n(x) \, \mathrm{d}x}, \quad f_n(x) = \frac{x^{t+\beta}}{x^{(n)}} \, \mathbb{1}_{(0,\epsilon)}(x), \tag{61}$$

where $\mathbb{1}_A$ stands for the indicator function of set A. We now provide three lemmas that will be useful to prove Proposition 10.

Lemma 8. *Let f and g be two pdf's on $\mathbb{R}$ such that $g(x)/f(x)$ is non-decreasing in x. Then $\int h(x)f(x)\mathrm{d}x \leq \int h(x)g(x)\mathrm{d}x$ for every non-decreasing $h : \mathbb{R} \to \mathbb{R}$.*

Proof. Let $X \sim f$ and $Y \sim g$. Since $g(x)/f(x)$ is non-decreasing we have $g(x_0)f(x_1) \leq g(x_1)f(x_0)$ for every $x_0 < x_1$. Thus we have

$$F_Y(x_1)f(x_1) = \int_{-\infty}^{x_1} g(x_0)f(x_1)\mathrm{d}x_0 \leq \int_{-\infty}^{x_1} g(x_1)f(x_0)\mathrm{d}x_0 = F_X(x_1)g(x_1)$$

and

$$\{1 - F_X(x_0)\}g(x_0) = \int_{x_0}^{\infty} g(x_0)f(x_1)\mathrm{d}x_1 \leq \int_{x_0}^{\infty} g(x_1)f(x_0)\mathrm{d}x_1 = \{1 - F_Y(x_0)\}f(x_0).$$

It follows

$$\frac{F_Y(x)}{F_X(x)} \leq \frac{g(x)}{f(x)} \leq \frac{1 - F_Y(x)}{1 - F_X(x)},$$

for every $x \in \mathbb{R}$, which implies

$$\frac{F_Y(x)}{1 - F_Y(x)} \leq \frac{F_X(x)}{1 - F_X(x)}.$$

Thus, Y stochastically dominates X, i.e. the corresponding cdf's satisfy $F_Y(x) \leq F_X(x)$ for every $x \in \mathbb{R}$, which implies that $E[h(X)] \leq E[h(Y)]$ for every non-decreasing h. □

Lemma 9. *Under assumptions A1 and A2, for every $n - t > s \geq 1$ it holds*

$$\frac{\gamma[t+s+\beta, \epsilon\{\log(n)+1\}]}{\delta^2\gamma[t+\beta, \epsilon\{\log(n)+1\}]}\{\log(n)+1\}^{-s} \leq \frac{\int_0^\epsilon \frac{\alpha^{t+s}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^{t}}{\alpha^{(n)}}\pi(\alpha)\,\mathrm{d}\alpha} \leq \frac{\delta^2\gamma\{t+s+\beta, \epsilon\log(n)\}}{\gamma\{t+\beta, \epsilon\log(n)\}}\{\log(n)/(1+\epsilon)\}^{-s},$$

where $\gamma(x,y)$ is the lower incomplete Gamma function and we recall that $\epsilon, \delta, \beta > 0$ are such that for every $\alpha \in (0,\epsilon)$ it holds $\frac{1}{\delta}\alpha^\beta \leq \pi(\alpha) \leq \delta\alpha^\beta$.

Proof. By (60) it suffices to find suitable bounds of $E_n[\alpha^s]$. For the upper inequality we apply Lemma 8 with $f = p_n$, $g(\alpha) \propto (cn)^{-\alpha}\alpha^{t+\beta-1}\mathbb{1}_{(\alpha\in[0,\epsilon])}$ with $c = (1+\epsilon)^{-1}$ and $h(\alpha) = \alpha^s$. To verify that $g(\alpha)/p_n(\alpha)$ is non-decreasing for $\alpha \in (0,\epsilon]$ we compute

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}\log\left\{\frac{g(\alpha)}{p_n(\alpha)}\right\} &= -\log\left(\frac{n}{1+\epsilon}\right) + \sum_{i=1}^{n-1}\frac{1}{\alpha+i} \\ &\geq -\log\left(\frac{n+\epsilon}{1+\epsilon}\right) + \sum_{i=1}^{n-1}\frac{1}{i+\epsilon} \geq 0,\end{aligned}$$

where the last inequality follows from

$$\int_1^k \frac{1}{x+\epsilon}\,dx < \sum_{i=1}^{k-1}\frac{1}{i+\epsilon}$$

for every $k > 1$. Thus, since $h(\alpha) = \alpha^s$ is non-decreasing in α it follows by Lemma 8 that

$$\begin{aligned} E_n[\alpha^s] \leq & \frac{\int_0^\epsilon \alpha^{t+s+\beta-1}(cn)^{-\alpha} \mathrm{d}\alpha}{\int_0^\epsilon \alpha^{t+\beta-1}(cn)^{-\alpha}\, \mathrm{d}\alpha} = \frac{\{\log(cn)\}^{-s} \int_0^{\epsilon \log(cn)} z^{t+s+\beta-1} e^{-z} \mathrm{d}z}{\int_0^{\epsilon \log(cn)} z^{t+\beta-1} e^{-z}\, \mathrm{d}z} \\ = & \frac{\{\log(cn)\}^{-s} \gamma\{t+s+\beta, \epsilon \log(cn)\}}{\gamma\{t+\beta, \epsilon \log(cn)\}}. \end{aligned}$$

The lower bound again follows from Lemma 8 with $f(\alpha) \propto (en)^{-\alpha} \alpha^{t+\beta-1} \mathbb{1}_{(\alpha \in [0,\epsilon])}$, $g(\alpha) = p_n(\alpha)$ and $h(\alpha) = \alpha^s$. To verify that $p_n(\alpha)/f(\alpha)$ is non-decreasing for $\alpha \in (0, \epsilon]$ we compute

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\alpha} \log \left\{ \frac{p_n(\alpha)}{f(\alpha)} \right\} &= -\sum_{i=1}^{n-1} \frac{1}{\alpha + i} + \log(n) + 1 \\ &\geq -\sum_{i=1}^{n-1} \frac{1}{i} + \log(n) + 1 \geq 0, \end{aligned}$$

where the last inequality follows from

$$\sum_{i=1}^{k} \frac{1}{i} \leq \log(k) + 1$$

for every $k \geq 1$. Thus, since $h(\alpha) = \alpha^s$ is non-decreasing in α, we have

$$\begin{aligned} E_n[\alpha^s] \geq & \frac{\int_0^\epsilon \alpha^{t+s+\beta-1}(en)^{-\alpha} \mathrm{d}\alpha}{\int_0^\epsilon \alpha^{t+\beta-1}(en)^{-\alpha}\, \mathrm{d}\alpha} = \frac{\{\log(en)\}^{-s} \int_0^{\epsilon \log(en)} z^{t+s+\beta-1} e^{-z} \mathrm{d}z}{\int_0^{\epsilon \log(en)} z^{t+\beta-1} e^{-z}\, \mathrm{d}z} \\ = & \frac{\{\log(en)\}^{-s} \gamma\{t+s+\beta, \epsilon \log(en)\}}{\gamma\{t+\beta, \epsilon \log(en)\}}. \end{aligned}$$

The proof is completed by combining the bounds with (60). □ □

Lemma 10. *For every $\epsilon > 0$, there exists $M > 0$ such that, for every $n \geq 1$, it holds*

$$M \int_0^\epsilon \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha \geq \int_\epsilon^\infty \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha\,.$$

Proof. Define $p = \frac{\int_\epsilon^\infty \alpha^t \pi(\alpha)\, \mathrm{d}\alpha}{\int_0^{\frac{\epsilon}{2}} \alpha^t \pi(\alpha)\, \mathrm{d}\alpha}$. Then

$$\begin{aligned} \int_0^\epsilon \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha - \int_\epsilon^\infty \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha &= \int_0^\epsilon \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha - \int_0^{\frac{\epsilon}{2}} p \frac{\alpha^t}{\epsilon^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha \\ &\geq \int_0^{\frac{\epsilon}{2}} \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha - \int_0^{\frac{\epsilon}{2}} p \frac{\alpha^t}{\epsilon^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha. \end{aligned}$$

Choose m such that $\left(\frac{\epsilon}{2}\right)^{(m)} < \frac{\epsilon^{(m)}}{p}$, which is always possible because $\left\{\epsilon^{(m)}\right\}^{-1} \left(\frac{\epsilon}{2}\right)^{(m)} \to 0$ as $m \to \infty$. Thus

$$\int_0^\epsilon \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha \geq \int_\epsilon^\infty \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha)\, \mathrm{d}\alpha, \quad n \geq m$$

and it suffices to set $M = \max(P, 1)$ with

$$P = \max_{1 \le i \le m} \left\{ \frac{\int_\epsilon^\infty \frac{\alpha^t}{\alpha^{(i)}} \pi(\alpha) \, \mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^t}{\alpha^{(i)}} \pi(\alpha) \, \mathrm{d}\alpha} \right\}.$$

□ □

Proof of Proposition 10. We first prove the upper bound. We have

$$C(n,t,t+s) \le \frac{\int_0^\infty \frac{\alpha^{t+s}}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha} = \frac{\int_0^\epsilon \frac{\alpha^{t+s}}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha} + \frac{\int_0^\epsilon \frac{\alpha^{t+s}}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha} \frac{\int_\epsilon^\infty \frac{\alpha^{t+s}}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^{t+s}}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha}.$$

Moreover, it holds

$$\frac{\int_\epsilon^\infty \frac{\alpha^{t+s}}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^{t+s}}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha} \le \frac{\int_\epsilon^\infty \alpha^{t+s-1} \pi(\alpha) \, \mathrm{d}\alpha}{\int_0^\epsilon \alpha^{t+s-1} \pi(\alpha) \, \mathrm{d}\alpha} \le \delta \frac{\int_\epsilon^\infty \alpha^{t+s-1} \pi(\alpha) \, \mathrm{d}\alpha}{\int_0^\epsilon \alpha^{t+s+\beta-1} \, \mathrm{d}\alpha} \le \delta \, E(\alpha^{t+s-1}) \frac{t+s+\beta}{\epsilon^{t+s+\beta}},$$

where the first inequality follows since $\alpha^{(n)} \ge \epsilon^{(n)}$ for $\alpha \in (\epsilon, \infty)$ and $\alpha^{(n)} \le \epsilon^{(n)}$ for $\alpha \in (0, \epsilon)$, while the second one follows from assumption $A2$. Moreover, E stands for the expected value with respect to π. Thus from Lemma 13 it holds

$$C(n,t,t+s) \le \frac{\delta^2 \left\{ 1 + E(\alpha^{t+s-1}) \frac{t+s+\beta}{\epsilon^{t+s+\beta}} \right\} \gamma\{t+s+\beta, \epsilon \log(n)\}}{\gamma\{t+\beta, \epsilon \log(n)\}} \{\log(n)/(1+\epsilon)\}^{-s}.$$

Then choose $G = \frac{4\delta^2}{\epsilon^{t+\beta} \gamma(t+\beta, \epsilon \log 2)}$ to obtain the upper bound. For the lower bound, apply Lemma 13 and Lemma 10 to get

$$C(n,t,t+s) \ge \frac{1}{M+1} \frac{\int_0^\epsilon \frac{\alpha^{t+s}}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha}{\int_0^\epsilon \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha} \ge \frac{1}{M+1} \frac{\gamma[t+s+\beta, \epsilon\{\log(n)+1\}]}{\delta^2 \gamma[t+\beta, \epsilon\{\log(n)+1\}]} \{\log(n)+1\}^{-s}.$$

Then choose $F = \frac{1}{(M+1)\delta^2 \gamma(t+\beta)}$. □ □

The following corollary of Proposition 10 will be useful.

Corollary 3. *Suppose π satisfies assumptions $A1$ and $A2$. Then $G > 0$ as in Proposition 10 is such that for every $0 < s < n$ and $n \ge 4$ it holds*

$$C(n,t,t+s) \le \frac{G \Gamma(t+\beta+1) 2^s s}{\epsilon} \mathbb{E}[\alpha^{t+s-1}] \log\{n/(1+\epsilon)\}^{-1}.$$

Proof. By Proposition 10 we have

$$C(n,t,t+s) \le \frac{Gs}{\epsilon^s} \mathbb{E}[\alpha^{t+s-1}] \frac{\gamma\{t+s+\beta, \epsilon \log(n)\}}{\log\{n/(1+\epsilon)\}^s}.$$

Note that

$$\gamma\{t+s+\beta, \epsilon \log(n)\} = \int_0^{\epsilon \log(n)} x^{t+s+\beta-1} e^{-x} \, \mathrm{d}x \le \epsilon^{s-1} \{\log(n)\}^{s-1} \Gamma(t+\beta+1),$$

that implies

$$\frac{\gamma\{t+s+\beta, \epsilon\log(n)\}}{\epsilon^s \log^s\{n/(1+\epsilon)\}} \leq \frac{\Gamma(t+\beta+1)}{\epsilon}\left[\frac{\log(n)}{\log\{n/(1+\epsilon)\}}\right]^{s-1} \log\{n/(1+\epsilon)\}^{-1}.$$

Moreover, since $\epsilon < 1$, we have $\log\{n/(1+\epsilon)\} \geq \frac{1}{2}\log(n)$ for every $n \geq 4$. Combining the inequalities above we obtain the desired result. □ □

Proof of Lemma 6

Proof. We need to study $R(n,1,s)$ as in (2.52). Taking the expectation with respect to the data generating distribution we have

$$\begin{aligned}
\mathbb{E}[R(n,1,s)] &= \sum_{A\in\mathcal{T}_s(n)} \frac{\prod_{j=1}^s (a_j-1)!}{(n-1)!}\mathbb{E}\left[\frac{\prod_{j=1}^s m(X_{A_j})}{m(X_{1:n})}\right] \\
&= \sum_{\boldsymbol{a}\in\mathcal{F}_s(n)} \binom{n}{a_1\cdots a_j}\frac{\prod_{j=1}^s (a_j-1)!}{s!(n-1)!}\mathbb{E}\left[\frac{\prod_{j=1}^s m(X_{A_j^{\boldsymbol{a}}})}{m(X_{1:n})}\right] \\
&= \sum_{\boldsymbol{a}\in\mathcal{F}_s(n)} \frac{n}{s!\prod_{j=1}^s a_j}\mathbb{E}\left[\frac{\prod_{j=1}^s m(X_{A_j^{\boldsymbol{a}}})}{m(X_{1:n})}\right].
\end{aligned}$$

□ □

Proof of Lemma 4

Proof. Assumptions $A1$ and $A2$ are immediately satisfied in all three cases discussed in the statement of the lemma. We thus focus on proving that $A3$ is satisfied, considering each of the three cases separately. Suppose first that the support of the density π is contained in $[0,c]$ with $c > 0$. Then

$$\int_0^\infty \alpha^s \pi(\alpha)\,\mathrm{d}\alpha \leq c^s\,.$$

Thus in this case assumption $A3$ is satisfied for every $\rho > 0$ because $c^s < D\rho^{-s}\Gamma(s+1)$ with $D = \max_{s\in\mathbb{N}} \frac{(c\rho)^s}{\Gamma(s+1)}$ for every $\rho > 0$. Suppose now the prior is given by a Generalized Gamma distribution, so that

$$\int_0^\infty \alpha^s \pi(\alpha)\,\mathrm{d}\alpha = \frac{p}{a^d\Gamma\left(\frac{d}{p}\right)}\int_0^\infty \alpha^{d+s-1}e^{-\left(\frac{\alpha}{a}\right)^p}\,\mathrm{d}\alpha\,.$$

The condition $p > 1$ implies that, for every fixed $\rho > 0$ and $a > 0$, there exists $k > 0$ such that $\rho\alpha \leq \left(\frac{\alpha}{a}\right)^p$ for every $\alpha \geq k$. Thus

$$\begin{aligned}
\int_0^\infty \alpha^{d+s-1}e^{-\left(\frac{\alpha}{a}\right)^p}\,\mathrm{d}\alpha &\leq \int_0^k \alpha^{s+d-1}e^{-\left(\frac{\alpha}{a}\right)^p}\,\mathrm{d}\alpha + \int_k^\infty \alpha^{s+d-1}e^{-\rho\alpha}\,\mathrm{d}\alpha \\
&\leq k^{s+d-1}e^{-\left(\frac{k}{a}\right)^p} + \rho^{-d-s}\Gamma(s+d).
\end{aligned}$$

Also,

$$\int_0^\infty \alpha^s \pi(\alpha)\,\mathrm{d}\alpha \leq \frac{p}{a^d \Gamma\left(\frac{d}{p}\right)} \Gamma(s+d) \left\{ \frac{k^{s+d-1} e^{-\left(\frac{k}{a}\right)^p}}{\Gamma(s+d)} + \rho^{-d-s} \right\} \leq$$
$$\leq D\rho^{-s}\Gamma(s+d),$$

with $D = \max_{s\in\mathbb{N}} \frac{p}{a^d\Gamma\left(\frac{d}{p}\right)} \left\{ \frac{k^{s+d-1}e^{-\left(\frac{k}{a}\right)^p}\rho^s}{\Gamma(s+d)} + \rho^{-d} \right\}$, so that also in this case assumption $A3$ is satisfied for every $\rho > 0$. Finally, in the case of Gamma distribution we get

$$\int_0^\infty \alpha^s \pi(\alpha)\,\mathrm{d}\alpha = \frac{\Gamma(\nu+s)}{\Gamma(\nu)}\rho^{-s}$$

and assumption $A3$ holds. □

Proof of Theorem 10

Through a linear rescaling, we may assume $[a, b] = [-c, c]$ without loss of generality. We rewrite the assumptions on g and Q_0 as

$T1$. $\exists\, m, M$ such that $0 < m \leq g(x) \leq M < \infty$ for every $x \in [-c, c]$;

$T2$. g is differentiable on $(-c, c)$ and $\exists\, R$ such that $|\frac{g'(x)}{g(x)}| \leq R < \infty$ for every $x \in (-c, c)$;

$T3$. $\exists\, U > 0$ such that $h(y) = q_0(y) + q_0(-y) \leq U$ for every $y \in [0, 2c]$;

$T4$. $\exists\, L > 0$ such that $q_0(\theta) \geq L$ for every θ in a neighborhood of θ_j^*, for every j.

Denote with $f(x) = \sum_{j=1}^t p_j k(x \mid \theta_j^*)$ the density of the data generating $P = \sum_{j=1}^t p_j R_j$, with $t \in \mathbb{N}$, $p_j \in (0, 1)$ and $\sum_{j=1}^t p_j = 1$. Since $\theta^* = (\theta_1^*, \ldots, \theta_t^*)$ is completely separated and $X^\infty \sim P_*^{(\infty)}$, each point x has non-null density for at most one component of the mixture, i.e.

$$x \in [\theta_i^* + a, \theta_i^* + b] \quad \Rightarrow \quad f(x) = p_i k(x \mid \theta_i^*) = p_i g(x - \theta_i^*).$$

Therefore we can define

$$C_j = \left\{ i \in \{1, \ldots, n\} \,:\, x_i \in [\theta_j^* + a, \theta_j^* + b] \right\}, \quad n_j = |C_j|.$$

Notice that $C_i \cap C_j = \emptyset$ for every $i \neq j$ and $\{1, \ldots, n\} = \bigcup_{j=1}^t C_j$, so that $\sum_{j=1}^t n_j = n$. Moreover, defining

$$C^{(n)} = \{ n_j > 0 \text{ for every } j \},$$

for every $x_{1:n} \in C^{(n)}$ it holds

$$\begin{aligned} &\sum_{A\in\tau_s(n)} \prod_{j=1}^s (a_j - 1)! \prod_{j=1}^s m(x_{A_j}) = 0 \quad \text{for every } s < t, \\ &\sum_{B\in\tau_t(n)} \prod_{j=1}^t (b_j - 1)! \prod_{j=1}^t m(x_{B_j}) = \prod_{j=1}^t (n_j - 1)! \prod_{j=1}^t m(x_{C_j}). \end{aligned} \tag{62}$$

Since $p_j > 0$ for every $j = 1, \dots, s$, we have $P_*^{(n)}(C^{(n)}) \to 1$ as $n \to \infty$. We need a technical lemma.

Lemma 11. *Let Ω_n be a sequence of sets depending on $X_{1:n}$, and let Z_n be random variables on the same probability space such that $P^{(\infty)}(\Omega_n) \to 1$ and*

$$Z_n \mathbb{1}_{\Omega_n} \to 0$$

in $P_*^{(\infty)}$-probability as $n \to \infty$. Then $Z_n \to 0$ in $P_*^{(\infty)}$-probability as $n \to \infty$.

Proof. By assumption $P_*^{(\infty)} \left(\mathbb{1}_{\Omega_n} Z_n > \epsilon \right) \to 0$ as $n \to \infty$. Thus, we have

$$P_*^{(\infty)} \left(Z_n > \epsilon \right) \leq P_*^{(\infty)} \left\{ \left(Z_n > \epsilon \right) \cap \Omega_n \right\} + P_*^{(\infty)} \left(\Omega_n^c \right) \to 0$$

as $n \to \infty$. □ □

Thus by Lemma 11 it suffices to study

$$\frac{\mathbb{P}(K_n = s \mid X_{1:n})}{\mathbb{P}(K_n = t \mid X_{1:n})} \mathbb{1}_{C^{(n)}} = \frac{\int \frac{\alpha^s}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha}{\int \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha) \, \mathrm{d}\alpha} \frac{\sum_{A \in \tau_s(n)} \prod_{j=1}^s (a_j - 1)! \prod_{j=1}^s m(X_{A_j})}{\sum_{B \in \tau_t(n)} \prod_{j=1}^t (b_j - 1)! \prod_{j=1}^t m(X_{B_j})} \mathbb{1}_{C^{(n)}}. \tag{63}$$

By (62), we have

$$\frac{\mathbb{P}(K_n = s \mid X_{1:n})}{\mathbb{P}(K_n = t \mid X_{1:n})} \mathbb{1}_{C^{(n)}} = 0$$

for every $s < t$. Let us now consider the case $s > t$. Again by complete separability, $A \in \tau_s(n)$ yields positive marginal density only if A is a refinement of the partition $\{C_1, \dots, C_t\}$, i.e. if

$$A \in \tilde{\tau}_s(n) = \left\{ A \in \tau_s(n) \, : \, \forall \, i = 1, \dots, s \text{ there exists } j \in \{1, \dots, t\} \text{ such that } A_i \subset C_j \right\}.$$

Therefore, if $A \in \tilde{\tau}_s(n)$, we write the j-the element as $A_j = (A_1^j, \dots, A_{s_j}^j)$ with $a_k^j = |A_k^j|$, so that

$$\sum_{A \in \tilde{\tau}_s(n)} \prod_{j=1}^s (a_j - 1)! \prod_{j=1}^s m(X_{A_j}) = \sum_{\mathbf{s} \in \mathbf{S}} \prod_{j=1}^t \sum_{A_j \in \tau_{s_j}(n_j)} \prod_{k=1}^{s_j} (a_k^j - 1)! \prod_{k=1}^{s_j} m(X_{A_k^j}),$$

where $\mathbf{S} = \left\{ (s_1, \dots, s_t) \, : \, 1 \leq s_j \leq n_j, \, \forall j, \text{ and } \sum_{j=1}^t s_j = s \right\}$. By the above and (62) we can rewrite (63) as

$$\begin{aligned} \frac{\mathbb{P}(K_n = s \mid X_{1:n})}{\mathbb{P}(K_n = t \mid X_{1:n})} \mathbb{1}_{C^{(n)}} &= C(n, t, s) \frac{\sum_{A \in \tilde{\tau}_s(n)} \prod_{j=1}^s (a_j - 1)! \prod_{j=1}^s m(X_{A_j})}{\prod_{j=1}^t (n_j - 1)! \prod_{j=1}^t m(X_{C_j})} \mathbb{1}_{C^{(n)}} \\ &= C(n, t, s) \sum_{\mathbf{s}} \prod_{j=1}^t \sum_{A_j \in \tau_{s_j}(n_j)} \frac{\prod_{k=1}^{s_j} (a_k^j - 1)!}{(n_j - 1)!} \frac{\prod_{k=1}^{s_j} m(X_{A_k^j})}{m(A_{C_j})} \mathbb{1}_{C^{(n)}}, \end{aligned} \tag{64}$$

where

$$m(X_{C_j}) = \int_{\mathbb{R}} \prod_{i \in C_j} k(X_i \mid \theta_j) \, Q_0(\mathrm{d}\theta_j) = \int_{\mathbb{R}} \prod_{i \in C_j} g(X_i - \theta_j) \, Q_0(\mathrm{d}\theta_j)$$

and

$$m(X_{A_h^j}) = \int_{\mathbb{R}} \prod_{i \in A_h^j} k(X_i \mid \theta_h)\, Q_0(\mathrm{d}\theta_h) = \int_{\mathbb{R}} \prod_{i \in A_h^j} g(X_i - \theta_h)\, Q_0(\mathrm{d}\theta_h),$$

with $h = 1, \dots, s_j$. We divide and multiply by

$$\prod_{i=1}^{n} f(X_i) = \prod_{j=1}^{t} \prod_{i \in C_j} p_j k(X_i \mid \theta_j^*) = \prod_{j=1}^{t} \prod_{h=1}^{s_j} \prod_{i \in A_h^j} p_j k(X_i \mid \theta_j^*),$$

so that the sum on the right hand side of (64) becomes

$$\sum_{\mathbf{s}} \prod_{j=1}^{t} \sum_{A_j \in \tau_{s_j}(n_j)} \frac{\prod_{k=1}^{s_j}(a_k^j - 1)!}{(n_j - 1)!} \frac{\prod_{k=1}^{s_j} \int_{\mathbb{R}} \prod_{i \in A_k^j} \frac{g(X_i - \theta_k)}{p_j g(X_i - \theta_j^*)}\, Q_0(\mathrm{d}\theta_k)}{\int_{\mathbb{R}} \prod_{i \in C_j} \frac{g(X_i - \theta_j)}{p_j g(X_i - \theta_j^*)}\, Q_0(\mathrm{d}\theta_j)} \mathbb{1}_{C^{(n)}}, \quad \text{for } s > t. \tag{65}$$

We start with the denominator. The next lemma specifies the behaviour of the maximum for each group, where $X_{(r)}^j$ denotes the r-th order statistic of X_{C_j}.

Lemma 12. *For every $j = 1, \dots, t$ it holds*

$$Y_{n_j}^j := \min\left[1, n_j(\log(n))^{\frac{1}{2t}}\{c + \theta_j^* - X_{(n_j)}^j\}\right] \to 1$$

in $P_*^{(\infty)}$-probability as $n \to \infty$.

Proof. First, notice that $n_j \to \infty$ $P_*^{(\infty)}$-almost surely as $n \to \infty$. By definition $Y_{n_j}^j \leq 1$, so we have to prove that $\forall \epsilon > 0$

$$P_*^{(\infty)}\left(1 - Y_{n_j}^j > \epsilon\right) \to 0$$

as $n_j \to \infty$. Without loss of generality assume $\theta_j^* = 0$. Thus, by definition we have

$$\begin{aligned} P_*^{(\infty)}(1 - Y_{n_j}^j > \epsilon) &= P_*^{(\infty)}\left[n_j(\log(n))^{\frac{1}{2t}}\{c - X_{(n)}^j\} \leq 1 - \epsilon\right] = P_*^{(\infty)}\left\{X_{(n)}^j \geq c - \frac{1-\epsilon}{n_j(\log(n))^{\frac{1}{2t}}}\right\} \\ &= 1 - \left\{1 - \int_{c - \frac{1-\epsilon}{n_j(\log(n))^{\frac{1}{2t}}}}^{c} g(x)\,\mathrm{d}x\right\}^n. \end{aligned}$$

Thus, by $T1$ we have that $\int_{c - \frac{1-\epsilon}{n_j(\log(n))^{\frac{1}{2t}}}}^{c} g(x)\,\mathrm{d}x \leq \frac{M(1-\epsilon)}{n_j(\log(n))^{\frac{1}{2t}}}$, so that

$$P_*^{(\infty)}(1 - Y_{n_j}^j > \epsilon) \leq 1 - \left\{1 - \frac{M(1-\epsilon)}{n_j(\log(n))^{\frac{1}{2t}}}\right\}^n = 1 - e^{-\frac{M(1-\epsilon)}{(\log(n))^{\frac{1}{2t}}} + n_j\, o\left(\frac{1}{n_j(\log(n))^{\frac{1}{2t}}}\right)} \to 0,$$

as $n \to \infty$, by the Taylor expansion of the logarithmic function. □

Lemma 13. *For every $j = 1, \dots, t$ it holds*

$$\prod_{i \in C_j} \frac{g(x_i - \theta_j)}{g(x_i)} \geq e^{-R} \mathbb{1}_{[0, \frac{1}{n_j}]}(|\theta_j - \theta_j^*|) \mathbb{1}_{[x^j_{(n_j)} - c, x^j_{(1)} + c]}(\theta_j - \theta_j^*).$$

with R defined in $T2$ and $x^j_{(r)}$ denotes the r-th order statistic of x_{C_j}.

Proof. Without loss of generality assume $\theta_j^* = 0$. Define $p(x) := \log g(x)$, with $x \in [-c, c]$, so that $p'(x) = \frac{g'(x)}{g(x)}$. By $T2$ and the Fundamental Theorem of Integral Calculus

$$|p(y) - p(x)| = \left| \int_x^y p'(t)\, \mathrm{d}t \right| \leq \int_x^y \left| \frac{g'(t)}{g(t)} \right| \mathrm{d}t \leq R|y - x|, \quad -c < x \leq y < c.$$

Thus, we have

$$\frac{g(x - \theta_j)}{g(x)} = e^{p(x - \theta_j) - p(x)} = e^{-\{p(x) - p(x - \theta_j)\}} \geq e^{-R|\theta_j|}, \quad x \in [-c, c].$$

Finally, we get

$$\begin{aligned} \prod_{i \in C_j} \frac{g(x_i - \theta_j)}{g(x_i)} &\geq e^{-R n_j |\theta_j|} \mathbb{1}_{[x^j_{(n_j)} - c, x^j_{(1)} + c]}(\theta_j) \geq e^{-Rn|\theta_j|} \mathbb{1}_{[0, \frac{1}{n_j}]}(|\theta_j|) \mathbb{1}_{[x^j_{(n_j)} - c, x^j_{(1)} + c]}(\theta_j) \\ &\geq e^{-R} \mathbb{1}_{[0, \frac{1}{n_j}]}(|\theta_j|) \mathbb{1}_{[x^j_{(n_j)} - c, x^j_{(1)} + c]}(\theta_j). \end{aligned}$$

□ □

Lemma 14. *For every $j = 1, \dots, t$ there exists $K > 0$ and $N_j \in \mathbb{N}$ such that for all $n_j \geq N_j$ it holds*

$$\int_{\mathbb{R}} \prod_{i \in C_j} \frac{g(X_i - \theta_j)}{g(X_i - \theta_j^*)} q_0(\theta_j)\, \mathrm{d}\theta_j \geq \frac{K^{\frac{1}{t}} Y^j_{n_j}}{n_j (\log(n))^{\frac{1}{2t}}},$$

with $Y^j_{n_j}$ defined in Lemma 12.

Proof. Without loss of generality assume $\theta_j^* = 0$. Notice that, by $T4$, there exists $N_j \in \mathbb{N}$ such that $q_0(\theta) \geq L$ for every $\theta \in \left[-\frac{1}{N_j}, 0\right]$. Thus, applying Lemma 13 and considering $n_j \geq N_j$, we get

$$\begin{aligned} \int_{\mathbb{R}} \prod_{i \in C_j} \frac{g(X_i - \theta_j)}{g(X_i)} q_0(\theta_j)\, \mathrm{d}\theta_j &\geq e^{-R} \int_{\mathbb{R}} \mathbb{1}_{[0, \frac{1}{n_j}]}(|\theta_j|) \mathbb{1}_{[X^j_{(n_j)} - c, x^j_{(1)} + c]}(\theta_j)\, q_0(\theta_j)\, \mathrm{d}\theta_j \\ &\geq e^{-R} \int_{-\frac{1}{n_j}}^0 \mathbb{1}_{\{X^j_{(n_j)} \leq \theta_j + c\}}\, q_0(\theta_j)\, \mathrm{d}\theta_j \geq L e^{-R} \min\left\{ \frac{1}{n_j}, c - X^j_{(n_j)} \right\}, \end{aligned}$$

with L defined in $T4$. Thus, multiplying both the numerator and the denominator by $n_j (\log(n))^{\frac{1}{2t}}$, with $n \geq N$, we have

$$\int_{\mathbb{R}} \prod_{i \in C_j} \frac{g(X_i - \theta_j)}{g(X_i)} q_0(\theta_j) \, \mathrm{d}\theta_j \geq 2Le^{-R} \min \left\{ \frac{1}{n_j}, c - X^j_{(n_j)} \right\}$$
$$\geq \frac{K^{\frac{1}{t}} \min \left[1, n_j (\log(n))^{\frac{1}{2t}} \{ c - X_{(n)} \} \right]}{n_j (\log(n))^{\frac{1}{2t}}} = \frac{K^{\frac{1}{2t}} Y_n}{n_j (\log(n))^{\frac{1}{2t}}},$$

with $K = (2Le^{-R})^t$. □

Define the event

$$\Omega_n = \left\{ \text{for every } j = 1, \ldots, t \text{ it holds: } n_j \geq N_j, Y^j_{n_j} \in [1/2, 1] \right\}, \tag{66}$$

such that $P^{(n)}_*(\Omega_n) \to 1$ thanks to Lemma 12 and Lemma 14. Thus, an upper bound of (65) with Ω_n in place of $C^{(n)}$ is given by

$$T^{(n)} := \frac{2^t \sqrt{\log(n)}}{K} \sum_{\mathbf{s}} \prod_{j=1}^{t} \sum_{A_j \in \tau_{s_j}(n_j)} n_j \frac{\prod_{k=1}^{s_j} (a^j_k - 1)!}{(n_j - 1)!} \prod_{h=1}^{s_j} \int_{\mathbb{R}} \prod_{i \in A^j_h} \frac{g(X_i - \theta_h)}{g(X_i - \theta^*_j)} Q_0(\mathrm{d}\theta_h) \mathbb{1}_{\Omega_n}, \tag{67}$$

for $s > t$. Now we apply the expected value with respect to the values of each group, as shown in the next lemma.

Lemma 15. *Under $X_{1:n} \sim P^{(n)}_*$, for every $j = 1, \ldots, t$, $s_j \geq 1$ and $(\theta_1, \ldots, \theta_{s_j}) \in \mathbb{R}^{s_j}$, we have*

$$\mathbb{E} \left[\prod_{h=1}^{s_j} \int_{\mathbb{R}^{s_j}} \prod_{i \in A^j_h} \frac{g(X_i - \theta_h)}{g(X_i - \theta^*_j)} q_0(\theta_h) \, \mathrm{d}\theta_h \right] \leq \left(\frac{U}{m} \right)^{s_j} \prod_{h=1}^{s_j} \frac{1}{a^j_h + 1},$$

with m and U defined in $T1$ and $T3$.

Proof. Without loss of generality assume $\theta^*_j = 0$. Taking the expectation under $P^{(n)}_*$ we have

$$\mathbb{E} \left[\int_{\mathbb{R}^{s_j}} \prod_{h=1}^{s_j} \prod_{i \in A^j_h} \frac{g(X_i - \theta_h)}{g(X_i)} q_0(\theta_h) \, \mathrm{d}\theta_h \right] = \int_{\mathbb{R}^{s_j}} \int_{[-c,c]^{n_j}} \prod_{h=1}^{s_j} \prod_{i \in A^j_h} g(x_i - \theta_h) q_0(\theta_h) \, \mathrm{d}x_i \, \mathrm{d}\theta_h, \tag{68}$$

By the change of variables $z = x - \theta_h$, we have

$$\int_{-c}^{c} g(x - \theta_h) \mathbb{1}_{[\theta_h - c, \theta_h + c]}(x) \, \mathrm{d}x = \int_{-c-\theta_h}^{c-\theta_h} g(z) \mathbb{1}_{[-c,c]}(z) \, \mathrm{d}z.$$

If $\theta_h > 0$, then

$$\int_{-c-\theta_h}^{c-\theta_h} g(z) \mathbb{1}_{[-c,c]}(z) \, \mathrm{d}z = \mathbb{1}_{[0,2c]}(\theta_h) \int_{-c}^{c-\theta_h} g(z) \, \mathrm{d}z$$
$$= \mathbb{1}_{[0,2c]}(\theta_h) \left(1 - \int_{c-\theta_h}^{c} g(z) \, \mathrm{d}z \right) \leq \mathbb{1}_{[0,2c]}(|\theta_h|) \left(1 - m|\theta_h| \right).$$

Similarly, if $\theta_h < 0$ we get

$$\begin{aligned}\int_{-c-\theta_h}^{c-\theta_h} g(z)\mathbb{1}_{[-c,c]}(z)\,\mathrm{d}z &= \mathbb{1}_{[-2c,0]}(\theta_h)\int_{-c-\theta_h}^{c} g(z)\,\mathrm{d}z \\ &= \mathbb{1}_{[-2c,0]}(\theta_h)\left(1-\int_{-c}^{-c-\theta_h} g(z)\,\mathrm{d}z\right) \leq \mathbb{1}_{[0,2c]}(|\theta_h|)\left(1-m|\theta_h|\right).\end{aligned}$$

Thus

$$\int_{-c}^{c} g(x-\theta_h)\mathbb{1}_{[\theta_h-c,\theta_h+c]}(x)\,\mathrm{d}x \leq \mathbb{1}_{[0,2c]}(|\theta_h|)\left(1-m|\theta_h|\right), \quad h=1,\dots,s_j,$$

which implies

$$\prod_{h=1}^{s_j}\prod_{i\in A_h^j}\int_{-c}^{c} g(x-\theta_h)\mathbb{1}_{[\theta_h-c,\theta_h+c]}(x)\,\mathrm{d}x \leq \prod_{h=1}^{s_j}\mathbb{1}_{[0,2c]}(|\theta_h|)\left(1-m|\theta_h|\right).$$

Considering h defined as in $T3$, we have

$$\int_{\mathbb{R}}\mathbb{1}_{[0,2c]}(|\theta_h|)\left(1-m|\theta_h|\right)q_0(\theta_h)\,\mathrm{d}\theta_h = \int_0^{2c}\left(1-m|\theta_h|\right)h(\theta_h)\,\mathrm{d}\theta_h, \quad h=1,\dots,s_j.$$

Combining the above with (68) we get

$$\begin{aligned}\mathbb{E}\left[\int_{\mathbb{R}^{s_j}}\prod_{h=1}^{s_j}\prod_{i\in A_h^j}\frac{g(X_i-\theta_h)}{g(X_i)}q_0(\theta_h)\,\mathrm{d}\theta_h\right] &= \int_{\mathbb{R}^{s_j}}\int_{[-c,c]^{n_j}}\prod_{h=1}^{s_j}\prod_{i\in A_h^j} g(x_i-\theta_h)q_0(\theta_h)\,\mathrm{d}x_i\,\mathrm{d}\theta_h \\ &\leq \prod_{h=1}^{s_j}\int_0^{2c}\left(1-m|\theta_h|\right)h(\theta_h)\,\mathrm{d}\theta_h.\end{aligned} \tag{69}$$

With U defined as in $T3$, we have

$$\int_0^{2c}(1-my)^{a_h^j}h(y)\,\mathrm{d}y \leq U\int_0^{2c}(1-my)^{a_h^j}\,\mathrm{d}y.$$

Now consider the change of variables $u = 1-my$ and compute

$$\int_0^{2c}(1-my)^{a_h^j}\,\mathrm{d}y = \frac{1}{m}\int_{1-2mc}^{1}u^{a_h^j}\,\mathrm{d}u = \frac{1-(1-2mc)^{a_h^j+1}}{m(a_h^j+1)} \leq \frac{1}{m(a_h^j+1)}.$$

Finally, through (69), we have

$$\begin{aligned}\mathbb{E}\left[\int_{\mathbb{R}^{s_j}}\prod_{h=1}^{s_j}\prod_{i\in A_h^j}\frac{g(X_i-\theta_h)}{g(X_i)}q_0(\theta_h)\,\mathrm{d}\theta_h\right] &\leq \prod_{h=1}^{s_j}\int_0^{2c}\left(1-m|\theta_h|\right)h(\theta_h)\,\mathrm{d}\theta_h \\ &\leq \left(\frac{U}{m}\right)^{s_j}\prod_{h=1}^{s_j}\frac{1}{a_h^j+1},\end{aligned}$$

as desired. □

We have the next two technical lemmas.

Lemma 16. *Let* $p^* = \min_{j\in\{1,\dots,t\}} p_j \in (0,1)$. *It holds*

$$\sum_{\boldsymbol{s}\in\boldsymbol{S}} \frac{s!}{\prod_{j=1}^t s_j!} = \sum_{\boldsymbol{s}} \binom{s}{s_1,\dots,s_t} \le (p^*)^{-s},$$

where $\boldsymbol{S} = \Big\{(s_1,\dots,s_t) \,:\, s_j \le n_j \text{ and } \sum_{j=1}^t s_j = s\Big\}$.

Proof. The result follows immediately from

$$\begin{aligned}\sum_{\mathbf{s}\in\mathbf{S}} \binom{s}{s_1,\dots,s_t} &\le (p^*)^{-s} \sum_{\mathbf{s}\in\mathbf{S}} \binom{s}{s_1,\dots,s_t} \prod_{j=1}^t p_j^{s_j} \\ &\le (p^*)^{-s} \sum_{\mathbf{s}\in R_t} \binom{s}{s_1,\dots,s_t} \prod_{j=1}^t p_j^{s_j} = (p^*)^{-s},\end{aligned}$$

where $R_t = \Big\{(s_1,\dots,s_t) \,:\, \sum_{j=1}^t s_j = s\Big\}$, since the sum on the right-hand side is the sum of the probabilities over all the possible values of a multinomial distribution with parameters $(s, p_1,\dots,p_t)$. □

Lemma 17. *For every* $p>1$ *and for every integers* $s\ge 2$ *and* $n \ge s$ *it holds*

$$\sum_{\boldsymbol{a}\in\mathcal{F}_s(n)} \left(\frac{n}{\prod_{j=1}^s a_j}\right)^p < C_p^{s-1},$$

where $\mathcal{F}_s(n) = \Big\{\boldsymbol{a} \in \{1,\dots,n\}^s : \sum_{j=1}^s a_j = n\Big\}$ *and* $C_p = 2^p\zeta(p)$, *with* $\zeta(p) = \sum_{a=1}^\infty \frac{1}{a^p} < \infty$.

Proof. We prove the result by induction. Consider the base case $s=2$. By the strict convexity of $x \mapsto x^p$ for $p>1$ we have

$$\sum_{\boldsymbol{a}\in\mathcal{F}_2(n)} \left(\frac{n}{a_1a_2}\right)^p = \sum_{a=1}^{n-1}\left\{\frac{n}{a(n-a)}\right\}^p = 2^p\sum_{a=1}^{n-1}\left(\frac{1}{2}\frac{1}{a} + \frac{1}{2}\frac{1}{n-a}\right)^p < 2^p\sum_{a=1}^{n-1}\frac{1}{a^p} < C_p,$$

for every $n\ge 2$. For the induction step, assume that for some $s\ge 3$ we have

$$\sum_{\boldsymbol{a}\in\mathcal{F}_{s-1}(n)} \left(\frac{n}{\prod_{j=1}^{s-1} a_j}\right)^2 < C_p^{s-2}$$

for all $n \geq s-1$. Then

$$\begin{aligned}\sum_{\boldsymbol{a}\in\mathcal{F}_s(n)}\left(\frac{n}{\prod_{j=1}^{s}a_j}\right)^p &= \sum_{a_s=1}^{n-s+1}\sum_{(a_1,\ldots,a_{s-1})\in\mathcal{F}_{s-1}(n-a_s)}\left(\frac{n}{\prod_{j=1}^{s}a_j}\right)^p\\ &= \sum_{a_s=1}^{n-s+1}\left\{\frac{n}{(n-a_s)a_s}\right\}^p\sum_{(a_1,\ldots,a_{s-1})\in\mathcal{F}_{s-1}(n-a_s)}\left(\frac{n-a_s}{\prod_{j=1}^{s-1}a_j}\right)^p\\ &\leq C_p^{s-2}\sum_{a_s=1}^{n-s+1}\left\{\frac{n}{(n-a_s)a_s}\right\}^p < C_p^{s-1}\end{aligned}$$

and thus the thesis follows by induction. □ □

In the following we will drop the subscript in C_p when the value of p is clear from the context, thus denoting $C = C_p$.

Lemma 18. *Consider the setting of* (2.42) *with* (f, k, q_0) *as in Theorem 10. Moreover, assume* $\pi(\alpha)$ *satisfies assumptions A1, A2, and A3. Then, under* $X_{1:\infty} \sim P_*^{(\infty)}$ *we have*

$$E\left[\mathbb{1}_{\Omega_n}\sum_{s=1}^{n-t}\frac{\mathbb{P}(K_n = t+s \,|\, X_{1:n})}{\mathbb{P}(K_n = t \,|\, X_{1:n})}\right] \to 0$$

as $n \to \infty$, *with* Ω_n *as in* (66).

Proof. Applying Lemma 15 we can upper bound the expected value of $T^{(n)}$ in (67) as follows

$$\begin{aligned}\mathbb{E}\left[T^{(n)}\right] &\leq \frac{2^t\sqrt{\log(n)}}{K}\left(\frac{U}{m}\right)^s\sum_{\mathbf{s}}\prod_{j=1}^{t}\sum_{A_j\in\tau_{s_j}(n_j)}\frac{n_j}{(n_j-1)!\prod_{k=1}^{s_j}(a_k^j+1)}\\ &\leq \frac{2^t\sqrt{\log(n)}}{K}\left(\frac{U}{m}\right)^s\sum_{\mathbf{s}}\prod_{j=1}^{t}\frac{1}{s_j!}\sum_{\mathbf{a}_j\in\mathcal{F}_{s_j}(n_j)}\left(\frac{n_j}{\prod_{k=1}^{s_j}a_k^j}\right)^2,\end{aligned}$$

where the last inequality follows from Lemma 6. Moreover, from Lemma 17 we have

$$\sum_{\mathbf{a}_j\in\mathcal{F}_{s_j}(n_j)}\left(\frac{n_j}{\prod_{k=1}^{s_j}a_k^j}\right)^2 < C^{s_j},$$

with constant $C < 7$. Thus

$$\mathbb{E}\left[T^{(n)}\right] \leq \frac{2^t\sqrt{\log(n)}}{K}\left(\frac{UC}{m}\right)^s\sum_{\mathbf{s}}\prod_{j=1}^{t}\frac{1}{s_j!}. \tag{70}$$

Moreover, from Corollary 3 and $A3$ we have

$$\begin{aligned}C(n,t,t+s) &\leq \frac{G\Gamma(t+\beta+1)2^s s}{\epsilon}E(\alpha^{t+s-})\log\{n/(1+\epsilon)\}^{-1}\\ &\leq \frac{DG\Gamma(t+\beta+1)2^s s}{\epsilon}\rho^{-(t+s-1)}\Gamma(\nu+t+s)\log\{n/(1+\epsilon)\}^{-1}, \quad n \geq 4\,.\end{aligned} \tag{71}$$

By (70), combined with Lemma 16, and (71) we finally have

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\Omega_n}\sum_{s=1}^{n-t}\frac{\mathbb{P}(K_n=s+t|X_{1:n})}{\mathbb{P}(K_n=t|X_{1:n})}\right] &= \sum_{s=1}^{n-t} C(n,t,t+s)\mathbb{E}[\mathbb{1}_{\Omega_n}R(n,t,t+s)] \\ &\leq \frac{2^t\rho^{1-t}(U/m)^t DG\Gamma(t+\beta+1)\sqrt{\log(n)}}{K\epsilon\log\{n/(1+\epsilon)\}}\underbrace{\sum_{s=1}^{n-1}\frac{s(2CUp^*/m)^s\rho^{-s}\Gamma(\nu+t+s)}{(s+1)!}}_{<\infty} \to 0,\end{aligned}$$

as $n\to\infty$, where finiteness follows by taking ρ sufficiently large. □

Proof of Theorem 10. First of all, assume $\pi(\cdot)$ satisfies $A1 - A3$. By Lemma 18 it holds

$$\mathbb{1}_{\Omega_n}\sum_{s=1}^{n-t}\frac{\mathbb{P}(K_n=t+s\,|\,X_{1:n})}{\mathbb{P}(K_n=t\,|\,X_{1:n})}\to 0$$

in $P_*^{(\infty)}$–probability as $n\to\infty$. The desired result then follows from Lemma 11 with $Z_n = \sum_{s=1}^{n-t}\frac{\mathbb{P}(K_n=t+s\,|\,X_{1:n})}{\mathbb{P}(K_n=t\,|\,X_{1:n})}$ and Ω_n as in (66).

Assume instead $\pi(\alpha)=\delta_{\alpha^*}(\alpha)$ with $\alpha^*>0$. By (64) we have

$$\frac{\mathbb{P}(K_n=t+1\mid X_{1:n})}{\mathbb{P}(K_n=t\mid X_{1:n})}\geq\alpha^*\sum_{\mathbf{s}\in\mathbf{S}}\prod_{j=1}^{t}\sum_{A_j\in\tau_{s_j}(n_j)}\frac{\prod_{k=1}^{s_j}(a_k^j-1)!}{(n_j-1)!}\frac{\prod_{k=1}^{s_j}m(X_{A_k^j})}{m(A_{C_j})}.$$

Notice that, with n high enough, $n_1>1$ almost surely. Then, denoting $i\in C_1$, we consider the special case

$$\mathbf{s}=(2,1,\dots,1),\quad A_1^1=\{i\}, A_2^1=A_{C_1}\backslash\{i\},$$

and $A_j=\{A_{C_j}\}$ for every $j\geq 2$. Thus we can write

$$\frac{\mathbb{P}(K_n=t+1\mid X_{1:n})}{\mathbb{P}(K_n=t\mid X_{1:n})}\geq\alpha^*\sum_{i\in C_1}\frac{1}{n_1-1}\frac{m(X_i)m\left(X_{C_1\backslash i}\right)}{m\left(X_{C_j}\right)}. \tag{72}$$

By $T1$ we have

$$\begin{aligned} m\left(X_{C_j}\right) &= \int_{\mathbb{R}}\prod_{j\in C_1} g(X_j-\theta)q_0(\theta)\,\mathrm{d}\theta \\ &\leq M\int_{\mathbb{R}}\prod_{j\in C_1\backslash i} g(X_j-\theta)q_0(\theta)\,\mathrm{d}\theta = M\,m\left(X_{C_1\backslash i}\right).\end{aligned}$$

Moreover, by $T4$ there exists $\epsilon>0$ such that

$$m(X_i)=\int_{\mathbb{R}} g(X_i-\theta)q_0(\theta)\mathrm{d}\theta\geq m\int_{\theta_1^*-\epsilon}^{\theta_1^*+\epsilon}q_0(\theta)\mathrm{d}\theta\quad\geq 2mL\epsilon.$$

Therefore, (72) becomes

$$\frac{\mathbb{P}(K_n=t+1\mid X_{1:n})}{\mathbb{P}(K_n=t\mid X_{1:n})}\geq\frac{2\alpha^*mL\epsilon}{M}\sum_{i\in C_1}\frac{1}{n_1-1}=\frac{2\alpha^*mL\epsilon}{M}\frac{n_1}{n_1-1},$$

and

$$\liminf_{n\to\infty} \sum_{s\neq t} \frac{\mathbb{P}(K_n = s \mid X_{1:n})}{\mathbb{P}(K_n = t \mid X_{1:n})} \geq \liminf_{n\to\infty} \frac{\mathbb{P}(K_n = t+1 \mid X_{1:n})}{\mathbb{P}(K_n = t \mid X_{1:n})} \geq \frac{\alpha^* m L \epsilon}{M} > 0.$$

Then

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(K_n = t \mid X_{1:n}) &= \limsup_{n\to\infty} \left\{ 1 + \sum_{s\neq t} \frac{\mathbb{P}(K_n = s \mid X_{1:n})}{\mathbb{P}(K_n = t \mid X_{1:n})} \right\}^{-1} \\ &= \frac{1}{1 + \liminf_{n\to\infty} \sum_{s\neq t} \frac{\mathbb{P}(K_n = s \mid X_{1:n})}{\mathbb{P}(K_n = t \mid X_{1:n})}} > 0,\end{aligned}$$

which completes the proof. □

Proof of Proposition 8

We adapt the proof of Theorem 2.1 in Cai et al. (2021). Denote by

$$\Psi = \{k(\cdot \mid \theta) : \theta \in \Theta \subseteq \mathbb{R}^p\}$$

the family of kernels, dominated by μ, either Lebesgue or counting measure, and with common domain $\mathbb{X} \subseteq \mathbb{R}^q$. Denote with $B_x(\epsilon)$ the closed ball of center $x \in \mathbb{X}$ and radius $\epsilon > 0$. Let $\bar{\Theta}$ be the closure of Θ and define the set

$$\mathbb{B} := \left\{ \bar{\theta} \in \bar{\Theta} \backslash \Theta \;:\; \lim_{\theta \to \bar{\theta}} \left\{ \sup_x k(x \mid \theta) \right\} = \infty \right\}.$$

Let $\mathbb{G}_s$ be the set of mixtures of exactly s elements in Ψ, that is

$$f \in \mathbb{G}_s \quad \Leftrightarrow \quad f = \sum_{j=1}^{s} q_j k(\cdot \mid \theta_j),$$

with $q_j > 0$ for every j, $\sum_{j=1}^{s} q_j = 1$ and $\theta_i \neq \theta_h$ for every $i \neq h$. Let $\mathcal{P}(G)$ be the set of probability measures on a generic space G; with a slight abuse of notation we will say $f \in \mathcal{P}(G)$ when f is the density of a probability measure $P \in \mathcal{P}(G)$. Therefore, given $P \in \mathbb{G}_t$, with weights $\{p_j\}_{j=1}^t$ and parameters $\{\theta_j^*\}_{j=1}^t$, we define the Kullback-Leibler neighborhoods of P as

$$KL_\epsilon(P) := \left\{ h \in P(\mathbb{X}) : \int \log \left\{ \frac{\sum_{j=1}^t p_j k(x \mid \theta_j^*)}{h(x)} \right\} P(\mathrm{d}x) < \epsilon \right\}, \tag{73}$$

for $\epsilon > 0$. We make the following assumptions:

$H1$. For every $\bar{\theta} \in \Theta \backslash \mathbb{B}$, for μ-almost every $x \in \mathbb{X}$ there exists $A := A(\bar{\theta}, x) \subset \Theta \backslash \mathbb{B}$ neighborhood of $\bar{\theta}$ so that the mapping $\theta \in A \to k(x \mid \theta)$ is continuous. Moreover $\mathbb{B}$ is closed;

$H2$. Let $\{\theta_i\}_{i=1}^\infty \subset \Theta$. If $||\theta_i|| \to \infty$ as $i \to \infty$, then for every compact set $K \subset \mathbb{X}$, $\int_K k(x \mid \theta_i)\, \mu(\mathrm{d}x) \to 0$, as $i \to \infty$. If $\theta_i \to \bar{\theta} \in \mathbb{B}$, then there exists $x^* \in \mathbb{X}$ such that $k(\cdot \mid \theta_i) \to \delta_{x^*}(\cdot)$ weakly as $i \to \infty$;

$H3$. If $f \in \mathbb{G}_t$, then there exist no $f' \in \mathbb{G}_s$, with $s < t$, such that $f(x) = f'(x)$ μ-almost surely;

$H4$. For every $P \in \mathbb{G}_t$, $t \geq 1$, with $\theta_1^*, \ldots, \theta_t^*$ belonging to the support of Q_0, we have $\mathrm{pr}(h \in K_\epsilon(P)) > 0$ for every $\epsilon > 0$, where h follows the prior distribution in (2.42).

Assumption $H2$ says that, when θ diverges or converges to elements in $\mathbb{B}$, the kernel k degenerates: it is satisfied for instance when the elements of θ are location or scale parameters. $H3$ instead implies that the clustering problem is not ill-posed, in the sense that different numbers of components always lead to different distribution. $H4$ finally requires that the finite mixtures of the kernel $k(\cdot \mid \theta)$ belongs to the Kullback-Leibler support of the prior. They are all weak requirements, satisfied by the most common kernels. Next Lemma shows that they are satisfied under assumptions $B1 - B3$.

Lemma 19. *Suppose the kernel $k(x \mid \theta)$ satisfies assumptions $B1 - B3$. Then $H1 - H4$ are fulfilled.*

Proof. Assumption $H3$ can be easily deduced from $B1$ and (2.47). As regards $H1$, since $\sup_{\theta\in\Theta, x\in\mathbb{X}} k(x \mid \theta) < \infty$, we have $\mathbb{B} = \emptyset$. Moreover, fix $\bar{\theta} \in \mathbb{R}$. If $x > \theta + b$, choose

$$A(\bar{\theta}, x) = \left(\bar{\theta} - \frac{x - \bar{\theta} - b}{2}, \bar{\theta} + \frac{x - \bar{\theta} - b}{2}\right),$$

so that $x > \theta + b$ that implies $k(x \mid \theta) = 0$ for every $\theta \in A(\bar{\theta}, x)$. Similarly, if $x < \theta + a$, choose

$$A(\bar{\theta}, x) = \left(\bar{\theta} - \frac{\bar{\theta} + a - x}{2}, \bar{\theta} + \frac{\bar{\theta} + a - x}{2}\right).$$

Finally, if $x \in (\bar{\theta} + a, \bar{\theta} + b)$, denoting $d = \min\{\bar{\theta} + b - x, x - \bar{\theta} - a\}$, choose

$$A(\bar{\theta}, x) = \left(\bar{\theta} - \frac{d}{2}, \bar{\theta} + \frac{d}{2}\right).$$

Then $k(x \mid \theta) = g(x - \theta)$ for every $\theta \in A(\bar{\theta}, x)$ and g is continuous on (a, b), by $B2$. Thus we can find the required neighborhood $A(\bar{\theta}, x)$ for every $x \notin \{\bar{\theta} + a, \bar{\theta} + b\}$, that is for μ-almost every x, since μ is the Lebesgue measure. Therefore $H1$ is satisfied.

$H2$ follows since θ is a location parameter and $\bar{\Theta} = \Theta$. We are left to show that $H4$ is satisfied: we prove the case $t = 1$ and the general setting follows similarly.

Recall that assumptions $B1 - B3$ can be rewritten as $T1 - T4$ in the proof of Theorem 10 and let $f(x) = k(x \mid \theta^*)$ be the density function of P. Fix $\delta > 0$, $\epsilon > 0$ and denote $r = 1 - exp(\epsilon/4)$. Define the set

$$\begin{aligned}\mathbb{F}(\delta, r) := \Bigg\{ p(x) = \sum_{j=1}^{\infty} q_j k(x \mid \theta_j) \ : q_1 \in [1 - r, 1], q_2 \in [r/2, 1], \\ 0 \le \theta^* - \theta_1 \le \delta, 0 \le \theta_2 - \theta^* \le \delta \Bigg\}.\end{aligned} \tag{74}$$

We denote $[a_j, b_j] := [a + \theta_j, b + \theta_j]$, with $j \ge 1$, and similarly $[a^*, b^*] := [a + \theta^*, b + \theta^*]$. Then we can choose δ small enough such that

$$[a_1, b_1] \cup [a_2, b_2] \supseteq [a^*, b^*],$$

for every θ_1 and θ_2 as in (74). Moreover, for every $x \in S_1 := [a_1, b_1] \cap [a^*, b^*]$ we have

$$\log\left\{\frac{g(x-\theta^*)}{q_1 g(x-\theta_1)}\right\} = -\log(q_1) + \log\left\{\frac{g(x-\theta^*)}{g(x-\theta_1)}\right\} \leq \epsilon/4 + \log\left\{\frac{g(x-\theta^*)}{g(x-\theta_1)}\right\}$$
$$\leq \epsilon/4 + R|\theta^* - \theta_1|$$

with $R > 0$ as in $T2$. Therefore we can choose δ small enough so that

$$\log\left\{\frac{g(x-\theta^*)}{q_1 g(x-\theta_1)}\right\} < \frac{\epsilon}{2} \tag{75}$$

for every $x \in S_1$. Similarly, we can choose δ small enough so that for every $x \in S_2 := [a^*, b^*] \setminus [a_1, b_1]$ we have

$$\int_{S_2} g(x-\theta^*)\log\left\{\frac{g(x-\theta^*)}{q_2 g(x-\theta_2)}\right\}\mathrm{d}x < \frac{\epsilon}{2}. \tag{76}$$

Indeed, since $g(x-\theta^*) \leq M$ and $m \leq g(x-\theta_2)$ for every x in S_2, with m and M as in $T1$, we have

$$g(x-\theta^*)\log\left\{\frac{g(x-\theta^*)}{q_2 g(x-\theta_2)}\right\} < M\log\{2M/(mr)\},$$

and S_2 has arbitrarily small length with δ small enough. For every $p \in \mathbb{F}(\delta, r)$, by applying (75) and (76), we have

$$\int_{a^*}^{b^*} g(x-\theta^*)\log\left\{\frac{g(x-\theta^*)}{\sum_{j=1}^\infty q_j g(x-\theta_j)}\right\}\mathrm{d}x =$$
$$\int_{S_1} g(x-\theta^*)\log\left\{\frac{g(x-\theta^*)}{\sum_{j=1}^\infty q_j g(x-\theta_j)}\right\}\mathrm{d}x + \int_{S_2} g(x-\theta^*)\log\left\{\frac{g(x-\theta^*)}{\sum_{j=1}^\infty q_j g(x-\theta_j)}\right\}\mathrm{d}x \leq$$
$$\int_{S_1} g(x-\theta^*)\log\left\{\frac{g(x-\theta^*)}{q_1 g(x-\theta_1)}\right\}\mathrm{d}x + \int_{S_2} g(x-\theta^*)\log\left\{\frac{g(x-\theta^*)}{q_2 g(x-\theta_2)}\right\}\mathrm{d}x \leq \epsilon.$$

Thus, $\mathbb{F}(\delta, r) \subseteq K_\epsilon(P)$ for δ small enough. Moreover, since θ^* belongs to the support of Q_0 and the Dirichlet process prior has full weak support on the space of probability weights $\{q_j\}_j$, we have that

$$\mathrm{pr}\{h \in K_\epsilon(P)\} \geq \mathrm{pr}\{h \in \mathbb{F}(\delta, r)\} > 0,$$

as desired. □ □

The proof of Proposition 8 will rely on the following Lemma.

Lemma 20. *Let assumption $H4$ be satisfied and let $P \in \mathbb{G}_t$ with parameters $\theta_1^*, \ldots, \theta_t^*$ belonging to the support of Q_0. Assume there exists $\mathcal{U}$ weak neighborhood of P such that $\mathcal{U} \cap \mathbb{G}_s = \emptyset$ for every $s < t$. Then*

$$pr\left(K_n < t \mid X_{1:n}\right) \to 0,$$

in $P^{(\infty)}$-probability as $n \to \infty$.

Proof. By assumption $H4$, the posterior distribution is consistent at P under the weak topology, in virtue of Schwartz theorem (see e.g.Theorem 6.16 and Example 6.20 in Ghosal and Van Der

Vaart (2017)), so that

$$\mathrm{pr}(h \in \mathcal{U}^c \mid X_{1:n}) \to 0, \tag{77}$$

in $P^{(\infty)}$-probability as $n \to \infty$. Moreover, we have

$$\mathrm{pr}(h \in \mathcal{U}^c \mid X_{1:n}) \geq \mathrm{pr}(h \in \mathcal{U}^c \mid X_{1:n}, K_n < t)\mathrm{pr}\left(K_n < t \mid X_{1:n}\right).$$

Notice that, conditional on $K_n < t$, the domain of the posterior distribution is a subset of $\cup_{s<t}\mathbb{G}_s$. Thus we have $\mathrm{pr}(h \in \mathcal{U}^c \mid X_{1:n}, K_n < t) = 1$ and

$$\mathrm{pr}(h \in \mathcal{U}^c \mid X_{1:n}) \geq \mathrm{pr}\left(K_n < t \mid X_{1:n}\right).$$

The result follows from (77). □

We need two technical Lemmas.

Lemma 21. *Assume a sequence $\{f_i\}_{i=1}^{\infty} \subset \cup_{s<t}\mathbb{G}_s$ is such that $f_i \to f \in \mathcal{P}(\mathbb{X})$ weakly as $i \to \infty$. Then there exist $s' < t$ and a sequence $\{f'_i\}_{i=1}^{\infty} \subset \mathbb{G}_{s'}$ such that $f'_i \to f$ weakly as $i \to \infty$.*

Proof. Define

$$a_s := \sup\{i \geq 1 \,:\, f_i \in \mathbb{G}_s\}$$

with $s < t$. By construction, there exists s' such that $a_{s'} = \infty$ and $\{f'_i\}$ is the subsequence of elements of $\{f_i\}$ that belong to $\mathbb{G}_{s'}$. □

Lemma 22. *Let $\left\{f_i = \sum_{j=1}^{s} q_{j,i}k(\cdot \mid \theta_{j,i})\right\}_{i=1}^{\infty} \subset \mathbb{G}_s$ be such that $f_i \to f \in \mathcal{P}(\mathbb{X})$ weakly as $i \to \infty$. Then there exist $s' \leq s$ and a sequence $\{f'_i\}_{i=1}^{\infty} \subset \mathbb{G}_{s'}$ such that $f'_i \to f$ weakly as $i \to \infty$ and*

$$\liminf_i q'_{j,i} > 0$$

for every $j = 1, \ldots, s'$.

Proof. If $\liminf_i q_{j,i} = 0$ for every $j = 1, \ldots, s$, the statement is true by taking $s := s'$ and $f'_i := f_i$ for every $i \geq 1$. Then assume there exists l such that $\liminf_i q_{l,i} = 0$. Consider a subsequence $\{\tilde{f}_i\}_{i=1}^{\infty}$, with weights $\{\tilde{q}_{j,i}\}_i$ and parameters $\{\tilde{\theta}_{j,i}\}_i$, such that $\lim_i \tilde{q}_{l,i} = 0$ and define

$$f'_i(x) = \sum_{j \neq l} \frac{\tilde{q}_{j,i}}{\sum_{r \neq l} \tilde{q}_{r,i}} k(x \mid \tilde{\theta}_{j,i}),$$

where $\sum_{r \neq l} \tilde{q}_{r,i} \to 1$, by construction. Let $A \subset \mathbb{X}$, then

$$\begin{aligned}\left|\int_A \tilde{f}_i(x)\mu(\mathrm{d}x) - \int_A f'_i(x)\mu(\mathrm{d}x)\right| &= \sum_{j \neq l}\left(\frac{\tilde{q}_{j,i}}{\sum_{r \neq l} \tilde{q}_{r,i}} - \tilde{q}_{j,i}\right)\int_A k(x \mid \tilde{\theta}_{j,i})\mu(\mathrm{d}x) \\ &+ \tilde{q}_{l,i}\int_A k(x \mid \tilde{\theta}_{l,i})\mu(\mathrm{d}x) \leq \sum_{j \neq l}\left(\frac{\tilde{q}_{j,i}}{\sum_{r \neq l} \tilde{q}_{r,i}} - \tilde{q}_{j,i}\right) + \tilde{q}_{l,i} \to 0,\end{aligned}$$

as $i \to \infty$. Therefore, since A is arbitrary and $\{\tilde{f}_i\}$ converges to f, also $\{f'_i\}$ converges weakly to f and $\{f'_i\}_{i=1}^{\infty} \in \mathbb{G}_{s-1}$. The result follows by applying recursively the above procedure for every l satisfying $\liminf_i q_{l,i} = 0$. □

Proof of Proposition 8. By Lemma 19 we can assume $H1 - H4$ and by Lemma 20, it suffices to prove the existence of a weak neighborhood $\mathcal{U}$ of P such that $\mathcal{U} \cap \mathbb{G}_s = \emptyset$, for every $s < t$. Assume by contradiction that no such $\mathcal{U}$ exists. Then, there exists a sequence $\{f_i\} \in \cap_{s<t} \mathbb{G}_s$ such that $f_i \to f$ weakly, as $i \to \infty$, where f is the density of P. By Lemmas 21 and 22 we can assume without loss of generality that $\{f_i\} \in \mathbb{G}_s$, with $s < t$, and $\liminf_i q_{j,i} > 0$ for every $j = 1, \dots, s$. We will consider three scenarios, of which at least one must hold: (i) there exists $l \in \{1, \dots, s\}$ such that $\limsup_i ||\theta_{l,i}|| = \infty$, (ii) the sequences $\{\theta_{j,i}\}_{i=1}^{\infty}$, with $j = 1, \dots, s$, belong to a compact set $C \subset \Theta \backslash \mathbb{B}$ for i large enough, (iii) the sequences $\{\theta_{j,i}\}_{i=1}^{\infty}$, with $j = 1, \dots, s$, belong to a compact set $C \subset \Theta$ and there exists $l \in \{1, \dots, s\}$ such that $\liminf_i \inf_{\theta \in \mathbb{B}} ||\theta_{l,i} - \theta|| = 0$.

First consider case (i) and assume there exists $1 \le l \le s$ such that $||\theta_{l,r(i)}|| \to \infty$ as $i \to \infty$ for a suitable subsequence $r(i)$. Fix $0 < \epsilon < \liminf_i q_{l,i}$ and choose $K \subset \mathbb{X}$ compact set such that $P(K) > 1 - \epsilon/4$. By assumption $H2$ we have

$$\int_{K^c} f_{r(i)}(x)\mu(\mathrm{d}x) > q_{l,r(i)} \int_{K^c} k(x \mid \theta_{l,r(i)})\mu(\mathrm{d}x) > \frac{\epsilon}{2},$$

for i large enough, which contradicts the weak convergence of $\{f_i\}_{i=1}^{\infty}$ to f.

Second, assume to be in case (ii) and there exists a compact set $C \subset \Theta \backslash \mathbb{B}$ such that $\theta_{i,j} \in C$ for every $i \ge 1$ and $j = 1, \dots, s$. Define the set

$$\mathbb{D}_s := \left\{ \nu(\mathrm{d}\theta) = \sum_{j=1}^{s} q_j \delta_{\theta_j}(\mathrm{d}\theta) \, : \, \theta_j \in C, q_j > 0, \sum_{j=1}^{s} q_j = 1 \right\} \subset \mathcal{P}(\Theta).$$

Since C is compact, we have that $\mathbb{D}_s$ is tight. By Prokhorov's Theorem $\mathbb{D}_s$ is also relatively compact, so that there exists a subsequence $r(i)$ such that

$$\nu_{r(i)} = \sum_{j=1}^{s} q_{j,r(i)} \delta_{\theta_{j,r(i)}} \to \nu \in \mathcal{P}(\Theta)$$

weakly as $i \to \infty$. By Lemma 4.1 in Cai et al. (2021) we have $\nu \in \mathbb{D}_s$, so that $\nu = \sum_{j=1}^{s} \tilde{q}_j \delta_{\tilde{\theta}_j}$ for some $\tilde{q}_j \in (0, 1)$, $\sum_{j=1}^{s} \tilde{q}_j = 1$ and $\tilde{\theta}_j \in C$, for $j = 1, \dots, s$. By $H1$ and $C \subset \Theta \backslash \mathbb{B}$, for μ-almost every $x \in \mathbb{X}$, we can find $C_j := C_j(x, \tilde{\theta}_j)$, with $j = 1, \dots, s$, closed neighborhood of $\tilde{\theta}_j$, so that $k(x \mid \theta)$ is continuous as a function of θ, with $\theta \in C_j$. Define $D := \left\{ \bigcup_{j=1}^{s} C_j \right\} \cap C$ compact set: notice that $D \neq \emptyset$, since $\tilde{\theta}_j \in C \cap C_j$, with $j = 1, \dots, s$. Moreover, by construction, the mapping $\theta \in D \to k(x \mid \theta)$ is continuous and therefore bounded, since D is compact. Since $\nu_i \to \nu$ weakly, as $i \to \infty$, there exists I such that for every $i \ge I$ we have $\theta_{j,r(i)} \in D$, for every $j = 1, \dots, s$. Thus, by definition of weak convergence we have

$$\sum_{j=1}^{s} q_{j,r(i)} k(x \mid \theta_{j,r(i)}) = \int k(x \mid \theta) \nu_{r(i)}(\mathrm{d}\theta) \to \int k(x \mid \theta) \nu(\mathrm{d}\theta) = \sum_{j=1}^{s} \tilde{q}_j k(x \mid \tilde{\theta}_j),$$

as $i \to \infty$. Since almost sure pointwise convergence of densities implies weak convergence, we have

$$f_{r(i)} \to \tilde{f} = \sum_{j=1}^{s} \tilde{q}_j k(\cdot \mid \tilde{\theta}_j)$$

weakly as $i \to \infty$. By uniqueness of the weak limit, $\tilde{f}(x) = f(x)$ for μ-almost every x, that contradicts $H3$.

Third, consider case (iii). Since $\theta_{j,i} \in C \subset \Theta$ compact set, for every $j = 1, \ldots, s$ and $i \geq 1$, there exists a suitable subsequence $r(i)$ such that $\theta_{l,r(i)} \to \bar{\theta}$. Since $\mathbb{B}$ is closed by $H1$, we have that $\bar{\theta} \in \mathbb{B}$. By definition of $\mathbb{B}$, this is not possible if μ is the counting measure, since $k(x \mid \theta) \leq 1$, for every $x \in \mathbb{X}$ and $\theta \in \Theta$. Thus, let μ be the Lebesgue measure. Then we can fix $\epsilon > 0$ such that

$$P(B_{x^*}(\epsilon)) < \frac{\liminf_i q_{l,i}}{4},$$

with x^* as in $H2$. Then by $H2$ we have

$$\int_{B_{x^*}(\epsilon)} f_{r(i)}(x)\mu(\mathrm{d}x) > q_{l,r(i)} \int_{B_{x^*}(\epsilon)} k(x \mid \theta_{l,r(i)})\mu(\mathrm{d}x) > \frac{\liminf_i q_{l,i}}{2},$$

for i large enough, that again contradicts the weak convergence of $\{f_i\}_{i=1}^{\infty}$ to f. □

Proof of Theorem 11

The marginal distribution is available and given by the following lemma.

Lemma 23. *Consider k and q_0 as in* (2.49)*. Then it holds*

$$m(x_{1:n}) = \frac{2c - \{\max(x_{1:n}, \theta^*) - \min(x_{1:n}, \theta^*)\}}{(2c)^{n+1}}, \qquad (x_{1:n} \in [\theta^* - c, \theta^* + c]^n).$$

Proof. Note that $x_i \in (\theta - c, \theta + c)$ for all $i \in \{1, \ldots, n\}$ if and only if $\theta \in (\max(x_{1:n}) - c, \min(x_{1:n}) + c)$. Thus

$$\begin{aligned} m(x_{1:n}) &= \frac{1}{(2c)^{n+1}} \int_{\Theta} \prod_{i=1}^{n} \mathbb{1}_{(\theta - c, \theta + c)}(x_i) \mathbb{1}_{(\theta^* - c, \theta^* + c)}(\theta)\mathrm{d}\theta \\ &= \frac{1}{(2c)^{n+1}} \int_{\Theta} \mathbb{1}_{(\max(x_{1:n}) - c, \min(x_{1:n}) + c)}(\theta) \mathbb{1}_{(\theta^* - c, \theta^* + c)}(\theta)\mathrm{d}\theta \\ &= \frac{2c - \{\max(x_{1:n}, \theta^*) - \min(x_{1:n}, \theta^*)\}}{(2c)^{n+1}}. \end{aligned}$$

□ □

Define $\mathrm{Range}(X_A) = \max_{i \in A}(X_i) - \min_{i \in A}(X_i)$. Lemma 23 has an important corollary, that is stated after a technical lemma.

Lemma 24. *Let $A \subset \{1, \ldots, n\}$ such that $|A| = a$, Then it holds:*

$$\frac{2c - \{\max(X_A, \theta^*) - \min(X_A, \theta^*)\}}{(2c)^{a+1}} \leq \frac{2c - \mathrm{Range}(X_A)}{(2c)^{a+1}}.$$

Proof. The result follows immediately from $\max(X_A, \theta^*) \geq \max(X_A)$ and $\min(X_A, \theta^*) \leq \min(X_A)$. □

Corollary 4. *In the setting of* (2.42) *with (f, k, q_0) as in* (2.49)*, define*

$$\Omega_n = \{x \in X^{\infty} : \max(x_{1:n}) \geq \theta^* \textit{ and } \min(x_{1:n}) \leq \theta^*\}.$$

Then

$$\frac{\prod_{j=1}^{s+1} m(X_{A_j})}{m(X_{1:n})} \mathbb{1}_{\Omega_n}(X_{1:\infty}) \leq \frac{\prod_{j=1}^{s+1}\{2c - \mathrm{Range}(X_{A_j})\}}{(2c)^s\{2c - \mathrm{Range}(X_{1:n})\}}, \tag{78}$$

for every $A \in \tau_{s+1}(n)$ *.*

Proof. As regards the numerator, apply firstly Lemma 23 and then Lemma 24 to get

$$m(X_{A_j}) = \frac{2c - \{\max(X_{A_j}, \theta^*) - \min(X_{A_j}, \theta^*)\}}{(2c)^{a_j+1}} \leq \frac{2c - \mathrm{Range}(X_{A_j})}{(2c)^{a_j+1}}, \quad j = 1, \dots, s+1\,.$$

Apply Lemma 23 to $m(x_{1:n})$ for every $x \in \Omega_n$, to get

$$\begin{aligned} m(X_{1:n})\mathbb{1}_{\Omega_n}(X_{1:\infty}) &= \frac{2c - \{\max(X_{1:n}, \theta^*) - \min(X_{1:n}, \theta^*)\}}{(2c)^{n+1}} \mathbb{1}_{\Omega_n}(X_{1:\infty}) \\ &= \frac{2c - \{\max(X_{1:n}) - \min(X_{1:n})\}}{(2c)^{n+1}} \mathbb{1}_{\Omega_n}(X_{1:\infty}), \end{aligned}$$

as desired. □

The lemma below shows that, in order to prove Theorem 11, it is sufficient to show

$$\mathbb{1}_{\Omega_n}(X_{1:\infty}) \sum_{s=1}^{n-1} \frac{\mathrm{pr}(K_n = s+1 | X_{1:n})}{\mathrm{pr}(K_n = 1 | X_{1:n})} \to 0$$

in $P_*^{(\infty)}$-probability.

Lemma 25. *Consider* f *as in* (2.49) *and define* $\Omega_n = \{x \in X^\infty : \max(x_{1:n}) \geq \theta^*$ *and* $\min(x_{1:n}) \leq \theta^*\}$*. Let* $\{Y_n\}$ *be a sequence of positive random variables. Thus,* $Y_n \mathbb{1}_{\Omega_n}(X_{1:\infty}) \to 0$ *in* $P_*^{(\infty)}$*-probability implies* $Y_n \to 0$ *in* $P_*^{(\infty)}$*-probability.*

Proof. First of all, by definition of f we have

$$\max(X_{1:n}) \to \theta^* + c, \quad \min(X_{1:n}) \to \theta^* - c$$

almost surely with respect to $P_*^{(\infty)}$ as $n \to \infty$. Then $P_*^{(\infty)}(\Omega_n) \to 1$, as $n \to \infty$, by definition of Ω_n. Thus, fix $\epsilon > 0$ and notice that

$$P_*^{(\infty)}(Y_n > \epsilon) = P_*^{(\infty)}\{(Y_n > \epsilon) \cap \Omega_n\} + P_*^{(\infty)}\{(Y_n > \epsilon) \cap \Omega_n^c\}\,.$$

The first term on the right-hand side goes to 0, since $Y_n \mathbb{1}_{\Omega_n}(X_{1:\infty}) \to 0$ in $P_*^{(\infty)}$-probability, while the second vanishes because $P^{(\infty)}(\Omega_n^c) \to 0$, both as $n \to \infty$. □

Combining Corollary 4 and Lemma 25 we are ready to prove Theorem 11.

Proof of Theorem 11. For every $s \geq 1$ and $A \in \tau_s(n)$, from Corollary 4 we have

$$\frac{\prod_{j=1}^{s} m(X_{A_j})}{m(X_{1:n})} \mathbb{1}_{\Omega_n}(X_{1:\infty}) \leq \frac{\prod_{j=1}^{s}\{2c - \mathrm{Range}(X_{A_j})\}}{(2c)^{s-1}\{2c - \mathrm{Range}(X_{1:n})\}}.$$

Note that $\{2c - \text{Range}(X_{A_j})\}/(2c) \sim \text{Beta}(2, a_j - 1)$ independently for $j = 1, \ldots, s$. Moreover, recall that if $Z \sim \text{Beta}(\alpha, \beta)$ then for $p > -\alpha$

$$\mathbb{E}[Z^p] = \frac{\Gamma(\alpha+p)\Gamma(\alpha+\beta)}{\Gamma(\alpha+p+\beta)\Gamma(\alpha)}.$$

Thus, by Hölder's inequality with exponents 3 and 3/2 we get

$$\begin{aligned}
\mathbb{E}\left[\frac{\prod_{j=1}^s m(X_{A_j})}{m(X_{1:n})}\right] &\leq \mathbb{E}\left[\prod_{j=1}^s m(X_{A_j})^3\right]^{1/3} \mathbb{E}\left[m(X_{1:n})^{-3/2}\right]^{2/3} \\
&= \left\{\frac{\Gamma(5)}{\Gamma(2)}\right\}^{s/3} \left\{\frac{\Gamma(1/2)}{\Gamma(2)}\right\}^{2/3} \left\{\prod_{j=1}^s \frac{\Gamma(1+a_j)}{\Gamma(a_j+4)}\right\}^{1/3} \left\{\frac{\Gamma(1+n)}{\Gamma(n-1/2)}\right\}^{2/3}.
\end{aligned}$$

By the recursive definition of the Gamma function and recalling that $\Gamma(1/2) = \pi^{1/2}$, the upper bound above becomes

$$\begin{aligned}
\mathbb{E}\left[\frac{\prod_{j=1}^s m(X_{A_j})}{m(X_{1:n})}\right] &\leq 24^{s/3}\pi^{1/3} \left\{\prod_{j=1}^s \frac{\Gamma(1+a_j)}{\Gamma(a_j+4)}\right\}^{1/3} \left\{\frac{\Gamma(1+n)}{\Gamma(n-1/2)}\right\}^{2/3} \\
&= 24^{s/3}\pi^{1/3} \left\{\prod_{j=1}^s \frac{1}{(a_j+3)(a_j+2)(a_j+1)}\right\}^{1/3} \left\{\frac{(n-1/2)\Gamma(1+n)}{\Gamma(n+1/2)}\right\}^{2/3}.
\end{aligned}$$

Moreover, exploiting again the recursive definition of the Gamma function, Gautschi's Inequality, i.e. $\frac{\Gamma(1+n)}{\Gamma(n+1/2)} \leq (n+1)^{1/2}$, and $(n+1)/(a_j+1) < n/a_j$, we have

$$\mathbb{E}\left[\frac{\prod_{j=1}^s m(X_{A_j})}{m(X_{1:n})}\right] \leq 24^{s/3} K \left\{\prod_{j=1}^s \frac{(n+1)^3}{(a_j+1)^3}\right\}^{1/3} \leq 24^{s/3} K \left(\frac{n^3}{\prod_{j=1}^s a_i^3}\right)^{1/3} = 24^{s/3} K \frac{n}{\prod_{j=1}^s a_j}.$$

Thus, applying Lemma 6 and Lemma 17 with $p = 2$ and $C = 4\zeta(2) < 7$ we get

$$\mathbb{E}[R(n,1,s)] \leq \frac{24^{s/3}K}{s!} \sum_{a \in \mathcal{F}_s(n)} \left(\frac{n}{\prod_{j=1}^s a_j}\right)^2 < \frac{C^{s-1} 24^{s/3} K}{s!},$$

where $R(n,1,s)$ is defined as in (2.52). From Corollary 3 we have

$$C(n,1,s+1) \leq \frac{G\Gamma(2+\beta)2^s s}{\epsilon} E(\alpha^s) \log\{n/(1+\epsilon)\}^{-1}, \quad n \geq 4\,.$$

Thus, combining the inequalities above with (2.52) and assumption $A3$ we have

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\Omega_n}(X_{1:\infty}) \sum_{s=1}^{n-1} \frac{\text{pr}(K_n = s+1 | X_{1:n})}{\text{pr}(K_n = 1 | X_{1:n})}\right] &= \sum_{s=1}^{n-1} C(n,1,s+1) E\{\mathbb{1}_{\Omega_n}(X_{1:\infty}) R(n,1,s+1)\} \\
&\leq \frac{24^{1/3} D G K \Gamma(2+\beta)}{\epsilon \log\{n/(1+\epsilon)\}} \underbrace{\sum_{s=1}^{n-1} \frac{s(2C24^{1/3})^s \rho^{-s} \Gamma(\nu+s+1)}{(s+1)!}}_{<\infty} \to 0 \qquad \text{as } n \to \infty\,,
\end{aligned}$$

where finiteness follows from $\rho \geq 38 > 24^{1/3} \times 2C$. This implies that

$$\sum_{s=1}^{n-1} \frac{\mathbb{P}(K_n = s+1 | X_{1:n})}{\mathbb{P}(K_n = 1 | X_{1:n})} \to 0$$

in L^1 and thus in $P_*^{(\infty)}$-probability as $n \to \infty$. Lemma 25 with $Y_n = \sum_{s=1}^{n-1} \frac{\mathrm{pr}(K_n = s+1 | X_{1:n})}{\mathrm{pr}(K_n = 1 | X_{1:n})}$ concludes the proof. □

Proof of Theorem 12

We first need the following result.

Lemma 26. *Let k and q_0 be as in* (2.50) *and $x_1 = \cdots = x_n = \theta^*$ for some $\theta^* \in \mathbb{R}$. Then*

$$\frac{\prod_{j=1}^s m(x_{A_j})}{m(x_{1:n})} = \left\{ \frac{n+1}{\prod_{j=1}^s (a_j+1)} \right\}^{1/2} \exp\left\{ \frac{\theta^{*2}}{2} \left(-\frac{n^2}{n+1} + \sum_{j=1}^s \frac{a_j^2}{a_j+1} \right) \right\} < \left(\frac{n}{\prod_{j=1}^s a_j} \right)^{1/2},$$

for every $s = 1, \dots, n$ and every partition $A = \{A_1, \dots, A_s\} \in \tau_s(n)$.

Proof. Since the marginal likelihood can be rewritten as

$$m(x_{A_j}) = (a_j+1)^{-1/2} q_0(\theta^*)^{a_j} \exp\left\{ \frac{\theta^{*2}}{2} \frac{a_j^2}{a_j+1} \right\},$$

the first equality is obtained. The inequality follows from

$$\begin{aligned} -\frac{n^2}{n+1} + \sum_{j=1}^s \frac{a_j^2}{a_j+1} = n - \frac{n^2}{n+1} + \sum_{j=1}^s \left(\frac{a_j^2}{a_j+1} - a_j \right) = \frac{n}{n+1} - \sum_{j=1}^s \frac{a_j}{a_j+1} = \\ = \sum_{j=1}^s a_j \left(\frac{1}{n+1} - \frac{1}{a_j+1} \right) \leq 0 \end{aligned}$$

and

$$\frac{n+1}{\prod_{j=1}^s (a_j+1)} \leq \frac{n}{\prod_{j=1}^s a_j},$$

which easily follows from $a_j \leq n$, for every $j = 1, \dots, s$. □ □

Proof of Theorem 12. First, we study $R(n,1,s)$ as defined in (2.52). Since all the observations are almost surely equal, we have

$$R(n,1,s) = \sum_{\boldsymbol{a} \in \mathcal{F}_s(n)} \frac{n}{s! \prod_{j=1}^s a_j} \frac{\prod_{j=1}^s m(X_{A_j^{\boldsymbol{a}}})}{m(X_{1:n})},$$

where $A^{\boldsymbol{a}}$ is an arbitrary partition in $\tau_s(n)$ such that $|A_j^{\boldsymbol{a}}| = a_j$ for $j = 1, \dots, s$. By application of Lemma 26 and Lemma 17 with $p = 3/2$, it turns out that the constant $C = 2^{\frac{3}{2}} \zeta\left(\frac{3}{2}\right) < 8$ is such that

$$R(n,1,s) < \frac{1}{s!} \sum_{\boldsymbol{a} \in \mathcal{F}_s(n)} \left(\frac{n}{\prod_{j=1}^s a_j} \right)^{3/2} < \frac{C^{s-1}}{s!}.$$

From Corollary 3 we have

$$C(n,1,s+1) \leq \frac{G\Gamma(2+\beta)2^s s}{\epsilon} E(\alpha^s) \log\{n/(1+\epsilon)\}^{-1}, \quad n \geq 4\,. \tag{79}$$

Thus, combining the inequalities above with (2.52) and assumption $A3$ we have

$$\begin{aligned}\sum_{s=1}^{n-1} \frac{\mathrm{pr}(K_n = s+1|X_{1:n})}{\mathrm{pr}(K_n = 1|X_{1:n})} &= \sum_{s=1}^{n-1} C(n,1,s+1)R(n,1,s+1) \\ &\leq \frac{DG\Gamma(2+\beta)}{\epsilon \log\{n/(1+\epsilon)\}} \underbrace{\sum_{s=1}^{n-1} \frac{s(2C)^s \rho^{-s}\Gamma(\nu+s+1)}{(s+1)!}}_{<\infty} \to 0 \qquad \text{as } n \to \infty\,,\end{aligned} \tag{80}$$

where the finiteness follows from $\rho > 16 > 2C$. Then we conclude applying a variation of Lemma 5 with equalities and limits in probability replaced by almost sure equalities and limits (the proof of Lemma 5 extends trivially to that case). □

Proof of Proposition 9

Proof. Under (2.42), for every $\epsilon > 0$ we have

$$\begin{aligned}\mathbb{P}(\alpha < \epsilon \mid X_{1:n}) &= \sum_{s=1}^{n} \mathbb{P}(\alpha < \epsilon \mid K_n = s)\ \mathbb{P}(K_n = s \mid X_{1:n}) = \\ &\geq \mathbb{P}(\alpha < \epsilon \mid K_n = t)\ \mathbb{P}(K_n = t \mid X_{1:n}).\end{aligned}$$

By assumption, $\mathbb{P}(K_n = t \mid X_{1:n}) \to 1$ in $P_*^{(\infty)}$-probability as $n \to \infty$. Moreover, by Proposition 10 with $s = 1$ we get

$$\mathbb{E}[\alpha \mid K_n = t] = C(n,t,t+1) \to 0,$$

as $n \to \infty$. It follows $\mathbb{P}(\alpha < \epsilon \mid K_n = t) \to 1$ in $P_*^{(\infty)}$-probability as $n \to \infty$, as desired. □

References

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

Ascolani, F., Lijoi, A., Rebaudo, G., and Zanella, G. (2023a). Clustering consistency with dirichlet process mixtures. *Biometrika*, 110(2):551–558.

Ascolani, F., Lijoi, A., and Ruggiero, M. (2021). Predictive inference with fleming–viot-driven dependent dirichlet processes. *Bayesian Analysis*, 16(2):371–395.

Ascolani, F., Lijoi, A., and Ruggiero, M. (2023b). Smoothing distributions for conditional fleming–viot and dawson–watanabe diffusions. *Bernoulli*, 29(2):1410–1434.

Barrientos, A. F., Jara, A., and Quintana, F. A. (2012). On the support of maceachern' dependent dirichlet processes and extensions. *Bayesian Analysis*, 739:277–310.

Beal, M., Ghahramani, Z., and Rasmussen, C. (2001). The infinite hidden markov model. *Advances in neural information processing systems*, 14.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Anal.*, 1:121–143.

Cai, D., Campbell, T., and Broderick, T. (2021). Finite mixture models do not reliably learn the number of components. *38th Int. Conf. Mach. Learn.*, 139:1158–1169.

Canale, A. and Ruggiero, M. (2016). Bayesian nonparametric forecasting of monotonic functional time series. *Electronic Journal of Statistics*, 10:3265–3286.

Cappé, O., Moulines, E., and Rydén, T. (2009). Inference in hidden markov models. In *Proceedings of EUSFLAT conference*, pages 14–16.

Caron, F., Davy, M., and Doucet, A. (2007). Generalized pólya urn for time-varying dirichlet process mixtures. *Proc. 23rd Conf. on Uncertainty in Artificial Intelligence*, page Vancouver.

Caron, F., Neiswanger, W., Wood, F., Doucet, A., and Davy, M. (2017). Generalized pólya urn for time varying pitman-yor processes. *Journal of Machine Learning Research*, 18(27).

Caron, F. and Teh, Y. (2012). Bayesian nonparametric models for ranked data. *Advances in Neural Information Processing Systems*, 25.

Chaleyat-Maurel, M. and Genon-Catalot, V. (2006). Computable infinite-dimensional filters with applications to discretized diffusion processes. *Stochastic Processes and Applications*, 116:1447–1467.

Cinlar, E. (2011). *Probability and stochastics*. Springer.

de Finetti, B. (1938). Sur la condition d'equivalence partielle. *Actualités Scientifiques et Industrielles*, 739:5–18, Translated In: Studies in Inductive and Probability, II. Jeffrey, R. (ed.) University of California Press: Berkeley 1980.

Dunson, D. B. (2006). Bayesian dynamic modeling of latent trait distributions. *Biostatistics*, 7(4):551–568.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.*, 90:577–588.

Escobar, M. D. and West, M. (1998). Computing nonparametric hierarchical models. In *Pract. nonparametric semiparametric Bayesian Stat.*, pages 1–22. Springer, New York, NY.

Ethier, S. N. and Griffiths, R. (1993). The transition function of a fleming-viot process. *The Annals of Probability*, pages 1571–1590.

Ethier, S. N. and Kurtz, T. G. (1993). Fleming–viot processes in population genetics. *SIAM Journal on Control and Optimization*, 31(2):345–386.

Favaro, S., Ruggiero, M., and Walker, S. G. (2009). On a gibbs sampler based random process in bayesian nonparametrics. *Electronic Journal of Statistics*, 3:1556–1566.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Stat.*, 27:143–158.

Ghosal, S. and Van der Vaart, A. W. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Stat.*, 35:697–723.

Ghosal, S. and Van Der Vaart, A. W. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press.

Griffin, J. E. and Steel, M. F. (2011). Stick-breaking autoregressive processes. *Journal of econometrics*, 162(2):383–396.

Griffiths, R. C. (1980). Lines of descent in the diffusion approximation of neutral wright-fisher models. *Theoretical population biology*, 17(1):37–50.

Gutiérrez, L., Mena, R. H., and Ruggiero, M. (2016). A time dependent bayesian nonparametric model for air quality analysis. *Computational Statistics & Data Analysis*, 95:161–175.

Jenkins, P. A. and Spano, D. (2017). Exact simulation of the wright–fisher diffusion. *Annals of Applied Probability*, 3:1478–1509.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival analysis: techniques for censored and truncated data*, volume 1230. Springer.

Kon Kam King, G., Canale, A., and Ruggiero, M. (2020). Bayesian functional forecasting with locally-autoregressive dependent processes. *Bayesian Analysis*, 14:1121–1141.

Kon Kam King, G., Papaspiliopoulos, O., and Ruggiero, M. (2021). Exact inference for a class of hidden markov models on general state spaces. *Electronic Journal of Statistics*, 15:2832–2875.

Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *Ann. Stat.*, 24:911–930.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Stat.*, 12:351–357.

MacEachern, S. N. (1999). Dependent nonparametric processes. pages Alexandria, VA: American Statistical Association.

MacEachern, S. N. (2000). Dependent dirichlet processes. *Technical Report,*, page The Ohio State University.

MacEachern, S. N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *J. Comput. Graph. Stat.*, 7:223–238.

McAuliffe, J. D., Blei, D. M., and Jordan, M. I. (2006). Nonparametric empirical Bayes for the Dirichlet process mixture model. *Stat. Comput.*, 16:5–14.

Mena, R. H. and Ruggiero, M. (2016). Dynamic density estimation with diffusive dirichlet mixtures. *Bernoulli*, 22:901–926.

Miller, J. W. and Harrison, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. *Adv. Neural Inf. Process. Syst.*, 26:199–206.

Miller, J. W. and Harrison, M. T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *J. Mach. Learn. Res.*, 15:3333–3370.

Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the number of components. *J. Am. Stat. Assoc.*, 113:340–356.

Müller, P., Quintana, F. A., and Page, G. L. (2017). Nonparametric Bayesian inference in applications. *Stat. Methods Appt.*, 27:175–206.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.*, 9:249–265.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Stat.*, 41:370–400.

Ohn, I. and Lin, L. (2023). Optimal bayesian estimation of gaussian mixtures with growing number of components. *Bernoulli*, 29(2):1195–1218.

Papaspiliopoulos, O. and Ruggiero, M. (2014). Optimal filtering and the dual process. *Bernoulli*, 20:1999–2019.

Papaspiliopoulos, O., Ruggiero, M., and Spano, D. (2016). Conjugacy properties of time-evolving dirichlet and gamma random measures. *Electronic Journal of Statistics*, 10:3452–3489.

Rodriguez, A. and Ter Horst, E. (2008). Bayesian dynamic density estimation. *Bayesian Analysis*, 3:339–366.

Sethuraman, J. (1994). A constructive definition of the dirichlet process prior. *Statistica Sinica*, 2:639–650.

Stepleton, T., Ghahramani, Z., Gordon, G., and Lee, T.-S. (2009). The block diagonal infinite hidden markov model. In *Artificial intelligence and statistics*, pages 552–559. PMLR.

Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical population biology*, 26(2):119–164.

Van Gael, J., Saatci, Y., Teh, Y. W., and Ghahramani, Z. (2008). Beam sampling for the infinite hidden markov model. In *Proceedings of the 25th international conference on Machine learning*, pages 1088–1095.

Walker, S. G., Hatjispyros, S. J., and Nicoleris, T. (2007). A fleming–viot process and bayesian nonparametrics. *Annals of Applied Probability*, 17:67–80.

Yang, C.-Y., Xia, E., Ho, N., and Jordan, M. I. (2019). Posterior distribution for the number of clusters in Dirichlet process mixture models. *Preprint at arXiv:1905.09959.*

Yau, C., Papaspiliopoulos, O., Roberts, G. O., and Holmes, C. (2011). Bayesian non-parametric hidden markov models with applications in genomics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):37–57.

Zeng, C. and Duan, L. L. (2020). Quasi-Bernoulli stick-breaking: infinite mixture with cluster consistency. *Preprint at arxiv:2008.09938*.

Zhang, A., Zhu, J., and Zhang, B. (2014). Max-margin infinite hidden markov models. In *International Conference on Machine Learning*, pages 315–323. PMLR.

Chapter 3

Hierarchies beyond the Dirichlet process

3.1 Introduction

As discussed in the first Chapter, the model induced by the law of a Dirichlet process suffers from few limitations. First of all, the weights of the predictive distribution depend on the past datapoints only through the number of observations: moreover the concentration parameter θ is the only tunable parameter that affects the predictive and clustering behaviour. This implies a quite rigid structure, where for example the growth of the number of clusters is always logarithmic regardless of the specification. In this Chapter we explore some extensions which deal with such limitations and their usage in hierarchical models.

In the next Section we employ models based on Completely Random measures (Kingman, 1967; Regazzini et al., 2003; Barrios et al., 2013) to study the different types of borrowing of information which can be induced by Bayesian nonparametric models; in the last Section instead we construct trees of random probability measures, based on the Pitman–Yor process (Pitman and Yor, 1997; Pitman, 2006), with applications to partially exchangeable data. The two Sections are based on the works of Ascolani et al. (2023a) and Ascolani et al. (2023b), respectively.

3.2 Full range borrowing of information priors

3.2.1 Introduction

As discussed in the first Chapter, real phenomena often present a level of heterogeneity that makes exchangeability unrealistic: collected data may refer to different features, populations, or, in general, may be collected under different experimental conditions. Such situations entail a significant level of heterogeneity and opportunities for borrowing information, that can be exploited through the notion of partial exchangeability, which implies exchangeability within each experimental condition, but not across. Two sequences of observations $X = (X_i)_{i\geq 1}$ and $Y = (Y_j)_{j\geq 1}$, taking values in a space $\mathbb{X}$, are partially exchangeable if and only if, for all sample sizes (n, m) and all permutations (π_1, π_2),

$$\Big((X_i)_{i=1}^n, (Y_j)_{j=1}^m\Big) \overset{d}{=} \Big((X_{\pi_1(i)})_{i=1}^n, (Y_{\pi_2(j)})_{j=1}^m\Big).$$

with $\overset{d}{=}$ denoting equality in distribution. From an inferential point of view, partial exchangeability entails that the order of the observations within each sample is non-informative, while the belonging to a specific sample is relevant and has to be taken into account. Moreover, by de Finetti's representation theorem (de Finetti, 1938) X and Y are partially exchangeable if and only if there exist random probabilities (P_1, P_2) such that for any $i, j = 1, \ldots, n$

$$(X_i, Y_j) \mid P_1, P_2 \overset{\text{i.i.d.}}{\sim} P_1 \times P_2 \qquad (P_1, P_2) \sim Q \tag{3.1}$$

with Q playing the role of the prior. The dependence induced by Q at the level of the observables defines the Bayesian learning mechanism and it connects to the notion of borrowing of information. This term was first coined by John Tukey (Brillinger, 2002) and popularized with reference to Stein's paradox and empirical Bayes techniques in Efron and Morris (1977). More generally, statisticians refer to borrowing of information when many samples contribute to inference related to just one sample. Imagine to collect the samples $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$, while being interested only in the parameter P_1 associated to X. The simplest approach could be to disregard the second sample $(Y_j)_{j=1}^m$, with the drawback of losing potentially useful information. The typical borrowing instead consists in shrinking the estimates for different samples towards each other: shrinkage is justified by the fact that distributions of different, but related, populations are expected to be similar in terms of shape and/or location. However, many contexts may still require borrowing of information between $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$, but without necessarily resulting in shrinkage. Indeed, one's available prior information may imply that the responses in different groups have a negative association and, thus, tend to be dissimilar in location, which makes shrinkage undesirable. Similarly, when there is no pre-experimental knowledge on the dependence between X_i and Y_j, a flexible prior specification allowing also for negative association would be more appropriate. A toy parametric example to further clarify that borrowing does not necessary imply classic shrinkage is provided in Section A2. Some applied scenarios of borrowing of information not resulting in shrinkage are, for instance, the study of survival times and abundances of competitive species (Lee et al., 2020), the incorporation of retrospective data to study associations between biomarkers (Gong et al., 2021), the association between dental caries and dental fluorosis (Lorenz et al., 2018), the analysis of stocks and bonds returns (see Bhardwaj and Dunsby, 2013, and Section 3.2.6), and the clustering of multivariate responses with missing entries (see Section 3.3.8). In this paper we introduce a class of nonparametric priors that allows for a more general version of borrowing, which includes shrinkage as a special case. These can be used as core building blocks for models tailored to specific applications.

Starting from the pioneering works of Cifarelli and Regazzini (1978) and MacEachern (1999, 2000), Bayesian nonparametric contributions for non–exchangeable data have grown substantially, see Foti and Williamson (2013), Müller et al. (2015) and Quintana et al. (2022) for insightful reviews. The vast majority of nonparametric models for partially exchangeable data entails that the random probabilities in (26) are such that

$$\begin{cases} P_1 \overset{a.s.}{=} \sum_{k \geq 1} \bar{J}_k \delta_{\theta_k} \\ P_2 \overset{a.s.}{=} \sum_{k \geq 1} \bar{W}_k \delta_{\phi_k} \end{cases} \qquad \theta_k \overset{\text{i.i.d.}}{\sim} Q_0, \quad \phi_k \overset{\text{i.i.d.}}{\sim} Q_0 \tag{3.2}$$

where the random weights $\Big((\bar{J}_k), (\bar{W}_k)\Big)$ and the atoms $((\theta_k), (\phi_k))$ are independent and $\theta_k \perp \phi_h$ for $k \neq h$. In this paper we focus on this class of models and, for ease of exposition, take P_1 and P_2 with the same marginal distribution.

A first prominent strategy for defining Q is to explicitly assign the distribution of the weights and the atoms in (3.2) so to create dependence between P_1 and P_2: this approach led to dependent Dirichlet processes (MacEachern, 1999, 2000; Quintana et al., 2022), dependent stick-breaking processes, kernel stick-breaking processes (Dunson and Park, 2008), probit stick-breaking processes (Rodriguez and Dunson, 2011) and others. Despite their flexibility and the availability of posterior sampling schemes, the derivation of analytical results is very difficult for these models; it is often not clear how the dependence of the series reflects at the level of the observables and therefore such methods may lack transparency.

A second popular strategy, analytically more tractable, relies on completely random measures (CRMs) either working directly on the law of multi-dimensional vectors of CRMs (Epifani and Lijoi, 2010; Griffin and Leisen, 2017; Riva-Palacio and Leisen, 2021) or combining conditionally independent CRMs, using additive structures (Müller et al., 2004; Griffin et al., 2013; Lijoi and Nipoti, 2014; Lijoi et al., 2014a,b), nested structures (Rodriguez et al., 2008; Camerlenghi et al., 2019a), or hierarchical structures (Teh et al., 2006; Camerlenghi et al., 2019b). CRMs are then suitably transformed to obtain the random probabilities in (3.2).

Dependent random probabilities clearly induce dependence across groups of observations. The simplest and most intuitive way to quantify the dependence structure is through correlations. Therefore, when considering correlations among observables, we will implicitly assume real-valued X_i's and Y_j's, namely $\mathbb{X} = \mathbb{R}$. All other results and concepts are valid for general spaces $\mathbb{X}$. A first result in this direction shows that, regardless of the specific dependent model, observations in different groups cannot be more correlated (in absolute sense) than the ones in the same group.

Proposition 11. *Suppose X and Y are partially exchangeable sequences, such that P_1 and P_2 in* (26) *have the same marginal distribution. Then*

$$-\text{Corr}(X_i, X_{i'}) \leq \text{Corr}(X_i, Y_j) \leq \text{Corr}(X_i, X_{i'}),$$

for any i, i' and j.

Due to exchangeability within each group, the upper bound in Proposition 11 is always non–negative and it can be shown that, for all the models as in (3.2), the correlation between observations in the same sample, $\text{Corr}(X_i, X_{i'})$, is determined by the probability of a tie. As for the correlation across samples $\text{Corr}(X_i, Y_j)$, we show that a similar result holds true, with *hyper-ties*, the new notion we introduce, replacing ties.

Moreover, note that for known models based on CRMs, which allow for the computation of the correlation, $\text{Corr}(X_i, Y_j)$ turns out to be positive. This implies that the literature available to date is focused on models that attain a limited range of possible values of the correlation, when it can be evaluated. Here we aim to overcome this limitation and introduce a novel class of priors which yield a wider range of correlation values among the observables, including those with negative sign. The next result shows that the sign of the correlation is only determined by the dependence structure between the atoms.

Proposition 12. *Suppose X and Y are partially exchangeable sequences, such that the underlying P_1 and P_2 are as in* (3.2)*. Moreover, for any k and k', let* $\text{Corr}(\theta_k, \phi_{k'}) \geq 0$*. Then* $\text{Corr}\,(X_i, Y_j) \geq 0$*, for any i and j.*

For instance, hierarchical processes (Teh et al., 2006; Camerlenghi et al., 2019b), which represent one of the most popular dependent models, induce dependence by the sharing of atoms across groups. However, by Proposition 12, this means that achieving negative correlation is impossible. Hence, a flexible joint distribution for the sequence of atoms must be specified. This task is accomplished by our proposal, termed normalized CRMs with Full-Range Borrowing of Information (n-FuRBI), that allows to attain any possible value for the correlation specified in Proposition 11. Moreover, it encompasses many previous constructions as special cases. We will show that it nicely combines the flexibility of the random series construction with the analytical tractability featured by CRMs. Our proposal allows to consider any interesting choice of borrowing of information: independence, classical shrinkage, but also repulsion of estimates for different samples, generating what we term *full–range borrowing of information.* Note that the repulsive behaviour of n-FuRBI is different from the one featured by the priors introduced in Petralia et al. (2012) and Quinlan et al. (2017), that induce repulsion among the atoms of a single random probability measure.

3.2.2 General results on dependent processes

The vast majority of dependent processes introduced in the literature are almost surely discrete and therefore admit a series representation as in (3.2). A key preliminary step leading to the definition of hyper-tie and n-FuRBI priors is the observation that the random probabilities in (3.2) can be embedded into

$$\begin{cases} \tilde{P}_1 \overset{a.s.}{=} \sum_{k\geq 1} \bar{J}_k \delta_{(\theta_k,\phi_k)} \\ \tilde{P}_2 \overset{a.s.}{=} \sum_{k\geq 1} \bar{W}_k \delta_{(\theta_k,\phi_k)} \end{cases} \qquad (\theta_k,\phi_k) \overset{i.i.d.}{\sim} G_0, \tag{3.3}$$

with G_0 a probability distribution on $\mathbb{X}\times\mathbb{X}$, whose marginals equal Q_0. While $\tilde{P}_1$ and $\tilde{P}_2$ share the same atoms, the weights and the atoms are independent and the pair of random probability measures P_1 and P_2 in (3.2) are obtained as the projections over different coordinates of $\tilde{P}_1$ and $\tilde{P}_2$, namely $P_1(\cdot) = \tilde{P}_1(\cdot\times\mathbb{X})$ and $P_2(\cdot) = \tilde{P}_2(\mathbb{X}\times\cdot)$. The structure of popular models is recovered by letting either $G_0 = Q_0^2$, which corresponds to independence, or $G_0(\mathrm{d}\theta,\mathrm{d}\phi) = Q_0(\mathrm{d}\theta)\delta_{\{\theta\}}(\mathrm{d}\phi)$, that is $\theta_k = \phi_k$ for any k as happens for, e.g., hierarchical processes (see Camerlenghi et al., 2019b). Almost sure discreteness implies that a sample from the random probability measure P_1 (or P_2) will display ties with positive probability. The probability of a tie, i.e. a coincidence of any two observations i and j in the same sample, is

$$\beta := \mathbb{P}(X_i = X_j) = \sum_{k\geq 1}\mathbb{E}(\bar{J}_k^2) = \sum_{k\geq 1}\mathbb{E}(\bar{W}_k^2) = \mathbb{P}(Y_i = Y_j) \tag{3.4}$$

with $(\bar{J}_k)_{k\geq 1}$ and $(\bar{W}_k)_{k\geq 1}$ equal in distribution since we are assuming, for simplicity, that P_1 and P_2 are equal in distribution. When considering jointly the two samples, the concept of tie can be replaced by the one of *hyper-tie*, that is two observations in different samples coinciding with components having the same label. According to (26), its probability is

$$\gamma := \sum_{k\geq 1}\mathbb{P}(X_i = \theta_k,\, Y_j = \phi_k) = \sum_{k\geq 1}\mathbb{E}(\bar{J}_k\bar{W}_k). \tag{3.5}$$

Sampling from components with the same label is equivalent to sampling the same atom at the level of the underlying $(\tilde{P}_1,\tilde{P}_2)$ in (3.3). Clearly, when the atoms are shared between P_1 and P_2, i.e. $G_0(\mathrm{d}\theta,\mathrm{d}\phi) = Q_0(\mathrm{d}\theta)\delta_{\{\theta\}}(\mathrm{d}\phi)$, a hyper-tie corresponds to an actual tie between observations in different samples.

The next result shows the relationship between β and γ, the probabilities of a tie and hyper-tie, respectively: in particular, the probability of a tie is always larger and equality is attained if and only if the probability weights of $\tilde{P}_1$ and $\tilde{P}_2$ are almost surely equal.

Proposition 13. *Let* (P_1, P_2) *be as in* (3.2) *and* β, γ *as in* (3.4) *and* (3.5), *respectively. Then* $0 \leq \gamma \leq \beta$ *and* $\beta = \gamma$ *if and only if* $\bar{W}_k \overset{a.s.}{=} \bar{J}_k$ *for any* k.

Hyper-ties play a crucial role in determining the dependence between observables across groups, as the ties do for the dependence between observables within groups, as shown by the next proposition.

Proposition 14. *Consider model* (26) *with* (P_1, P_2) *as in* (3.2). *Then, for any* $i \neq i'$ *and any* $j \neq j'$

$$\operatorname{Corr}(X_i, X_{i'}) = \operatorname{Corr}(Y_j, Y_{j'}) = \beta \qquad\qquad \operatorname{Corr}(X_i, Y_j) = \gamma\, \rho_0$$

with ρ_0 *the correlation between two random variables jointly sampled from* G_0.

Thus, while the correlation between observations in the same sample equals the probability of a tie, the correlation between observations from different samples is determined by the probability of a hyper-tie, corrected by the correlation between atoms. Clearly a suitable choice of the joint distribution of the atoms makes the latter negative. Thus, by choosing G_0 appropriately, for instance as a bivariate normal, it is easy to tune the correlation according to the available prior knowledge. The following Corollary shows the values that can be attained, once the marginal law is specified.

Corollary 5. *Consider model* (26) *with* (P_1, P_2) *as in* (3.2). *If the marginal distribution of* $\tilde{p}_1$ *and* $\tilde{p}_2$ *is fixed, then* $\operatorname{Corr}(X_i, Y_j) \in [-\beta, \beta]$ *and the extreme values are attained if and only if the jumps are equal and* $\rho_0 = \pm 1$.

Interestingly, with equal weights and jumps, which corresponds to full exchangeability, one achieves the extreme case of $\operatorname{Corr}(X_i, Y_j) = \beta$. Null correlation, instead, is attained when atoms are uncorrelated or when the probability of hyper-ties is zero. Lastly, maximum negative correlation $\operatorname{Corr}(X_i, Y_j) = -\beta$ is attained with equal weights and negatively correlated atoms and can be thought of as the opposite case with respect to exchangeability, at least in terms of correlation. Ties and hyper-ties play a similar role also in the predictive structure, as the next result shows.

Proposition 15. *Consider model* (26) *with* (P_1, P_2) *as in* (3.2). *Then*

$$\mathbb{P}\left(X_1 \in A, X_2 \in B\right) = \beta Q_0(A \cap B) + (1 - \beta)\, Q_0(A) Q_0(B).$$

and

$$\mathbb{P}\left(X_1 \in A, Y_1 \in B\right) = \gamma G_0(A \times B) + (1 - \gamma)\, Q_0(A) Q_0(B).$$

The result is indeed quite intuitive. If X_1 and Y_1 form a hyper-tie (with probability γ) they come from the same pair of atoms and need to be sampled jointly; otherwise they refer to different atoms and are sampled independently. The same happens inside each group, where X_1 and X_2 are equal with probability β.

Example 1. The hierarchical Dirichlet process (Teh et al., 2006) is characterized by the hierarchical representation $P_i \mid \tilde{P}_0 \overset{\text{i.i.d.}}{\sim} \text{DP}(\theta, P_0)$, with $P_0 \sim \text{DP}(\theta_0, Q_0)$, where Q_0 is a diffuse measure and $\text{DP}(\alpha, H)$ denotes the law of a Dirichlet process with concentration parameter $\alpha > 0$ and baseline distribution H. Since the P_i's share the atoms, an hyper-tie corresponds to an actual tie between observations in different samples, so that with simple computations we get

$$\beta = \mathrm{Corr}(X_i, X_j) = 1 - \frac{\theta\theta_0}{(1+\theta)(1+\theta_0)}, \qquad \gamma = \mathrm{Corr}(X_i, Y_j) = \frac{1}{1+\theta_0}.$$

Thus, the correlation among the observables is forced to be positive, with θ_0 tuning the dependence; see Example 1 in Camerlenghi et al. (2019b) for more details.

Given the above results and considerations, it should be clear that γ defined in (3.5) is crucial for tuning the level of dependence. However, closed form expressions of γ are available only for a few cases and, in fact, we are facing a trade–off: on the one hand we have dependent processes based on the stick-breaking representation, that allow for high flexibility while sacrificing the availability of analytical results; on the other hand we have constructions based on CRMs, for which an extensive theory has been developed, though they are not as effective for tuning the dependence, since all the existing instances produce non-negative correlation across samples. In the following we combine the best of both approaches through n-FuRBI: they are flexible processes that can attain any value for the correlation between the observables, while at the same time a posterior representation can be derived. Their construction is based on CRMs and completely random vectors, reviewed in the next section.

3.2.3 Some basics on completely random measures

As shown in Lijoi and Prünster (2010), many Bayesian nonparametric models can be obtained as suitable transformations of CRMs; among others, these include the Dirichlet process, the Pitman-Yor process and the neutral-to-the-right priors. The extension of CRMs to the bivariate setting is provided by *completely random vectors* $\mu = (\mu_1, \mu_2)$, whose components take values in the space of boundedly finite measures on $\mathbb{X}$ and are such that, for every collection of pairwise disjoint sets $(A_i)_{i\geq 1}^n$, the random vectors $(\mu_1(A_1), \mu_2(A_1)), \dots, (\mu_1(A_n), \mu_2(A_n))$ are mutually independent. We focus on the case of no fixed atoms and no deterministic component, so that the marginal CRMs μ_1 and μ_2 are almost surely discrete and can be written as sum of $\mathbb{X}$–valued random atoms with random weights, i.e.

$$\mu_1 \overset{a.s.}{=} \sum_{i\geq 1} J_i \delta_{\kappa_i}, \quad \mu_2 \overset{a.s.}{=} \sum_{i\geq 1} W_i \delta_{\kappa_i}.$$

In the following section it will be convenient to use the reparametrization $\kappa_i = (\theta_i, \phi_i) \in \mathbb{X} = \mathbb{X}_1 \times \mathbb{X}_2$. Such completely random vectors are characterized by the Lévy-Khintchine representation

$$\mathbb{E}\left\{e^{-\mu_1(f_1) - \mu_2(f_2)}\right\} = \exp\left[-\int_{\mathbb{R}_+^2 \times \mathbb{X}} \{1 - e^{-s_1 f_1(x) - s_2 f_2(x)}\} \, v(\mathrm{d}s_1, \mathrm{d}s_2, \mathrm{d}x)\right] \tag{3.6}$$

where $\mu_i(f_i) = \int_{\mathbb{X}} f_i(x)\mu_i(\mathrm{d}x)$ for $\mathbb{R}^+$-valued f_i and $v(\mathrm{d}s_1, \mathrm{d}s_2, \mathrm{d}x)$ is the joint Lévy intensity. We shall focus on the homogeneous case, in which jumps $(J_j)_{j\geq 1}$ and locations $(X_j)_{j\geq 1}$ are independent. In terms of Lévy intensity it reads $v(\mathrm{d}s_1, \mathrm{d}s_2, \mathrm{d}x) = \rho(\mathrm{d}s_1, \mathrm{d}s_2)\alpha(\mathrm{d}x)$ for some finite measure α on $\mathbb{X}$ and measure ρ. Moreover, in the sequel we will also need the joint and marginal Laplace exponents given by

$$\psi_b(\lambda_1,\lambda_2) := \int\limits_{\mathbb{R}_+^2\times\mathbb{X}} (1-e^{-\lambda_1 s_1-\lambda_2 s_2})\rho(\mathrm{d}s_1,\mathrm{d}s_2)\alpha(\mathrm{d}x), \quad \lambda_1>0,\lambda_2>0.$$

$$\psi(\lambda) := \int\limits_{\mathbb{R}_+\times\mathbb{X}} (1-e^{-\lambda s})\rho(\mathrm{d}s)\alpha(\mathrm{d}x) \quad \lambda>0,$$

For an exhaustive account on CRMs, we refer to Kingman (1967, 1993). Completely random vectors and CRMs are often normalized to obtain random probability measures, as introduced in Regazzini et al. (2003), i.e. $P(\cdot)=\mu(\cdot)/\mu(\mathbb{X})$. Notice that in principle any random measure μ such that $\mathbb{P}(0<\mu(\mathbb{X})<\infty)=1$ can be normalized in order to define a random probability measure. However, the strength of completely random vectors and measures lies in their Lévy–Khintchine representations and unique correspondence with the associated Lévy intensity, which allow a high degree of analytical tractability. CRMs and the corresponding normalized probabilities have been extensively studied to model exchangeable data (see, for instance, James et al., 2006, 2009, 2010; Lijoi and Prünster, 2010; Favaro et al., 2016; Camerlenghi et al., 2018). Similarly, a completely random vector can be used to model dependence between two groups. For more details on completely random vectors and an interesting account of their dependence structure, we refer to Catalano et al. (2021, 2023). Since the two measures in the vector share all the atoms, by virtue of Proposition 12 the induced model yields non–negative correlation between samples. The issue is addressed in the next section, by means of a novel class of random probability measures that leverage the dependence structure specifed for the atoms.

3.2.4 Full-range borrowing of information nonparametric prior

In this section we introduce n-FuRBI and for simplicity we still consider only the case of two samples with the same a priori marginal distribution.

Definition 2. Consider a completely random vector $(\tilde{\mu}_1,\tilde{\mu}_2)$ on $\mathbb{X}^2$ with Lévy intensity $v(\mathrm{d}s_1,\mathrm{d}s_2,\mathrm{d}x_1,\mathrm{d}x_2)=\rho(\mathrm{d}s_1,\mathrm{d}s_2)\ \alpha(\mathrm{d}x_1,\mathrm{d}x_2)$, where $\alpha(\mathrm{d}x_1,\mathrm{d}x_2)=\theta G_0(\mathrm{d}x_1,\mathrm{d}x_2)$, where $\theta=\alpha(\mathbb{X}^2)\in(0,+\infty)$, and G_0 is a non-atomic probability measure on $\mathbb{X}^2$ such that $G_0(\cdot\times\mathbb{X})=G_0(\mathbb{X}\times\cdot)=Q_0(\cdot)$. Then μ_1 and μ_2 defined as

$$\mu_1(\cdot)=\tilde{\mu}_1(\mathbb{X}\times\cdot) \qquad \mu_2(\cdot)=\tilde{\mu}_2(\cdot\times\mathbb{X})$$

are CRMs with *Full-Range Borrowing of Information* (FuRBI CRMs) and underlying Lévy intensity v. The normalized versions $P_j(\cdot)=\mu_j(\cdot)/\mu_j(\mathbb{X})$ for $j=1,2$ are said *normalized CRMs with Full-Range Borrowing of Information* (n-FuRBI).

Essentially, first a pair of random measures endowed with the same locations is constructed on the product space $\mathbb{X}^2$; as a second step, the coordinates of each pair of atoms are split. Thus, the n-FuRBI admit a representation as in (3.2) and (3.3). In general FuRBI CRMs are not completely random vectors, because the joint sampling of the atoms forbids the independence of the vector evaluated on pairwise disjoint sets. However, the representation in terms of a completely random vector in the product space is useful to characterize the joint law of the FuRBI CRMs, as shown in the following proposition.

Proposition 16. *Let (μ_1,μ_2) be a vector of FuRBI CRMs. Then*

(i) *μ_1 and μ_2 are CRMs with intensity $\rho(\mathrm{d}s)\theta Q_0(\mathrm{d}x)$, where $\rho(\mathrm{d}s)=\int_{\mathbb{R}_+}\rho(\mathrm{d}s_1,\mathrm{d}s)$.*

(ii) *For any A and B, the following equality holds*

$$\mathbb{E}\Big[\mathrm{e}^{-\lambda_1\mu_1(A)-\lambda_2\mu_2(B)}\Big] = \exp\{-G_0(A\times B^c)\psi(\lambda_1) - G_0(A^c\times B)\psi(\lambda_2)\} \\ \times\ \exp\{-G_0(A\times B)\psi_b(\lambda_1,\lambda_2)\},$$

where ψ denotes the common marginal Laplace exponent and ψ_b the joint Laplace exponent of (μ_1,μ_2).

(iii) *The joint law of (μ_1,μ_2) is characterized by the joint Lévy intensity of $(\tilde\mu_1,\tilde\mu_2)$.*

The next proposition shows that the β and γ associated to any couple of n-FuRBI can be computed through their Laplace exponents.

Proposition 17. *Consider (P_1,P_2) n-FuRBI. Then the probability of a tie and of a hyper-tie are respectively*

$$\beta = -\int_{\mathbb{R}_+} u\left\{\frac{\mathrm{d}^2}{\mathrm{d}u^2}\psi(u)\right\}e^{-\psi(u)}\,\mathrm{d}u,\quad \gamma = -\int_{\mathbb{R}_+^2}\left\{\frac{\partial^2}{\partial u_1\partial u_2}\psi_b(u_1,u_2)\right\}e^{-\psi_b(u_1,u_2)}\,\mathrm{d}u_1\mathrm{d}u_2.$$

Thus, the crucial value of γ can be obtained by computing, analytically or numerically, a bivariate integral. The two results above show a recurrent trait of our approach: interesting quantities will be usually rewritten in terms of the original completely random vector, in order to exploit its analytical tractability. We conclude this section with two examples of FuRBI CRMs, that also show how some existing constructions can be obtained as special cases.

Example 2 (FuRBI CRMs with equal jumps)**.** Let $\rho(\mathrm{d}s_1)\delta_{s_1}(\mathrm{d}s_2)\,\theta\,G_0(\mathrm{d}x_1,\mathrm{d}x_2)$ be the underlying Lévy intensity. The series representation of the corresponding FuRBI CRMs is

$$\mu_1(\cdot)\overset{a.s.}{=}\sum_{k\geq 1}W_k\delta_{\theta_k}\qquad \mu_2(\cdot)\overset{a.s.}{=}\sum_{k\geq 1}W_k\delta_{\phi_k}\qquad \text{with }\ (\theta_k,\phi_k)\overset{i.i.d}{\sim}G_0.$$

Therefore, $\gamma=\beta$, so that a tie and a hyper–tie are observed with the same probability.

Example 3 (Extended Compound FuRBI CRMs)**.** *Consider the Lévy intensity*

$$v(\mathrm{d}s_1,\mathrm{d}s_2,\mathrm{d}x_1,\mathrm{d}x_2) = \int z^{-2}h(s_1/z,\ s_2/z)\,\mathrm{d}s_1\mathrm{d}s_2v^*(\mathrm{d}z)\,\theta\,G_0(\mathrm{d}x_1,\mathrm{d}x_2),$$

where h is some density and v^* is a Lévy intensity that satisfies

$$\int z^{-2}\int\min\{1,||s||\}h(s_1/z,\ s_2/z)\,\mathrm{d}s_1\mathrm{d}s_2v^*(\mathrm{d}z)<\infty,\quad ||s||=\sqrt{s_1^2+s_2^2}.$$

The series representation of the corresponding FuRBI CRMs is

$$\mu_1(\cdot)\overset{a.s.}{=}\sum_{k\geq 1}m_{1,k}W_k\delta_{\theta_k}\qquad \mu_2(\cdot)\overset{a.s.}{=}\sum_{k\geq 1}m_{2,k}W_k\delta_{\phi_k}$$

where $(\theta_k,\phi_k)\overset{i.i.d}{\sim}G_0$ and $(m_{1,k},m_{2,k})\overset{iid}{\sim}h$. When G_0 is degenerate on the main diagonal, one retrieves the class of compound random measures introduced by Griffin and Leisen (2017).

Correlation structure between n-FuRBI

In order to analyze the dependence between the marginal n-FuRBI priors P_1 and P_2, it is useful to compute the correlation of the random probability measures evaluated on the same set A. In all the existing CRM-based models such a correlation does not depend on the specific set considered and, hence, it is often used as a global measure of dependence. The next proposition provides the covariance structure between two n-FuRBI.

Proposition 18. *Let P_1 and P_2 be n-FuRBI. Then for any A, B, such that $0 \leq Q_0(A) \leq 1$ and $0 \leq Q_0(B) \leq 1$, we have $\mathrm{Cov}(P_1(A), P_2(B)) = \gamma\left[G_0(A \times B) - Q_0(A)Q_0(B)\right]$ and*

$$\mathrm{Corr}(P_1(A), P_2(B)) = \frac{\gamma}{\beta} \frac{G_0(A \times B) - Q_0(A)Q_0(B)}{\sqrt{Q_0(A)(1 - Q_0(A))Q_0(B)(1 - Q_0(B))}}.$$

By setting $A = B$, from the previous results one immediately deduces that $\mathrm{Cov}(P_1(A), P_2(A)) = \gamma\left[G_0(A \times A) - Q_0(A)^2\right]$ and

$$\mathrm{Corr}(P_1(A), P_2(A)) = \frac{\gamma}{\beta} \frac{G_0(A \times A) - Q_0(A)^2}{Q_0(A)(1 - Q_0(A))}.$$

Unlike what usually happens with existing models, here the correlation can be negative, when A is such that $G_0(A \times A) < Q_0(A)^2$, that is when G_0 exhibits a repulsive behaviour between the coordinates in $\mathbb{X}^2$. Moreover, the correlation depends on the specific set on which the two measures are evaluated and, therefore, it has to be interpreted as a local measure of dependence. See Section A2 for an illustration of this phenomenon on sets of the form $(-\infty, x)$. Note that here and in the following we use the prefix S to indicate sections of the supplementary material.

Example 4 (n-FuRBI with equal jumps)**.** *In this case, Proposition 13 entails $\beta = \gamma$. Therefore*

$$\mathrm{Corr}\left(P_1(A), P_2(A)\right) = \frac{G_0(A \times A) - Q_0(A)^2}{Q_0(A)(1 - Q_0(A))}.$$

Moreover, still by virtue of Proposition 13, for a given G_0 this is the highest possible correlation in absolute value.

Proposition 14 then provides the correlation between the observables, which is even more important from a modeling perspective.

Example 5 (Gamma n-FuRBI with equal jumps)**.** *If the common marginal is the law of a Dirichlet process, then $\mathrm{Corr}(X_i, Y_j) = \rho_0/(1 + \theta)$. Choosing appropriately ρ_0 and θ the entire range $(-1, 1)$ becomes available.*

Note that hyper-ties allow to perform a more general type of borrowing, compared to ties, even when the correlation is positive. While ties are a useful construction to model multiple samples that share certain values/latent parameters, hyper-ties can borrow information even when the two samples have no common values/latent parameter. This aspect will play a crucial role in the data-analyses of Sections 3.2.6 and 3.3.8; for these the assumption of common values would be highly unrealistic.

3.2.5 Inference

Posterior Characterization

Having provided an exhaustive description of the a priori properties of n-FuRBI, the following key step is to provide a tractable posterior characterization. Conjugacy is out of question here: even in the exchangeable context it is a property characterizing the Dirichlet process (see James et al., 2006). Nevertheless, conditional on a set of suitable latent variables, the posterior distribution of the original completely random vector (μ_1, μ_2) turns out to be again a completely random vector leading to a neat posterior characterization and viable methods for sampling.

Consider a sample of n observations $(X_i)_{i=1}^n$ from $\tilde{p}_1$ with unique values $\underline{X}_n^* = (X_1^*, \dots, X_k^*)$ and associated multiplicities $(n_1, \dots, n_k)$; analogously, consider m observations $(Y_j)_{j=1}^m$ from $\tilde{p}_2$ with unique values $\underline{Y}_m^* = (Y_1^*, \dots, Y_c^*)$ and multiplicities $(m_1, \dots, m_c)$. While it is immediate to check for ties, hyper-ties cannot be identified deterministically from the data. To this end, we define a latent random element p encoding the hyper-ties, such that $p = \{(i_l, j_l)\}_l$, where (i, j), with $1 \leq i \leq k$ and $1 \leq j \leq c$, denotes a hyper-tie between X_i^* and Y_j^*. Moreover $(i, 0)$, with $1 \leq i \leq k$, denotes that X_i^* does not form a hyper-tie with any value in $\underline{Y}_m^*$ and $(0, j)$, with $1 \leq j \leq c$, denotes that Y_j^* does not form an hyper-tie with any value in $\underline{X}_n^*$.

Therefore, if $(i, j) \in p$ with $i \neq 0$ and $j \neq 0$, it means that X_i^* and Y_j^* come from the same pair of atoms in representation (3.3). Instead, $(i, 0) \in p$ implies that X_i^* is the only value associated to a specific pair, and similarly for Y_j^* if $(0, j) \in p$. Since we are working with unique values, it is clear that each X_i^* and Y_j^* can form at most one hyper-tie, i.e. it is associated to a unique member of p. This justifies the following formal definition.

Definition 3. We say that $p = \{(i_l, j_l)\}_l$ is a *compatible hyper-ties structure* for $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$ if, firstly, for any $1 \leq i \leq k$, there exists exactly one i_l such that $i_l = i$, thus each element of $\underline{X}_n^*$ forms at most one hyper-tie; secondly, for any $1 \leq j \leq c$, there exists exactly one j_l such that $j_l = j$, thus each element of $\underline{Y}_m^*$ forms at most one hyper-tie; lastly, for any l, if $i_l = 0$ then $j_l \neq 0$, thus at least one coordinate refers to an element of $\underline{X}_n^*$ or $\underline{Y}_m^*$.

As a simple example, suppose that $\underline{X}_n$ and $\underline{Y}_m$ contain respectively 2 and 1 unique values. Then $k = 2$, $c = 1$ and the support of p is

$$\mathcal{P} = \left\{ \{(1,1), (2,0)\}, \{(1,0), (2,1)\}, \{(1,0), (2,0), (0,1)\} \right\}.$$

Once the latent structure p is identified, its elements can be conveniently partitioned into the set $\Delta_p = \{(i, j) \in p \mid i \neq 0 \text{ and } j \neq 0\}$, which includes all the hyper-ties, and the sets $\Delta_p^1 = \{(i, j) \in p \mid j = 0\}$ and $\Delta_p^2 = \{(i, j) \in p \mid i = 0\}$. If X_i^* and Y_j^* form a hyper-tie, it means that (X_i^*, Y_j^*) is an actual atom in representation (3.3). Instead, if X_i^* does not form a hyper-tie, we have a partial knowledge of the original pair: the unknown second coordinate can be sampled from $P_{X_i^*}(\cdot)$, that is the conditional distribution given X_i^*, induced by the joint measure G_0, which will henceforth be assumed to be non–atomic. A similar argument applies if Y_j^* does not form a hyper-tie.

In order to simplify notation, we set $g_{i,j} = g_0(X_i^*, Y_j^*)$, $g_{i,0} = p_0(X_i^*)$, and $g_{0,j} = q_0(Y_j^*)$, where g_0 and q_0 are the density functions of G_0 and Q_0 respectively, that we assume exist with respect to suitable dominating measures. Finally, we consider the following integrals

$$\tau_{n,m}(\underline{u}) = \int_{\mathbb{R}_+^2} e^{-u_1 s_1 - u_2 s_2} s_1^n s_2^m \, \rho(\mathrm{d}s_1, \mathrm{d}s_2), \quad \underline{u} = (u_1, u_2),$$

where often n and m will be equal to n_i and m_j, with $1 \leq i \leq k$ and $1 \leq j \leq c$. For consistency, we set $n_0 = m_0 = 0$.

The key result of the section relies on a latent structure that is identified by random variables whose conditional distributions, given $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$, are available. Indeed, these random variables are given by p, whose probability mass function is proportional to

$$\left(\prod_{(i,j)\in p} g_{i,j}\right)\int_{\mathbb{R}_+^2} u_1^{n-1}u_2^{m-1}\prod_{(i,j)\in p}\tau_{n_i,m_j}(\underline{u})\,\mathrm{e}^{-\psi_b(\underline{u})}\,\mathrm{d}\underline{u},$$

the vector (U_1, U_2), whose density on $\mathbb{R}_+^2$ is proportional to $u_1^{n-1}u_2^{m-1}\prod_{(i,j)\in p}\tau_{n_i,m_j}(\underline{u})e^{-\psi_b(u)}$, the variables $\{Z_i^x\}_i$, whose distribution is $P_{X_i^*}(\cdot)$, for any $i = 1, \dots k$, and $\{Z_j^y\}_j$, whose distribution is $P_{Y_j^*}(\cdot)$, for any $j = 1, \dots, c$. We are now ready to state the key posterior characterization.

Theorem 13. *Let$(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$ be from model* (26)*, with Q being the law of a n-FuRBI. Then, the distribution of $(\tilde{\mu}_1, \tilde{\mu}_2)$ conditional on $(X_i)_{i=1}^n$, $(Y_j)_{j=1}^m$ and the set of latent variables $(p, U_1, U_2, \{Z_i^x\}_i, \{Z_j^y\}_j)$ is*

$$(\hat{\mu}_1, \hat{\mu}_2) + \sum_{(i,j)\in\Delta_p} J_{i,j}\delta_{\left(X_i^*,Y_j^*\right)} + \sum_{(i,j)\in\Delta_p^1} J_{i,0}\delta_{\left(X_i^*,Z_i^x\right)} + \sum_{(i,j)\in\Delta_p^2} J_{0,j}\delta_{\left(Z_j^y,Y_j^*\right)},$$

where $(\hat{\mu}_1, \hat{\mu}_2)$ is a completely random vector with intensity $\mathrm{e}^{-U_1s_1-U_2s_2}\rho(\mathrm{d}s_1, \mathrm{d}s_2)G_0(\mathrm{d}x)$ and $J_{i,j} = (J_{i,j}^1, J_{i,j}^2)$, with $i = 0, \dots, k$ e $j = 0, \dots, c$, are jumps with density proportional to $s_1^{n_i}s_2^{m_j}e^{-U_1s_1-U_2s_2}\rho(\mathrm{d}s_1, \mathrm{d}s_2)$. Moreover $(\hat{\mu}_1, \hat{\mu}_2)$ and $J_{i,j}$ are independent.

Conditional on the latent variables, the structure is quite intuitive: the posterior is the law of a completely random vector with modified intensity and fixed locations, given by the pairs formed by the hyper-ties. This is somehow reminiscent of the posterior structures of exchangeable models (James et al., 2009; Lijoi and Prünster, 2010), with the key novelty played by the new notion of hyper-ties, in addition to the identification of a suitable latent structure.

The distribution of the latent variables admits a nice interpretation. For instance, the mass function of the latent structure p is the product of two terms: the probability of observing the number of hyper-ties identified by p times the likelihood that exactly those pairs are formed, through the density function g_0. Thus, thanks to the homogeneity of the original completely random vector, we observe a separate effect for jumps and locations on this hidden clustering structure. The next corollary shows how the posterior distribution of the normalized measures can be deduced from Theorem 13. The statement focuses on $\tilde{P}_1$, though an analogous representation holds also for $\tilde{P}_2$.

Corollary 6. *Under the same assumptions of* Theorem 13*, conditional on $(X_i)_{i=1}^n$, $(Y_j)_{j=1}^m$ and the latent variables $(p, U_1, U_2, \{Z_i^x\}_i, \{Z_j^y\}_j)$, the random probability measure $\tilde{P}_1$ in* (3.3) *equals in distribution*

$$w_1\frac{\hat{\mu}_1}{T_1} + w_2\frac{\sum_{(i,j)\in\Delta_p} J_{i,j}^1\delta_{\left(X_i^*,Y_j^*\right)}}{\sum_{(i,j)\in\Delta_p} J_{i,j}^1} + w_3\frac{\sum_{(i,j)\in\Delta_p^1} J_{i,0}^1\delta_{\left(X_i^*,Z_i^x\right)}}{\sum_{(i,j)\in\Delta_p^1} J_{i,0}^1} + w_4\frac{\sum_{(i,j)\in\Delta_p^2} J_{0,j}^1\delta_{\left(Z_j^y,Y_j^*\right)}}{\sum_{(i,j)\in\Delta_p^2} J_{0,j}^1},$$

where $T_1 = \hat{\mu}_1(\mathbb{X}\times\mathbb{X})$, while

$$w_1 \propto T_1, \quad w_2 \propto \sum_{(i,j)\in\Delta_p} J_{i,j}^1, \quad w_3 \propto \sum_{(i,j)\in\Delta_p^1} J_{i,0}^1, \quad w_4 \propto \sum_{(i,j)\in\Delta_p^2} J_{0,j}^1,$$

with the constraint $\sum_{i=1}^{4} w_i = 1$*.*

Predictive structure

Prediction of new observations arises naturally within the Bayesian framework, since it coincides with the estimate of the distribution under a square loss function. Moreover, it has the merit of providing intuition on how the model behaves and learns and it can be used to develop marginal algorithms that avoid the direct sampling of P_1 and P_2, which are infinite-dimensional objects. In Proposition 15 we saw how to sample the first pair of observations. The next result tackles the general case.

Theorem 14. *Consider samples* $(X_i)_{i=1}^n$ *and* $(Y_j)_{j=1}^m$ *from model* (26)*, with the same setting of Theorem 13. Then there exist probability weights* ξ_0*,* $\{\xi_i^x\}$ *and* $\{\xi_j^y\}$ *such that*

$$\mathbb{P}\Big(X_{n+1} \in C \mid (X_i)_{i=1}^n, (Y_j)_{j=1}^m\Big) = \xi_0 P_0(C) + \sum_{i=1}^{k} \xi_i^x \delta_{X_i^*}(C) + \sum_{j=1}^{c} \xi_j^y P_{Y_j^*}(C)\,.$$

Analogously, there exist probability weights η_0*,* $\{\eta_i^x\}$ *and* $\{\eta_j^y\}$ *such that for any* $C \in \mathcal{X}$

$$\mathbb{P}\Big(Y_{m+1} \in C \mid (X_i)_{i=1}^n, (Y_j)_{j=1}^m\Big) = \eta_0 P_0(C) + \sum_{j=1}^{c} \eta_j^y \delta_{Y_j^*}(C) + \sum_{i=1}^{k} \eta_i^x P_{X_i^*}(C)\,.$$

Explicit formulae for the weights are available in the proof of Theorem 14, in Section A1. In specific cases they can be computed in closed form, conditional to the latent variables: see e.g. example 10 in Section A2 for the Inverse Gaussian case with equal jumps.

Hence, the marginal predictive distributions have a quite intuitive form: they are linear combinations of the centering distribution Q_0, a weighted version of the empirical distribution and a last term that depends on the other sample. The crucial differences with respect to prediction rules arising in the exchangeable case (Lijoi and Prünster, 2010; De Blasi et al., 2015) is the addition of the last term, which clearly shows how posterior inference changes when incorporating heterogeneous information and performing borrowing of information.

Example 6 (n-FuRBI with equal atoms)**.** *If the joint distribution* G_0 *is degenerate such that the atoms are completely shared between* P_1 *and* P_2*, then* $P_Z(\cdot) = \delta_Z(\cdot)$*. Therefore, the last term in Theorem 14 becomes a weighted version of the empirical distribution relative to the other sample.*

Algorithms for posterior inference and prediction are derived in Section A2.

3.2.6 Numerical Illustrations and Real Data Analyses

Bayesian mixture models

Discrete Bayesian models, as the one specified in (26), are usually not employed directly on the data, but as a building block in hierarchical mixture models: in this setting X and Y are hidden values that describes the clustering structure within the data. Such models have been introduced by Lo (1984) for the Dirichlet processes and gained popularity thanks also to the availability of sampling methods for posterior inference (Escobar and West, 1995; Ishwaran and James, 2001; Neal, 2000). Suppose $\{f(\cdot \mid x) : x \in \mathbb{X}\}$ is a family of probability density kernels on a space $\mathbb{W}$. Then the model can be formulated as

$$\begin{matrix} W_i \mid X_i \overset{\text{ind}}{\sim} f(\cdot \mid X_i) \\ X_i \mid P_1 \overset{\text{i.i.d.}}{\sim} P_1 \end{matrix}, \quad \begin{matrix} V_j \mid Y_j \overset{\text{ind}}{\sim} f(\cdot \mid Y_j) \\ Y_j \mid P_2 \overset{\text{i.i.d.}}{\sim} P_2 \end{matrix}, \quad (P_1, P_2) \sim \text{n-FuRBI}.$$

where $(W_i)_{i=1}^n$ and $(V_j)_{j=1}^m$ are the observable samples and are assumed to be conditionally independent, given $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$. Integrating out the latent variables $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$, the data are random draws from suitable countable mixtures, i.e.

$$W_i \mid P_1 \overset{\text{i.i.d.}}{\sim} \int f(\cdot \mid x)\, P_1(\mathrm{d}x), \quad V_j \mid P_2 \overset{\text{i.i.d.}}{\sim} \int f(\cdot \mid y)\, P_2(\mathrm{d}y).$$

Example 7 (Gaussian mixtures)**.** *We assume $f(\cdot \mid x) := N(\cdot \mid x, \sigma^2)$, with σ^2 positive known constant, to be the normal density. Thus, the latent parameter is the mean, i.e. $\mathbb{X} = \mathbb{R}$. In this case $Cov(X_i, Y_j) = Cov(W_i, V_j)$, so that the joint behavior of the latent means is reflected on the observations: this shows the importance of the correlation structure given by Proposition 14 also for hierarchical models. Alternatively, the latent parameters could specify both the mean and the variance, with $\mathbb{X} = \mathbb{R} \times \mathbb{R}_+$.*

The goal is then to draw samples from the posterior distribution given $(W_i)_{i=1}^n$ and $(V_j)_{j=1}^m$: however this requires to integrate out all the possible partitions of the $n + m$ latent variables. As detailed in Section A2., it is possible to devise a Gibbs sampler for drawing from the posterior distribution of $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$. Once a posterior sample $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$ is generated, relevant quantities of interest can be approximated by exploiting the conditional independence of $(W_i)_{i=1}^n$ and $(V_j)_{j=1}^m$, given the latent variables.

Simulation study for density estimation

We consider a simple application with simulated data, in order to understand how inference changes when taking into account heterogeneous sources of information. Assume the following generating mechanism: $W_i \overset{\text{i.i.d.}}{\sim} N(\cdot \mid 10, 1)$, for $i = 1, \dots, 20$, and $V_j \overset{\text{i.i.d.}}{\sim} N(\cdot \mid -10, 1)$, for $j = 1, \dots, 100$. Supposing only the phenomenon associated to the first sample is of interest, hierarchical mixtures are considered to make prediction on the unknown density of W_i. The kernel considered is the one specified in Example 7, with known $\sigma^2 = 1$ and latent mean μ. Four different approaches for modelling dependence between $(W_i)_{i\geq 1}$ and $(V_i)_{i\geq 1}$ are devised: the exchangeable approach, according to which sequences W and V are supposed to form one exchangeable sequence, inducing the highest positive correlation between W_i and V_j; the independent approach, according to which the sample $(V_i)_{i\geq 1}$ is disregarded entirely, that is $(W_i)_{i\geq 1}$ and $(V_i)_{i\geq 1}$ are treated independently; the hierarchical approach, where we use a hierarchical Dirichlet process (see Example 1) that corresponds to a classical borrowing of information; the FuRBI approach, where the underlying random probability measures $\tilde{p}_1$ and $\tilde{p}_2$ are n-FuRBI with equal weights and the distribution on the atoms is $G_0(\cdot \mid \rho_0) = N_2\left(\cdot \mid \underline{0}, 1, \rho_0\right)$ with $\rho_0 \sim \text{Unif}([-1, 1])$, where $N_2(\cdot \mid \underline{m}, \sigma_0^2, \rho_0)$ denotes the bivariate normal distribution with mean vector $\underline{m}$, common variance σ_0^2 and correlation ρ_0. It can be proven that under this specification $\text{Corr}(W_i, V_j) = 0$, so that a priori W and V are marginally uncorrelated. The prior specification is purposely simple, especially regarding the base measure and the concentration parameter, in order to single out the effect of the borrowing between the two groups as much as possible.

For the first two cases and the n-FuRBI, the marginal distribution is given by a Dirichlet process with $\theta = 1$ and $Q_0(\cdot) = N(\cdot \mid 0, 1)$; instead for the hierarchical process the concentration parameters are fixed in order to match the expected number of different clusters with the other methods, for a fair comparison. As highlighted in Example 5, n-FuRBI with equal jumps lead to the most general setting in terms of achievable correlation between samples; moreover, choosing the marginal processes to derive from a Gamma process, we can achieve any value in the interval $(-1, 1)$, tuning appropriately the concentration parameter θ.

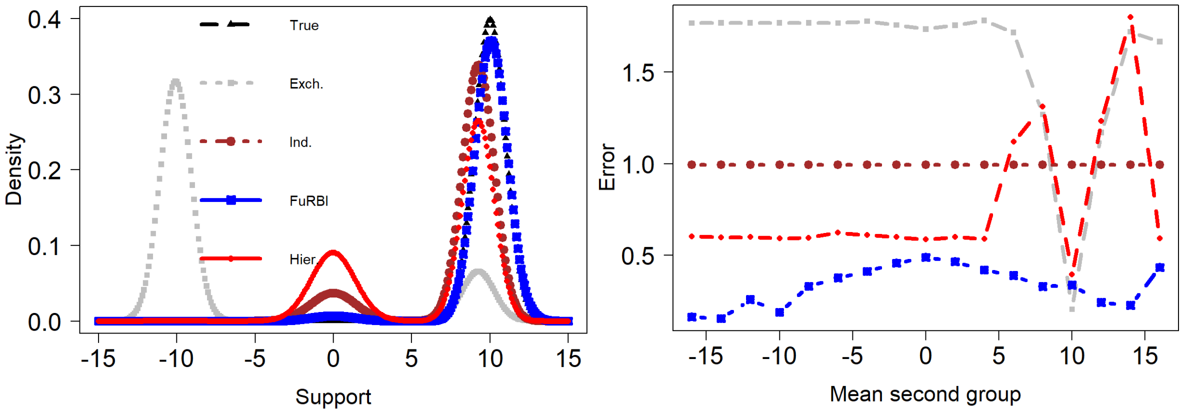


Figure 3.1: Left: mean posterior densities for the case with opposite true means. Right: mean integrated error (computed on a grid and as the median over 50 different samples) for the four estimates, varying the true mean of V.

The left panel of Figure 3.1 shows the performances of the four methods, after the application of the blocked Gibbs sampler provided in the supporting material: the mean posterior density (computed pointwise) is depicted. The exchangeable approach behaves very badly, as expected, because the two samples have clearly a different distribution. The independent choice leads to a reasonable estimate, even if it still overestimates the probability mass around the prior mean (because of the small sample size of the first sample). The hierarchical estimate is quite good, but our proposal, instead, fits almost perfectly the target density and seems to exploit the opposite behaviour of the two phenomena: this is clearly highlighted by the posterior distribution of ρ_0, whose approximated mean is close to -0.9.

One may wonder whether these superior performances follow from the precise specification above, with opposite true means. Therefore, we repeated the experiment keeping the same generating mechanism for W, but with the true mean of V ranging in the set $\{-16, -14, \ldots, 14, 16\}$: the mean integrated absolute error (computed on a grid and as the median over 50 different samples) is depicted in the right panel of Figure 3.1. It is apparent that the FuRBI approach almost always yields the smallest error, regardless of the true value. Its performance is close to the exchangeable case only when the two true means are equal, that is when exchangeability actually holds; analogously, the n-FuRBI priors yield the highest error when the mean of V corresponds to the prior mean, i.e., when the other group provides less additional information. The hierarchical process captures the right dependence when the two means coincide, but can be misled when they are close; finally, when the second sample is very far from the first one it performs better than the independent model, probably thanks to the different inner clustering structure. The results are also summarized in Table 3.1. Thus, n-FuRBI seem to be always capable of combining heterogeneous information in the right way; in particular, at least in this

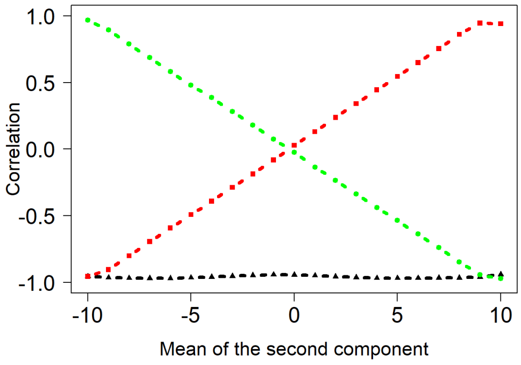


Figure 3.2: Posterior median of the correlation (obtained through 100 simulation studies) between the three unknown means. Black with triangular shapes: correlation between the first and third component. Red with square shapes: correlation between the first and second components. Green with circular shapes: correlation between the second and third component.

Mean of V	Exch.	Ind.	FuRBI	Hier.
-16	1.769	0.995	**0.163**	0.604
-10	1.769	0.995	**0.189**	0.592
0	1.737	0.995	**0.489**	0.587
10	**0.205**	0.995	0.338	0.397
16	1.666	0.995	**0.435**	0.592

Table 3.1: Mean integrated absolute error associated to the four methods for some values of the mean of V. The values in bold are the smallest ones for each row.

example, they recognize the most useful type of borrowing of information. In Section S5.1 similar experiments are conducted, using different data generating distributions: they show that the conclusions hold even when the data display significantly different features, as multimodality or heavy tails.

Finally, we consider a similar application with three groups, in order to see whether n-FuRBI are able to discern more complex types of dependence. We assume to observe $W_{1,i} \overset{\text{i.i.d.}}{\sim} N(\cdot \mid 10, 1)$, $W_{2,i} \overset{\text{i.i.d.}}{\sim} N(\cdot \mid x, 1)$, and $W_{3,i} \overset{\text{i.i.d.}}{\sim} N(\cdot \mid -10, 1)$, where $i = 1, \dots, 20$ and $x \in \{-10, -9, \dots, 10\}$. Then, for each value of x we apply the same n-FuRBI with the same weights described above, but where the atoms are distributed according to

$$G_0(\cdot) = N_3\left(\cdot \middle| \underline{0}, 1, \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix}\right),$$

where $N_3(\cdot \mid \mu_0, \sigma^2, \Psi)$ denotes a multivariate normal distribution with mean μ_0, all the variances equal to σ^2 and correlation matrix Ψ and $\rho_{12}, \rho_{13}, \rho_{23} \overset{\text{i.i.d.}}{\sim} \text{Unif}([-1, 1])$. The posterior medians of ρ_{12}, ρ_{13} and ρ_{23} are depicted in Figure 3.2, for any value of x. The results are in line with our intuition: the correlation between the first and second component is always close to -1 (indeed they have opposite behaviour relative to the prior), while ρ_{13} and ρ_{23} vary linearly with x, being positive when the means have the same sign.

Predicting stocks and bonds returns

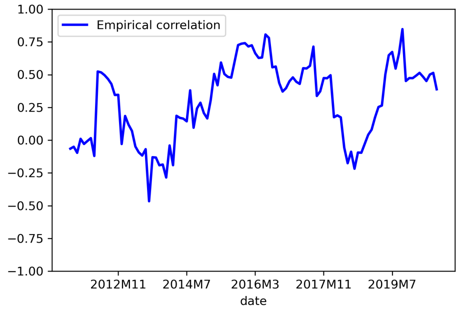


Figure 3.3: Empirical correlation between average stock return and average commodity return computed on a moving window of 12 months using data from March 2011 to January 2021.

Findings from the previous section and Section A2 suggest that n-FuRBI may be used to enhance density estimates and prediction in multi-sample data. Here, the performance is showcased on a real dataset of stocks and bonds returns. We collected monthly returns of January 2021 for a sample of 49 stocks portfolios from the Kenneth R. French's Data Library (data available at `http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`) and for a sample of 55 commodities from the Primary Commodity Prices Database of the International Monetary Fund (data available at `https://www.imf.org/en/Research/commodity-prices`).

We employ a Bayesian mixture model and assume that stock and bonds returns, denoted by W_i and V_j, respectively, are sampled from mixtures of normals where the mixing distributions act on mean and variance of the kernel, i.e.,

$$W_i \mid P_1 \overset{\text{i.i.d.}}{\sim} \int N(\cdot \mid x, \sigma_w^2)\, P_1(\mathrm{d}x, \mathrm{d}\sigma_w^2) \qquad V_j \mid P_2 \overset{\text{i.i.d.}}{\sim} \int N(\cdot \mid y, \sigma_v^2)\, P_2(\mathrm{d}y, \mathrm{d}\sigma_v^2).$$

Stocks and commodities exhibit correlation that largely varies over time ranging from positive to negative values (see, for instance, Bhardwaj and Dunsby, 2013, and Figure 3.3). As consequence, commodities returns contain useful information to make inference over the distribution of stocks portfolios, and viceversa. Thus, borrowing of information represents a natural strategy to improve inference. However, returns may differ even largely in value between the two sets of financial instruments, especially in periods of negative correlation. For instance, in our dataset, 53% of the observed stocks returns are negative, while only 16% of the bonds returns have negative sign. As such, classical nonparametric borrowing, consisting in sharing of mixture components, is not appropriate and, as shown in the following, possibly harmful. We instead make use of n-FuRBI models as prior distribution, i.e.,

$$\begin{aligned}(P_1, P_2) \mid \theta, z, G_0 &\sim \text{n-FuRBI}(\theta, \rho, G_0)\\ \theta &\sim \text{Gamma}(\alpha, \beta)\end{aligned}$$

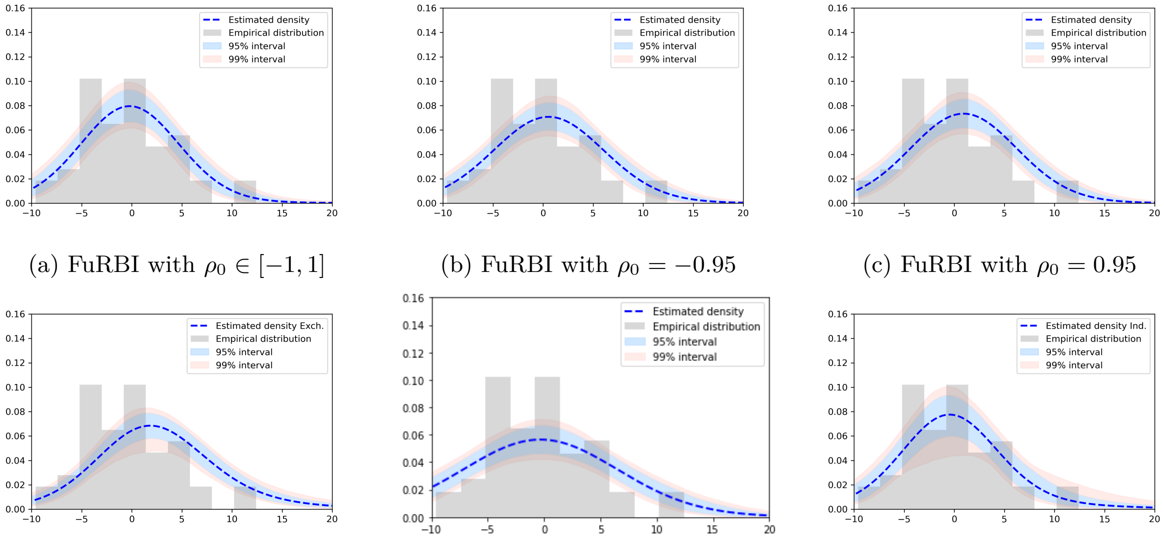
The base measure G_0 is chosen so that marginal distributions are given by normalized CRMs with conjugate Normal-InverseGamma base measure, i.e.

$$\begin{aligned}G_0(\mathrm{d}_x, \mathrm{d}_y, \mathrm{d}\sigma_w^2, \mathrm{d}\sigma_v^2 \mid \rho_0) = {}& N_2(\mathrm{d}x, \mathrm{d}y \mid m, \Sigma(\lambda_1, \lambda_2, \sigma_w^2, \sigma_v^2\rho_0))\\ & \times \text{InvGamma}(\mathrm{d}\sigma_w^2 \mid \alpha_1, \beta_1) \times \text{InvGamma}(\mathrm{d}\sigma_v^2 \mid \alpha_2, \beta_2)\end{aligned}$$

with

$$m = (m_1, m_2)' \qquad \text{and} \qquad \Sigma = \begin{bmatrix} \frac{\sigma_w^2}{\lambda_1} & \rho_0 \, \frac{\sigma_w}{\lambda_1^{1/2}} \frac{\sigma_v}{\lambda_2^{1/2}} \\ \rho_0 \, \frac{\sigma_w}{\lambda_1^{1/2}} \frac{\sigma_v}{\lambda_2^{1/2}} & \frac{\sigma_v^2}{\lambda_2} \end{bmatrix}$$

and we use the following joint underlying Lévy intensity $v(\mathrm{d}s_1, \mathrm{d}s_2, \mathrm{d}x_1, \mathrm{d}x_2) = \{z\,[\rho(\mathrm{d}s_1)\delta_0(\mathrm{d}s_2) + \rho(\mathrm{d}s_2)\delta_0(\mathrm{d}s_1)] + (1-z)\,\rho(\mathrm{d}s_1)\delta_{s_1}(\mathrm{d}s_2)\}\,\theta\, G_0(\mathrm{d}x_1, \mathrm{d}x_2)$, with $z \sim \text{Unif}([0,\,1])$. We term the re-



(a) FuRBI with $\rho_0 \in [-1, 1]$

(b) FuRBI with $\rho_0 = -0.95$

(c) FuRBI with $\rho_0 = 0.95$

(d) Exchangeable model

(e) GM-dependent model

(f) Independent model

Figure 3.4: Posterior density estimates for stocks returns.

sulting n-FuRBI *additive n-FuRBI*, since the series representation of the corresponding FuRBI CRMs is

$$\mu_1(\cdot) \overset{a.s.}{=} \sum_{k\geq 1} W_k \delta_{\theta_{0,k}} + \sum_{k\geq 1} J_k \delta_{\theta_{1,k}} \qquad\qquad \mu_2(\cdot) \overset{a.s.}{=} \sum_{k\geq 1} W_k \delta_{\phi_{0,k}} + \sum_{k\geq 1} V_k \delta_{\phi_{2,k}},$$

where $(\theta_{0,k}, \phi_{0,k}) \overset{i.i.d}{\sim} G_0$, $\theta_{1,k} \overset{i.i.d}{\sim} P_0$ and $\phi_{2,k} \overset{i.i.d}{\sim} P_0$. When G_0 is degenerate on the main diagonal (i.e. $\rho_0 = 1$), one retrieves GM-dependent completely random measures (Lijoi et al., 2014a,b; Lijoi and Nipoti, 2014). In order to obtain two Dirichlet processes marginally we set $\rho(s) = s^{-1}e^{-s}$, so that $\beta = 1/1 + \theta$ and $\gamma = (1-z)\,{}_3F_2(\theta - \theta\, z + 2,\, 1,\, 1;\, \theta + 2,\, \theta + 2;\, 1)\theta/(1+\theta)^2$, where ${}_3F_2$ is the generalized hypergeometric function.

The Bayesian paradigm requires the elicitation of a prior guess about the phenomenon, by tuning the hyperparameters of the model. In particular, we set the a priori expectations m_1 and m_2 in the two groups equal to the empirical averages of the two groups in December 2020, i.e., the month preceding the data collection, leading to $m_1 = 5.8591$ and $m_2 = 3.9731$. In the following, we say that a financial instrument is *outperforming* if its observed return is higher than its a priori expected value. In order to assign ρ_0, we use the results of Propositions 14 and 15. The elicited ρ_0 should reflect our prior opinion about the correlation, which means that it should induce a learning mechanism agreeing with the following principle: under positive/negative correlation, conditioning on the event of outperforming commodities, the prior probability of outperforming/underperforming stocks should increase. Prior opinion about the correlation can be formulated working with financial experts and, thanks to n-FuRBI, incorporated through an informative prior on the parameter ρ_0. Here, we consider three scenarios: in the first and second, we derive inferential results under a prior opinion of negative and positive correlation, respectively, while in the third scenario we assume that no information on the correlation is available. The three scenarios are obtained with, respectively, $\rho_0 = 0.95$, $\rho_0 = -0.95$, and using a uniform prior on ρ_0. After standardizing the data, we set the remaining hyperparameters in a weakly informative way, i.e. $\lambda_1 = \lambda_2 = 1$, $\alpha_1 = \alpha_2 = 2$, and $\beta_1 = \beta_2 = 4$. Sensitivity analysis, carried out in Section A2, shows that results are robust with respect to different choices for λ_j, α_j and β_j for $j = 1, 2$. We perform $50,000$ iterations of the marginal algorithm (Section A2) and discard the first $10,000$ as burn–in.

Finally, we compare our approach with three alternative models: the independent model and the exchangeable model, described in the previous section, and the GM-dependent model from Lijoi et al. (2014b),which performs classical borrowing based on ties and shares the same addictive structure of additive n-FuRBI.

	ALCPO	MLCPO
FuRBI $\rho_0 \in [-1, 1]$	**-1.2347**	**-0.9627**
FuRBI $\rho_0 = -0.95$	-1.2925	-1.0115
FuRBI $\rho_0 = 0.95$	-1.2896	-1.0149
Exch	-1.5024	-1.1521
GM-dep	-1.4864	-1.1557
Ind	-1.3495	-1.1017

Table 3.2: ALCPO and MLCPO under the three models. Best performance is highlighted in bold.

Figure 3.4 displays the posterior density estimates for stocks returns. The analogous figure for bonds returns can be found in Section A2. Models employing additive n-FURBI produce density estimates that better resemble the empirical distribution. The best performance is attained with a non-informative prior over the correlation ρ_0: this is probably due to the fact that the intensity and direction of the borrowing of information concentrate on the optimal value for the dataset. The FuRBI models with fixed ρ_0 perform worse compared to full-borrowing; nonetheless, thanks to their flexibility, they still produce better results than other competitors. The GM-dependent and the exchangeable models yield the worst density estimates in terms of resemblance of the histogram, as expected. Indeed, the type of borrowing they perform, based on ties, is not appropriate for the specific problem. Lastly, we note that the independent model appears to provide a reasonable density estimation, but presents significantly higher uncertainty. While Figure 3.4 provides insight on the model performance, an important caveat is in order: a too close resemblance of the empirical distribution may indicate overfitting. Note moreover that, given the low numerosity of the samples, the histogram is very much influenced by few observations unlike the density estimates: since this is due by the presence of a prior, a more refined analysis should include different choices of the baseline measure in order to assess the impact on the final estimate.

To evaluate the predictive performance, we resort to the conditional predictive ordinates (CPOs) statistics (see, e.g. Gelfand et al., 1992; Barrios et al., 2013). Essentially, for each value i, we train the model without the i-th observation and compute the predictive density at the observed point. For the first sample it reads $\text{CPO}_i^w = \tilde{f}(w_i \mid w^{-i}, v)$, for $i = 1, \ldots, n$ and analogously for the second sample we have $\text{CPO}_j^v = \tilde{f}(v_j \mid w, v^{-j})$, for $j = 1, \ldots, m$, where w and v denote the vectors of observed returns for, respectively, stocks and commodities. Table 3.2 displays the average logarithmic CPO (ALCPO) and the median logarithmic CPO (MLCPO) in the overall sample. Higher values correspond to a better performance, and the n-FuRBI exhibits the best performance.

Clustering of multivariate data with missing entries

We now show how to leverage on our methodology to perform borrowing of information and clustering with multivariate data affected by missing entries. The n-FuRBI priors are very well suited for this problem: indeed, incomplete observations can be interpreted as projections of latent complete observations and, in particular, hyper-ties between incomplete observations can be thought of as actual ties between complete observations.

We consider a P-variate ($P > 1$) dataset with missing entries and divide the dataset into distinct samples based on the missing entries: denote by $(\underline{W}_i^{(j_1,\dots,j_l)}, i = 1, \dots, n_{(j_1,\dots,j_l)})$ the sample where l outcomes with labels $(j_1, \dots, j_l)$ are missing. The dimension of the vector $\underline{W}_i^{(j_1,\dots,j_l)}$ is therefore $P_{j_1,\dots,j_l} = P - l$. Denote by $q_{j_1,\dots,j_l}$ the corresponding unknown distribution, i.e.,

$$\underline{W}_i^{(x)} \mid q_x \overset{\text{i.i.d.}}{\sim} q_x \qquad \text{for } i = 1, \dots, n_x \text{ and } x \in I,$$

where I is the index set of all the possible combinations of missing variables identifying different samples, which are at most $2^P - 1$. Independent analyses for each sample should clearly be avoided and classical nonparametric borrowing cannot even be specified because the support spaces of different samples differ one from the other.

To perform clustering, we assume that each q_x is a mixture of multivariate normal kernels with diagonal covariance matrix and mixing measure $\tilde{p}_x$ on locations, i.e.

$$\underline{W}_i^{(x)} \mid P_x, \underline{\sigma}^2 \overset{\text{i.i.d.}}{\sim} \int N_{P_x}(\cdot \mid \underline{\mu}_x, \underline{\sigma}_x^2)\, P_x(\mathrm{d}\underline{\mu}_x),$$

where $\underline{\sigma}^2 = \left(\sigma_1^2, \dots, \sigma_P^2\right)$, $\underline{\sigma}_x^2$ is the restriction of $\underline{\sigma}^2$ to all the elements besides x and $N_K(\cdot \mid \underline{\mu}, \underline{\tau}^2)$ denotes the K-variate normal distribution with mean vector $\underline{\mu}$ and diagonal covariance matrix given by $\underline{\tau}^2$. Independence of the kernel (implied by the diagonal covariance matrix) is a common assumption in clustering models for multivariate responses (see, for instance, Gao et al., 2020; Franzolini et al., 2023): in this way we are forcing the clustering structure to encode all the dependence across responses. The P_x are distributed as

$$(P_x, x \in I) \sim \text{additive n-FuRBI},$$

described in the last Section. The atoms of $(P_x, x \in I)$ are costrained so that an hyper-tie

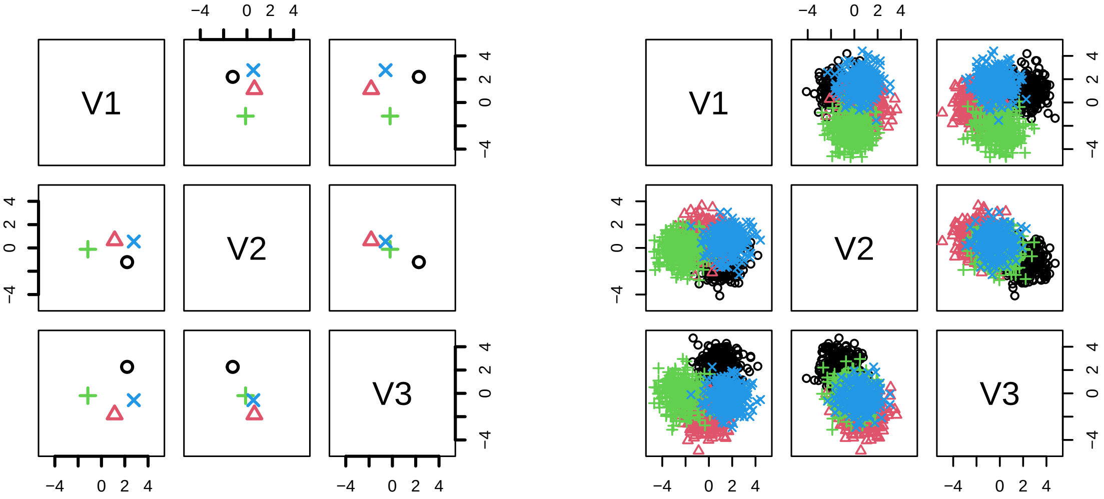


Figure 3.5: Simulated data: left panel shows true clusters locations, right panel shows complete simulated data for $n = 1000$ before applying the missingness mechanisms.

simul number	missing mechanism	% of missing entries	n-FuRBI $z = 0.2$	n-FuRBI $z = 0.5$	n-FuRBI $z = 0.8$	mice + k-means	mice + DPM
n.1	MCAR	16.1%	**0.7883**	0.7882	0.7881	0.7408	0.7734
n.2	MNAR	16.7%	0.7703	0.7704	**0.7706**	0.6323	0.7617
n.3	MCAR	35.9%	**0.7292**	0.7285	0.7283	0.6786	0.7165
n.4	MNAR	34%	0.7304	0.7301	**0.7432**	0.6391	0.7328

Table 3.3: Rand indexes for 5 competing methods: 3 n-FuRBI models with varying parameter z, mice+k-means and mice+DPM. For n-FuRBI and mice+DPM the posterior expected value is computed averaging over the Rand indexes of all clustering configurations visited by the MCMC chain after burn-in.

simul number	missing mechanism	% of missing entries	n-FuRBI $z = 0.2$	n-FuRBI $z = 0.5$	n-FuRBI $z = 0.8$	mice + k-means	mice + DPM
n.1	MCAR	16.1%	4.24	**4.19**	4.22	3	5.48
n.2	MNAR	16.7%	**4.59**	3.29	3.37	2	5.36
n.3	MCAR	35.9%	4.38	**4.18**	4.20	3	7.01
n.4	MNAR	34%	4.28	**4.17**	4.59	2	5.85

Table 3.4: Estimated number of clusters for 5 competing methods. The posterior mean is used for n-FuRBI and mice+DPM, while the number of clusters is selected by maximizing the average silhouette for mice+k-means. The true number of clusters is equal to 4.

can be interpreted as an actual tie between complete observations: moreover the choice of dependent weights allows to recover group-specific features, if the missingness mechanism is informative. Section A2 provides a discussion of this and contains the details about the choice of the hyperparameters.

First, we conduct a simulation study where data for $n = 1,000$ items, $P = 3$ responses, and $K = 4$ clusters are simulated from a mixture of Gaussian distributions. Figure 3.5 shows the locations of the true clusters and the complete simulated data before deleting entries. Then, different missingness mechanisms are applied to determine the entries to be treated as missing. Missing completely at random (MCAR) scenarios are obtained by sampling missing entries uniformly, while, in missing non at random (MNAR) scenarios the probability of being missing depends on the true cluster allocation. Different combinations of missing variables define different samples: the number of samples ranges from 3 to 6 among simulation scenarios. The detailed distributions of missing values are provided in Section A2. Different values of the hyperparameter z of the Lévy intensity are considered. Our results are compared with those obtained with two alternative approaches, called “mice + k-means" and “mice + DPM", which follow a two-steps procedure: first one imputes missing data by chained equations as implemented in the `R` package `mice` (van Buuren and Groothuis-Oudshoorn, 2011), then, the clustering structure is estimated with, respectively, k-means and a Dirichlet process mixture. Note that the number of clusters for k-means is chosen to maximize the average silhouette. For each run of the n-FuRBI model, we perform $25,000$ iterations of the MCMC chain and discard the first half as burn-in. Tables 3.3 and 3.4 summarize the performance of the models. The n-FuRBI priors outperform the alternatives in all scenarios considered, in term of estimating both the number of clusters and the clustering configuration, measured by Rand indexes between the estimated configuration and the true clustering structure. Moreover, the posterior distribution of n-FuRBI models

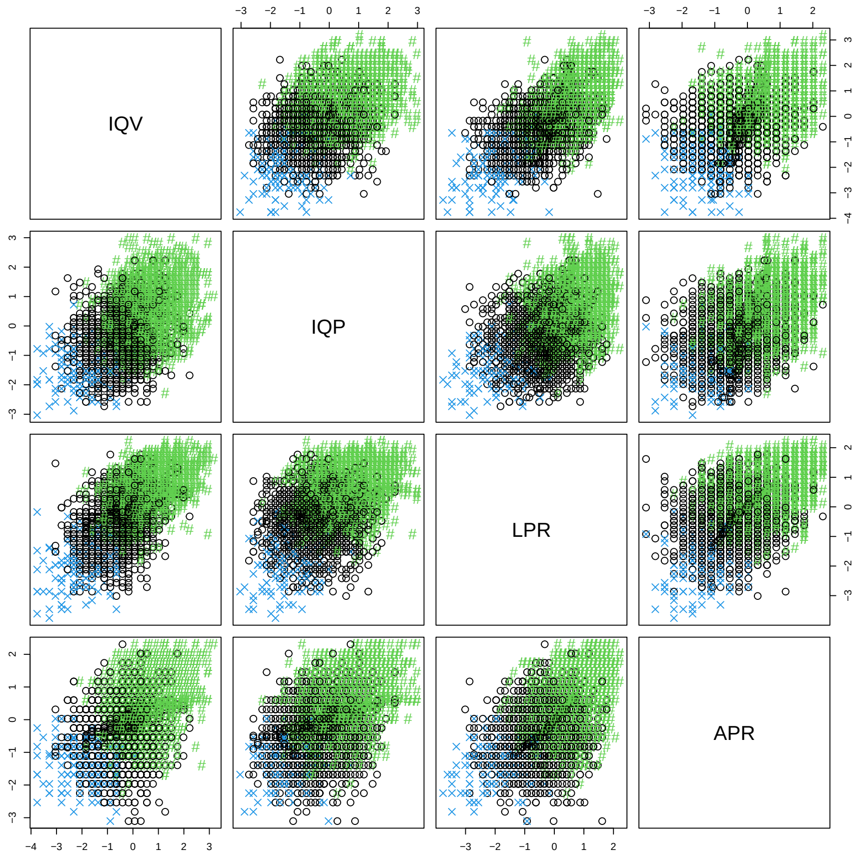


Figure 3.6: Scatter plots of the four scores (after standardization) for the brandsma dataset. Coordinates of missing data are set equal to their respective posterior median. Different colors and symbols denote the three estimated clusters obtained minimizing the variation of information loss with respect to the posterior distribution.

reflects uncertainty both about the estimated clustering configuration and about the imputation mechanism, which is instead ignored by two-steps procedures.

Finally, we apply the model also on the brandsma dataset (Snijders and Bosker (2012)), which refers to grade 8 students (age about 11 years) in elementary schools in the Netherlands (see, Brandsma and Knuver, 1989). The goal is to cluster $n = 4,106$ pupils, based on their IQ verbal score (IQV), IQ performance score (IQP), language score (LRP), and arithmetic score (APR). The number of subjects presenting missing entries is 339 out of 4,106 (i.e., 8.26%). As before, different combinations of missing variables define different samples: the number of samples is 7 in the brandsma dataset. In this real data analysis, the final clustering configuration provides a lower dimensional description of the data rather than an estimate of ideal true clusters. Data are standardized before running the model, so that the sample means and variances are equal to 0 and 1. Figure 3.6 shows the estimated clustering configuration obtained minimizing the variation of information loss with respect to the posterior distribution. The model identifies three clusters, which show as major tendency that groups of students performing above/below average for one of the four scores tends to perform above/below average also for the other scores. In particular, a first cluster includes 53% of the subjects, which have lower performances: indeed cluster averages of the standardized scores are IQV= -0.371, IQP= -0.398, LRP= -0.387, and APR= -0.435. Instead the second cluster, including 44% of the subjects, retains the best students: the cluster averages of the standardized scores are IQV= 0.609, IQP= 0.595, LRP= 0.629, and APR= 0.642. Finally, the students with worst scores are allocated to a third cluster whose averages are IQV= -2.01, IQP= -1.43, LRP= -1.90, and APR= -1.34.

3.2.7 Conclusion

We investigated the dependence induced across groups in a wide class of Bayesian nonparametric models, introducing the notion of hyper-tie. We showed how hyper-ties play a crucial role in driving the Bayesian learning mechanism and the borrowing of information across samples. We noted that existing nonparametric priors either do not allow an explicit evaluation of the value of the correlation or, when they do, they are able to induce only non-negative correlation. Thus we designed n-FuRBI, a novel class of dependent nonparametric priors, which may induce either positive or negative correlation between the random probabilities as well as across samples introducing a novel and flexible idea of borrowing of strength. This allows to achieve high flexibility as well as analytical tractability, while outperforming competing models in different scenarios. Our class of priors is immediately applicable to model multi-sample data through mixture models, as shown in the analysis of the financial dataset. Moreover, it allows also for a variety of interesting extensions since it can be seen as an effective building block to model non trivial dependencies in more complex data analyses as showcased in a clustering problem with mutivariate data in presence of missing entries.

3.3 Trees of random probability measures

3.3.1 Introduction

In the nonparametric setting, a common choice for the prior distribution is the law of a Dirichlet process (Ferguson, 1973), or suitable generalizations such as Pitman-Yor process (Pitman and Yor, 1997; Pitman, 2006) or processes derived from completely random measures (Kingman, 1967; James et al., 2006, 2009). In a partially exchangeable setting, a common approach is to combine distributions as above, in order to induce various types of dependence between groups: this leads to additive structures (Lijoi et al., 2014a), nested structures (Rodriguez et al., 2008) and hierarchical structures.

The latter, that are the starting point of this Section, work by creating a hierarchy of random measures that therefore turn out to be dependent. The graphical model is given by the left part of Figure 3.7: a common, latent random measure P_0 specifies the law of P_i, $i = 1, 2, 3$, that are associated to three distinct groups. When the random measures are discrete, as in the case of the hierarchical Dirichlet process (Teh et al., 2006) the induced clustering implies the presence of ties both within and across groups, leading to a nice and interpretable borrowing of information. This is particularly interesting for instance in topic modelling, where each document is described with a multinomial kernel and each parameter corresponds a topic (i.e. a distribution over all the possible words). Therefore, P_0 becomes a pool of common topics that are shared, with different relevance, by distinct documents. In this context the hierarchical Dirichlet process is the nonparametric extension of the well-known Latent Dirichlet Allocation (Blei et al., 2003), that describes the documents as a mixture of latent topics. Even if it is endowed with great analytical tractability, the Dirichlet process has well–known limitations, both in the exchangeable and non-exchangeable case: consequently, Camerlenghi et al. (2019b) provided a general theory

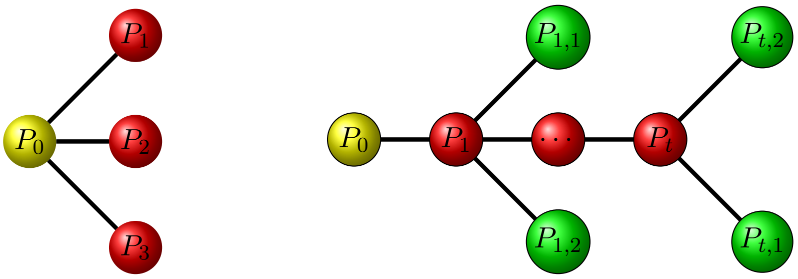


Figure 3.7: Graphical models of a hierarchical structure (left) and a tree structure (right)

for hierarchies of random measures, that allows the usage of various extensions of the Dirichlet process.
Notice that, looking again at Figure 3.7, a hierarchical model can be seen as a very special tree, where P_0 is the root and the observations are collected at the leaves. However, this structure is sometimes too simple to describe all the relevant features of the data. Still considering the topic modelling framework, we may be interested in a corpus of documents that grows in time (e.g. papers submitted each year to a specific conference): it is reasonable to believe that the documents, and the latent topics, yield a temporal dependence that could be exploited. Indeed, a graphical model as in the right part of Figure 3.7 would be more accurate: P_0 plays again the role of the common pool of topics, while the red nodes correspond to the distribution associated to the years of interest, linked to the random measures describing the single documents, given by the green nodes. Motivated by those applications, many models describing similar structures have been introduced, both parametric (Blei and Lafferty, 2006) and nonparametric (Caron et al., 2007,?; Teh, 2006; Wang et al., 2017). Similar probabilistic structures, but in different contexts, have been proposed in Gnedin and Iksanov (2020); Nieto-Barajas (2021). Such proposals focus mostly on temporal dependence and it is often not clear how to incorporate additional information: for instance, in the example above, we may want to distinguish papers belonging to different scientific fields. Moreover, documents may be seen as an ordered collection of chapters and sections, which the usual hierarchical structure is not able to capture. The usefulness of incorporating this order will be shown in Section 3.3.8.

In this work we propose a methodology to construct a generic tree of random probability measures, chosen to describe the underlying features of the dataset. In particular, to each node is associated a random measure endowed with the law of a Pitman-Yor process (Pitman and Yor, 1997; Pitman, 2006), and the edges are given by a hierarchical structure. The construction allows to collect observations at any node (not necessarily the leaves) and to handle properly missing data at every position of the tree. Moreover, thanks to the nice analytical properties of Pitman-Yor process, and its characterization through σ-stable processes, we are able to explicitly assess the impact of the geometry of the tree on the clustering properties of the model. Indeed, considering again the right part of Figure 3.7, it is reasonable that some topics at time 2 actually come from time 1 and this should be reflected on the dependence between P_1 and P_2. We show that our construction implies this behaviour and that such dependence can be suitably tuned using appropriate hyperparameters. Furthermore, the predictive distribution can be derived and allows to perform posterior inference. To summarize, this paper has three goals: (i) provide a general framework to encode various types of dependence through trees of random probability measures; (ii) give explicit expressions for prior quantities of interest (e.g. correlation across groups) and the predictive distribution; (iii) allow to collect data at different (possibly internal) nodes and handle missing data without additional complications.

Throughout the paper we will focus on topic modelling applications, mainly to show the implications of our proposal. However, the construction is completely general and the tree structure is appropriate beyond a corpus of documents. For instance in An et al. (2008) a similar structure, based on kernel stick breaking priors, is used for image analysis. Moreover, similar models can be applied to microbiome data: each document corresponds to a biological sample and each distinct term to a bacterial species. See Sankaran and Holmes (2019) for a review. In this context, tree structures arise naturally to describe compositional data (see Wang et al. (2021) for an example using Pólya trees).

Trees of Pitman-Yor processes have already appeared in language models called sequence memoizers (Wood and Teh, 2009; Teh, 2006). In this case the observations typically take values in a finite space (e.g. words in a dictionary), so that the base measure at the root of the tree is atomic. See also Johnson et al. (2006); Wood et al. (2011) for more details. Our treatment is different in the sense that the sampling space is completely general: moreover, the base measure at the root is diffuse, which is crucial to derive the clustering properties and the predictive distribution (see Theorem 15).

3.3.2 Pitman-Yor process

As discussed extensively, in this document we focus on discrete nonparameteric priors: indeed, considering structures as in Figure 3.7, discreteness allows to make ties both within and across groups, leading to a natural way of borrowing information. Therefore, we assume the realizations of the prior law Q to be almost surely discrete, i.e. $P \overset{\mathrm{d}}{=} \sum_{j \geq 1} W_j \delta_{X_j}$, where $\{W_j\}_j$ are random probability weights and X_j independent random atoms sampled from a suitable probability measure Q_0 on the sampling space $\mathbb{X}$. A popular choice for the distribution of the weights is given by

$$W_i = V_i \prod_{j=1}^{i-1} (1 - V_j), \quad V_i \sim \mathrm{Beta}(1 - \sigma, \theta + i\sigma), \tag{3.7}$$

with $\sigma \in [0, 1)$ and $\theta > 0$. This representation, often called *stick-breaking* construction, leads to the definition of the Pitman-Yor (PY) process (Pitman and Yor, 1997; Pitman, 2006). Notice that the choice $\sigma = 0$ corresponds to the well-known Dirichlet process (Ferguson, 1973).

Defining the process through the weights, as in (3.7), though often useful from a computational perspective, makes a theoretical analysis quite challenging. However, the PY process can be also defined through the predictive distribution, which reads

$$X_{n+1} \mid X_{1:n} \sim \frac{\theta + \sigma K_n}{\theta + n} Q_0 + \frac{1}{\theta + n} \sum_{i=1}^{K_n} (n_j - \sigma) \delta_{X_j^*}, \tag{3.8}$$

where K_n is the number of distinct values $\left(X_1^*, \dots, X_{K_n}^*\right)$ in the sample $X_{1:n} = (X_1, \dots, X_n)$, with multiplicities $(n_1, \dots, n_{K_n})$. Thus, the $(n+1)$-th observation can be either completely new from the baseline Q_0 either copies one of the already observed datapoints: it is then clear that a sample from model for exchangeable data with Q being the law of a PY process exhibits ties with positive probability. A special role is played by the parameter σ, which reinforces the probability of observing new values according to the number of distinct species. The predictive distribution (3.8) is an example of the general class of Gibbs-type priors, which stand out for analytical tractability: see De Blasi et al. (2015) for a recent review. A natural way to characterize such priors is through the Exchangeable Partition Probability Function (EPPF), which describes the

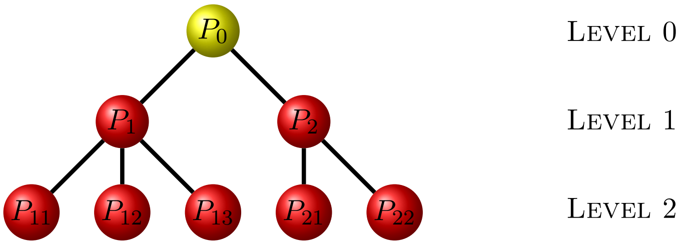


Figure 3.8: Graphical model of a generic tree structure.

induced law on the partitions of n elements (i.e. on the clustering): this representation will be thoroughly discussed in Section 3.3.5).

Finally, it is possible to represent a Pitman-Yor process through Completely Random Measures (CRMs). The latter are random objects on the space of discrete finite measures, which are endowed with nice analytical properties. Many Bayesian nonparametric priors can be described through CRMs (see e.g. Regazzini et al. (2003); James et al. (2006, 2009)), allowing to explicitly derive many posterior quantities of interest. In particular, the law of a PY process is given by a suitable change of measure of the normalized σ-stable process (Pitman and Yor, 1997). See Section 3.2.3 above or Section A3 in the Appendix for more details. This is the characterization we will use for the proofs of all the results in this document.

3.3.3 Building a tree

The graphical model of a tree is illustrated in Figure (3.8). The tree is divided in subsequent levels, where level 0 always corresponds to the root $\mathbf{p}_0$. All the other nodes are identified by a vector of integers, say $\mathbf{p}$, whose position is specified by its values and the level by its length, denoted $|\mathbf{p}|$. For instance in figure 3.8, $P_{2,1}$ is the random measure associated to node $\mathbf{p} = (2, 1)$ at level $|\mathbf{p}| = 2$.

In a tree structure, nodes are connected by edges, that define a children-parent relationship among the nodes. We denote with $(\mathbf{p}, i)$ the i-th child of $\mathbf{p}$ (counting from left to right) and with $\underline{\mathbf{p}}$ the father of $\mathbf{p}$, that is the vector of length $(|\mathbf{p}| - 1)$ derived from $\mathbf{p}$ by truncating the last component. Finally, considering $\mathbf{p} \in \mathcal{T}$, we denote with $\mathcal{C}(\mathbf{p}) \subset \mathcal{T}$ the set of children of $\mathbf{p}$. In order to define a proper tree, each node different from the root must have a parent, so that we can say that

$$\mathcal{T} \subset \bigcup_{k=1}^{\infty} \mathbb{N}_+^k \quad \text{is a tree if and only if} \quad \begin{cases} 0 \in \mathcal{T} \\ \forall \mathbf{p} \in \mathcal{T}, \text{ with } |\mathbf{p}| \geq 2, \exists\, \mathbf{q} \in \mathcal{T} \text{ s.t. } \mathbf{q} = \underline{\mathbf{p}} \end{cases} \tag{3.9}$$

In other words, a tree must contain the root and each other node must have a single edge connecting it to the lower level. Allowing to observe data at each node, the model reads

$$X_{\mathbf{p},i} \mid P_{\mathbf{p}} \overset{\text{i.i.d.}}{\sim} P_{\mathbf{p}}, \quad \{P_{\mathbf{p}}\,;\, \mathbf{p} \in \mathcal{T}\} \sim Q,$$

where $X_{\mathbf{p},i}$ denotes the i-th observation collected at node $\mathbf{p}$.

In order to define the law Q, we need to specify the distribution of the nodes and how the edges affect the dependence among them. In particular, we define the child-parent relation as

$$P_{\mathbf{p}} \mid P_{\underline{\mathbf{p}}} \overset{\text{i.i.d.}}{\sim} \text{PY}(\sigma_{\mathbf{p}}, \theta_{\mathbf{p}}, P_{\underline{\mathbf{p}}}),$$

using the notation introduced in the previous Section. In other words, children are conditionally independent given the father node, that plays the role of the baseline distribution, i.e. it provides the pool of available atoms. Endowing the root with the law of a PY with a diffuse baseline distribution Q_0, we can write the model in a recursive fashion as

$$X_{\mathbf{p},i} \mid P_{\mathbf{p}} \overset{\text{i.i.d.}}{\sim} P_{\mathbf{p}}, \quad P_{\mathbf{p}} \mid P_{\underline{\mathbf{p}}} \overset{\text{i.i.d.}}{\sim} \mathrm{PY}(\sigma_{\mathbf{p}}, \theta_{\mathbf{p}}, P_{\underline{\mathbf{p}}}), \quad P_0 \sim \mathrm{PY}(\sigma_0, \theta_0, Q_0). \tag{3.10}$$

Therefore, each edge corresponds to the creation of a new hierarchy. This makes the depencence between arbitrary nodes $\mathbf{p}$ and $\mathbf{q}$ far from trivial; indeed the strength of their relation will depend on the path of hierarchies that leads from the root to $\mathbf{p}$ and $\mathbf{q}$. This is the matter of next Section.

3.3.4 Prior properties and dependence between the nodes

Considering two nodes $\mathbf{p}$ and $\mathbf{q}$, we will call $\mathbf{m} \in \mathcal{T}$ the Most Recent Common Ancestor (MRCA) of $\mathbf{p}$ and $\mathbf{q}$, that is the the node on the lowest level that is a relative (e.g. connected through a series of edges) of both $\mathbf{p}$ and $\mathbf{q}$. More formally, if $\mathcal{P}(\mathbf{p}) \subset \mathcal{T}$ is the path of connected nodes from the root to $\mathbf{p}$, the MRCA is the element with the highest length belonging to the set $\mathcal{P}(\mathbf{p}) \cap \mathcal{P}(\mathbf{q})$.

For future reference, from now on we will denote

$$\gamma_{\mathbf{p}} = \frac{1 - \sigma_{\mathbf{p}}}{\theta_{\mathbf{p}} + 1}. \tag{3.11}$$

As shown in the next two propositions, the set $\{\gamma_{\mathbf{p}}\}_{\mathbf{p} \in \mathcal{T}}$ plays a crucial role in determining the dependence structure induced by model (3.10). We start from the relation between random measures located at arbitrary positions of the tree.

Proposition 19. *Let $\mathcal{T}$ be a tree with model* (3.10)*. Let $\boldsymbol{p}, \boldsymbol{q} \in \mathcal{T}$ be such that $\boldsymbol{m} \in \mathcal{T}$ is their MRCA. Then, for every set A we have*

$$\mathbb{E}[P_{\boldsymbol{p}}(A)] = Q_0(A), \quad \mathit{Corr}\left(P_{\boldsymbol{p}}(A), P_{\boldsymbol{q}}(A)\right) = \frac{1 - \prod_{\boldsymbol{l} \in \mathcal{P}(\boldsymbol{m})}(1 - \gamma_{\boldsymbol{l}})}{\sqrt{1 - \prod_{\boldsymbol{l} \in \mathcal{P}(\boldsymbol{p})}(1 - \gamma_{\boldsymbol{l}})}\sqrt{1 - \prod_{\boldsymbol{l} \in \mathcal{P}(\boldsymbol{q})}(1 - \gamma_{\boldsymbol{l}})}}.$$

If moreover $\gamma_{\boldsymbol{l}} = \gamma$, for every $\boldsymbol{l} \in \mathcal{T}$, we have

$$\mathit{Corr}\left(P_{\boldsymbol{p}}(A), P_{\boldsymbol{q}}(A)\right) = \frac{1 - (1-\gamma)^{|\boldsymbol{m}|+1}}{\sqrt{1 - (1-\gamma)^{|\boldsymbol{p}|+1}}\sqrt{1 - (1-\gamma)^{|\boldsymbol{q}|+1}}}.$$

Proposition 19 shows that the tree is centered around the baseline distribution of the root Q_0, in the sense that $Q_0(A)$ is the average of each node, for every A. Moreover, the correlation is always positive and independent of the specific set considered: this is reminiscent of most of the priors for partially exchangeable models (e.g. Camerlenghi et al. (2019b)). It is easy to see that the correlation is an increasing function of $\mathcal{P}(\mathbf{m})$, in the sense that the longer the path the stronger the dependence. The intuition is that a long path from the root to the common ancestor leads to a large number of nodes (i.e. information) shared between $\mathbf{p}$ and $\mathbf{q}$: in the context of topic modelling it implies a larger number of topics shared by documents $\mathbf{p}$ and $\mathbf{q}$, as expected. In this sense, our proposal induces the relationships discussed in the Introduction: as we go along the tree, nodes become more and more correlated.

Alternatively, it is possible to measure dependence at the level of the observations, as the next Propositions highlights.

Proposition 20. *Let $\mathcal{T}$ be a tree with model* (3.10)*. Let $\boldsymbol{p}, \boldsymbol{q} \in \mathcal{T}$ be such that $\boldsymbol{m} \in \mathcal{T}$ is their MRCA. Then for every i and j we have*

$$\mathbb{P}(X_{\boldsymbol{p},i} \in A) = Q_0(A), \quad Corr\left(X_{\boldsymbol{p},i}, X_{\boldsymbol{q},j}\right) = \mathbb{P}\left(X_{\boldsymbol{p},i} = X_{\boldsymbol{q},j}\right) = 1 - \prod_{\boldsymbol{l} \in \mathcal{P}(\boldsymbol{m})} (1 - \gamma_{\boldsymbol{l}}).$$

If moreover $\gamma_{\boldsymbol{l}} = \gamma$, for every $\boldsymbol{l} \in \mathcal{T}$, we obtain

$$Corr\left(X_{\boldsymbol{p},i}, X_{\boldsymbol{q},j}\right) = \mathbb{P}\left(X_{\boldsymbol{p},i} = X_{\boldsymbol{q},j}\right) = 1 - (1-\gamma)^{|\boldsymbol{m}|+1}.$$

The intuition is analogous to Proposition 19: dependence becomes stronger along the tree. Interestingly, the correlation between observations at different nodes depends only on the path to the MRCA: indeed the longer $\mathcal{P}(\mathbf{m})$ the higher the probability of a tie. Using the tools of Proposition 20 we can derive the joint distribution of a pair $(X_{\mathbf{p},i}, X_{\mathbf{q},j})$.

Corollary 7. *Let $\mathcal{T}$ be a tree with model* (3.10)*. Let $\boldsymbol{p}, \boldsymbol{q} \in \mathcal{T}$ be such that $\boldsymbol{m} \in \mathcal{T}$ is their MRCA. Then for every indexes i and j and sets A and B we have*

$$\mathbb{P}\left(X_{\boldsymbol{p},i} \in A, X_{\boldsymbol{q},j} \in B\right) = \left[1 - \prod_{\boldsymbol{l} \in \mathcal{P}(\boldsymbol{m})} (1 - \gamma_{\boldsymbol{l}})\right] Q_0(A \cap B) + \left[\prod_{\boldsymbol{l} \in \mathcal{P}(\boldsymbol{m})} (1 - \gamma_{\boldsymbol{l}})\right] Q_0(A) Q_0(B).$$

The sampling mechanism yields a nice interpretation: with probability
$1 - \prod_{\mathbf{l} \in \mathcal{P}(\mathbf{m})}(1 - \gamma_{\mathbf{l}})$, that intuitively quantifies the informations shared between $\mathbf{p}$ and $\mathbf{q}$, the observations are sampled together, otherwise they are collected independently. The next example shows how the above formulas simplify in the particular case of the Dirichlet process, i.e. when $\sigma_{\mathbf{p}} = 0$.

Example 8. *Consider model* (3.10) *with $\sigma_{\boldsymbol{p}} = 0$ for every $\boldsymbol{p} \in \mathcal{T}$. If $\boldsymbol{m}$ is the MRCA of nodes $\boldsymbol{p}$ and $\boldsymbol{q}$, it holds*

$$\gamma_{\boldsymbol{p}} = \frac{1}{1 + \theta_{\boldsymbol{p}}} \quad \textit{and} \quad Corr\left(X_{\boldsymbol{p},i}, X_{\boldsymbol{q},j}\right) = 1 - \prod_{\boldsymbol{l} \in \mathcal{P}(\boldsymbol{m})} \frac{\theta_{\boldsymbol{l}}}{1 + \theta_{\boldsymbol{l}}}.$$

Moreover, if $\theta_{\boldsymbol{p}} = \theta$ for every $\boldsymbol{p} \in \mathcal{T}$ we get

$$Corr\left(X_{\boldsymbol{p},i}, X_{\boldsymbol{q},j}\right) = 1 - \left(\frac{\theta}{1+\theta}\right)^{|\boldsymbol{m}|+1}, \quad Corr\left(P_{\boldsymbol{p}}(A), P_{\boldsymbol{q}}(A)\right) = \frac{1 - \left(\frac{\theta}{1+\theta}\right)^{|\boldsymbol{m}|+1}}{\sqrt{1 - \left(\frac{\theta}{1+\theta}\right)^{|\boldsymbol{p}|+1}} \sqrt{1 - \left(\frac{\theta}{1+\theta}\right)^{|\boldsymbol{q}|+1}}}.$$

If $\theta \to \infty$ the correlations vanish, since the law of the random measures $P_{\boldsymbol{p}}$ degenerate on the deterministic distribution Q_0; the opposite happens if $\theta \to 0$, leading to the maximal correlation.

Therefore, thanks to the nice analytical tractability of the PY process, the prior dependence can be suitably tuned by the researcher through simple formulas depending on $\{((\sigma_{\mathbf{p}}, \theta_{\mathbf{p}})\,;\, \mathbf{p} \in \mathcal{T}\}$. Moreover, using similar tools, it is possible to derive the full predictive structure of model (3.10), and therefore an algorithm for posterior sampling, as will be illustrated in the next Section.

3.3.5 Random partitions and the predictive distribution

A common way to explain the clustering induced by the exchangeable model consists in using a Chinese Reasturant metaphor: the clusters are thought as customers sitting at the same table, where the unique value associated to the cluster is the common dish served. With a hierarchical structure (see Camerlenghi et al. (2019b)), the metaphorical restaurant becomes a franchise: considering again the left part of Figure 3.7, the red nodes correspond to three distinct restaurants, whose customers are again subdivided in tables. However, now the dishes come from the same menu, that is given by the yellow node; since also P_0 is almost surely discrete, different tables may have the same dish. This metaphor is not only amusing, but is also useful to make explicit the dependence within and across nodes (i.e. restaurants): indeed, if the clustering at the restaurant level (i.e. red nodes) is available, sampling a new observation requires only to decide whether to open a new table (whose dish will be sampled from P_0). This is the key idea underlying the algorithms for posterior sampling with hierarchical structures (see Teh et al. (2006); Camerlenghi et al. (2019b)).

Luckily, it is possible to extend the culinary metaphor to model (3.10) associated to a tree $\mathcal{T}$. Indeed each node $\mathbf{p}$ corresponds to a restaurant, whose customers are subdivided in $l_{\mathbf{p}\bullet}$ tables (the notation will be clear in the following). The dishes associated to the tables come from the parent node $\underline{\mathbf{p}}$, that is itself a restaurant, so that they are clustered again. Proceeding recursively, the actual dishes come from the root P_0, that plays the role of the common menu available to all the restaurants. Notice that at each node $\mathbf{p}$ there are the proper customers (i.e. observations collected at $\mathbf{p}$) and dishes coming from the children nodes.

More formally, let $\mathcal{T}$ be a tree with d levels with n_j, $j = 1, \dots, d$, observations collected at each level. Denote with $\mathcal{L}_j \subset \mathcal{T}$ the set of nodes in level j. Then the metaphor becomes as follows

- at level d, the n_d customers are divided in $l_{d\bullet} = \sum_{\mathbf{p}\in\mathcal{L}_d} l_{\mathbf{p}\bullet} \leq n_d$ tables;
- at level $d-1$, the $n_{d-1} + l_{d\bullet}$ customers are divided in $l_{d-1\bullet} = \sum_{\mathbf{p}\in\mathcal{L}_{d-1}} l_{\mathbf{p}\bullet}$ tables;

$$\vdots$$

- at level 1, the $n_1 + l_{2\bullet}$ customers are divided in $l_{1\bullet} = \sum_{\mathbf{p}\in\mathcal{L}_1} l_{\mathbf{p}\bullet}$ tables;
- at level 0, the $l_{1\bullet}$ customers are divided in k tables, whose dishes are sampled from Q_0. Since Q_0 is diffuse, the dishes are almost surely different.

The dishes at level j become new customers at level $j-1$ and therefore observations are clustered in coarser partitions, as we go from the leaves to the root. The latter is the common restaurant that specifies the dishes available to all the customers, regardless of the position of the restaurant. Notice that this latent clustering is not observed in a sample from model (3.10). Indeed we only observe the k distinct dishes and which dish is associated to a customer: however, two customers may share the same dish without seating at the same table, as we discussed. It turns out that knowing the division in tables at each node greatly simplifies the computation, as will be shown in Theorem 15 below.

In order to formalize this, we need to evaluate the partial Exchangeable Partition Probability

Function (pEPPF) associated to model (3.10), defined as

$$\Pi_k^{(n)}\left(\mathbf{n_p}\,;\,\mathbf{p}\in\mathcal{T}\right)=\mathbb{E}\int_{\mathbb{X}_*^k}\prod_{j=1}^{k}\prod_{\mathbf{p}\in\mathcal{T}}P_{\mathbf{p}}^{n_{\mathbf{p},j}}(\mathrm{d}x_j),\tag{3.12}$$

where $\mathbf{n_p}=(n_{\mathbf{p},1},\ldots,n_{\mathbf{p},k})$ is a vector of positive integers such that $n_{\mathbf{p}}=\sum_{j=1}^{k}n_{\mathbf{p},j}$ and $n=\sum_{i=1}^{d}n_i$, with $n_i=\sum_{\mathbf{p}\in\mathcal{L}_i}n_{\mathbf{p}}$ number of observations per level. Moreover, $\mathbb{X}_*^k$ is the subset of $\mathbb{X}_*^k$ given by vectors with all distinct entries. Indeed, $\Pi_k^{(n)}\left(\mathbf{n_p}\,;\,\mathbf{p}\in\mathcal{T}\right)$ is the probability of observing exactly the partition $\{\mathbf{n_p}\,;\,\mathbf{p}\in\mathcal{T}\}$ when sampling $n_{\mathbf{p}}$ observations at node $\mathbf{p}$. In this context k is the number of distinct values in the overall sample and $\mathbf{n_p}$ the vector of associated multiplicities observed at node $\mathbf{p}$.

The pEPPF can be composed starting from the random partitions induced by an exchangeable sequence, described by the Exchangeable Partition Probability Function (EPPF). In particular, for any $p\in\{1,\ldots,n\}$ and any vector of positive integers $(r_1,\ldots,r_p)$ such that $\sum_{j=1}^{p}r_j=n$, we set

$$\Phi_{\mathbf{p},p}^{(n)}(r_1,\ldots,r_p)=\frac{\prod_{i=1}^{p-1}(\theta+i\sigma)}{(\theta_{\mathbf{p}}+1)_{n-1}}\prod_{i=1}^{p}(1-\sigma_{\mathbf{p}})_{r_i-1}\tag{3.13}$$

This is the EPPF induced by a Pitman-Yor process with parameters $(\sigma_{\mathbf{p}},\theta_{\mathbf{p}},Q_0)$, with Q_0 diffuse, see De Blasi et al. (2015); Pitman (2006). We are now able to evaluate the pEPPF (3.12).

Theorem 15. *Let $\mathcal{T}$ be a tree with d levels. Suppose the sequences $\{(X_{\boldsymbol{p},j})_{j\geq 1}\,:\,\boldsymbol{p}\in\mathcal{T}\}$ are partially exchangeable according to model* (3.10)*. Then*

$$\Pi_k^{(n)}\left(\boldsymbol{n_p}\,;\,\boldsymbol{p}\in\mathcal{T}\right)=\sum_{\boldsymbol{l}}\sum_{\boldsymbol{q}}\frac{1}{\boldsymbol{l}!}\binom{\boldsymbol{n}}{\boldsymbol{q}}\Phi_{0,k}^{(l_{1\bullet})}(l_{1,1},\ldots,l_{1,k})\prod_{i=1}^{d}\prod_{\boldsymbol{p}\in\mathcal{L}_i}\Phi_{\boldsymbol{p},l_{\boldsymbol{p}\bullet}}^{(n_{\boldsymbol{p}}+l_{\boldsymbol{p}+1\bullet})}(\boldsymbol{q}_{\boldsymbol{p},1},\ldots,\boldsymbol{q}_{\boldsymbol{p},k}),$$

with

$$\frac{1}{\boldsymbol{l}!}\binom{\boldsymbol{n}}{\boldsymbol{q}}=\prod_{i=1}^{d}\prod_{\boldsymbol{p}\in L_i}\prod_{j=1}^{k}\frac{1}{l_{\boldsymbol{p},j}!}\binom{n_{\boldsymbol{p},j}+l_{\boldsymbol{p}+1,j}}{q_{\boldsymbol{p},j,1},\ldots,q_{\boldsymbol{p},j,l_{\boldsymbol{p},j}}}$$

and where

1. $\sum_{\boldsymbol{l}}=\sum_{i=1}^{d}\sum_{\boldsymbol{p}\in\mathcal{L}_i}\sum_{\boldsymbol{l_p}}$, *where* $\sum_{\boldsymbol{l_p}}=\sum_{l_{\boldsymbol{p},1}=1}^{n_{\boldsymbol{p},1}+l_{\boldsymbol{p}+1,1}}\cdots\sum_{l_{\boldsymbol{p},k}=1}^{n_{\boldsymbol{p},k}+l_{\boldsymbol{p}+1,k}}$ *and* $l_{\boldsymbol{p}+1,j}=\sum_{\boldsymbol{g}\in\mathcal{C}_{\boldsymbol{p}}}l_{\boldsymbol{g},j}$, *with* $l_{\boldsymbol{p},j}\in\{1,\ldots,n_{\boldsymbol{p},j}+l_{\boldsymbol{p}+1,j}\}$;
2. $\sum_{\boldsymbol{q}}=\sum_{i=1}^{d}\sum_{\boldsymbol{p}\in\mathcal{L}_i}\sum_{\boldsymbol{q_p}}$, *where* $\boldsymbol{q_p}=(\boldsymbol{q}_{\boldsymbol{p},1},\ldots,\boldsymbol{q}_{\boldsymbol{p},k})$ *and* $\boldsymbol{q}_{\boldsymbol{p},j}$ *is a vector of positive integers such that* $\sum_{t=1}^{l_{p,j}}q_{\boldsymbol{p},j,t}=n_{\boldsymbol{p},j}+l_{\boldsymbol{p}+1,j}$;
3. $l_{\boldsymbol{p}\bullet}=\sum_{j=1}^{k}l_{\boldsymbol{p},j}$ *and* $\Phi_{r,\boldsymbol{p}}^{(n)}(\cdot)$ *is as in* (3.13).

Within the culinary metaphor, $\mathbf{l}_{\mathbf{p},j}$ is the number of tables in node $\mathbf{p}$ eating dish j, so that $l_{\mathbf{p}\bullet}=\sum_{j=1}^{k}l_{\mathbf{p},j}$ is the total number of tables at node $\mathbf{p}$, whose dishes are given by $\mathbf{q_p}$. In particular, $q_{\mathbf{p},j,t}$ is the number of customers in restaurant $\mathbf{p}$ eating dish j at table t, with $t=1,\ldots,l_{\mathbf{p},j}$. Notice that, given $\mathbf{q}$ and $\mathbf{l}$, the pEPPF reduces to the product

$$\Phi_{0,k}^{(l_{1\bullet})}(l_{1,1},\ldots,l_{1,k})\prod_{i=1}^{d}\prod_{\mathbf{p}\in\mathcal{L}_i}\Phi_{\mathbf{p},l_{\mathbf{p}\bullet}}^{(n_{\mathbf{p}}+l_{\mathbf{p}+1\bullet})}(\mathbf{q}_{\mathbf{p},1},\ldots,\mathbf{q}_{\mathbf{p},k}),\tag{3.14}$$

which displays the random partitions associated to each node of the tree. The product form above makes the predictive distribution very explicit: for instance the probability to a sample a completely new value at node $\mathbf{p}$, conditional on $\mathbf{q}$ and $\mathbf{l}$, becomes

$$\prod_{\mathbf{r}\in\mathcal{P}(\mathbf{p})} \frac{\Phi^{(n_{\mathbf{r}}+l_{\mathbf{r}+1\bullet}+1)}_{\mathbf{r},l_{\mathbf{r}\bullet}+1}(\mathbf{q}_{\mathbf{r},1},\dots,\mathbf{q}_{\mathbf{r},k},1)}{\Phi^{(n_{\mathbf{r}}+l_{\mathbf{r}+1\bullet})}_{\mathbf{r},l_{\mathbf{r}\bullet}}(\mathbf{q}_{\mathbf{r},1},\dots,\mathbf{q}_{\mathbf{r},k})} = \prod_{\mathbf{r}\in\mathcal{P}(\mathbf{p})} \frac{\theta_{\mathbf{r}}+\sigma_{\mathbf{r}} l_{\mathbf{r}\bullet}}{\theta_{\mathbf{r}}+n_{\mathbf{r}}+l_{\mathbf{r}+1\bullet}},$$

that is exactly the probability of creating a new table in each node (i.e. restaurant) in the path from $\mathbf{p}$ to the root.

Then it becomes natural to include the sampling of $\mathbf{l}$ and $\mathbf{q}$ in the algorithm for posterior inference. Therefore, we introduce a set of latent variables $\mathbf{T} = \{\mathbf{T}_{\mathbf{p}}\,;\,\mathbf{p}\in\mathcal{T}\}$, that describes the clustering structure. In terms of the Chinese restaurant metaphor, $T_{\mathbf{p},j}$ is the label of the table of customer (observation) j at restaurant (node) $\mathbf{p}$. Consequently, a sample $\mathcal{X} = \{(X_{\mathbf{p},i})_{i=1}^{n_{\mathbf{p}}}\,;\,\mathbf{p}\in\mathcal{T}\}$, with unqiue values $\{X_1^*,\dots,X_k^*\}$ and multiplicities $\mathbf{n}_{\mathbf{p}}$, will be endowed with latent variables of the form

$$\mathbf{T}_{\mathbf{p}} = \left(T_{\mathbf{p},1},\dots,T_{\mathbf{p},n_{\mathbf{p}}+l_{\mathbf{p}+1\bullet}}\right),$$

for every $\mathbf{p}\in\mathcal{T}$. In particular $T_{\mathbf{p},i}$ is the label associated with observation $X_{\mathbf{p},i}$, with $i = 1,\dots,n_{\mathbf{p}}$, while $T_{\mathbf{p},i}$ with $i > n_{\mathbf{p}}+1$ refers to one of the tables of $\mathbf{p}$'s children, whose common value we denote again with $X_{\mathbf{p},i}$. Moreover we denote with $T^*_{\mathbf{p},r}$ the r-th unique label in $\mathbf{T}_{\mathbf{p}}$, i.e. one of the tables in node $\mathbf{p}$, associated to the value (dish) $X_{T^*_{\mathbf{p},r}}$. Thus, $\mathbf{l}$ and $\mathbf{q}$ can be recovered from $\mathbf{T}$ by

$$\mathbf{l}_{\mathbf{p},j} = \left\{\text{number of unique values } T^*_{\mathbf{p},r} \text{ such that } X_{T^*_{\mathbf{p},r}} = j\right\}$$

and

$$q_{\mathbf{p},j,T^*_{\mathbf{p},r}} = \begin{cases} \text{number of labels } T_{\mathbf{p},i} \text{ such that } T_{\mathbf{p},i} = T^*_{\mathbf{p},r} & \text{if } X_{T^*_{\mathbf{p},r}} = j \\ 0 \quad \text{if } X_{T^*_{\mathbf{p},r}} \neq j \end{cases}$$

Therefore, it is possible to devise a Gibbs sampler on the augmented space $\{(\mathcal{X},\mathbf{T})\,;\,\mathbf{p}\in\mathcal{T}\}$ to perform posterior inference: more explicit details are given in the next Section. Moreover, the pEPPF is not only useful to unveil the clustering structure and devise suitable algorithms for posterior inference, but it also allows to derive interesting properties of model (3.10). In Section 3.3.7 we use it to derive the asymptotic behaviour of the number of clusters.

3.3.6 Posterior sampling

Assume to collect a set of observations $\mathcal{X} = \{(X_{\mathbf{p},i})_{i=1}^{n_{\mathbf{p}}}\,;\,\mathbf{p}\in\mathcal{T}\}$ from model (3.10). In order to perform statistical inference it is necessary to evaluate the distribution of $\{P_{\mathbf{p}}\,;\,\mathbf{p}\in\mathcal{T}\}$ conditional on the sample $\mathcal{X}$. However the latter is not available in closed form, due to the complexity arising by the hierarchical strcuture: thus we need to provide approximations through MCMC methods. The algorithms presented in this Section can be easily extended to mixture models.

A first approach is called *conditional* and consists in simulating trajectories of $P_{\mathbf{p}}$ by its posterior distribution. The latter is a difficult task, since $P_{\mathbf{p}}$ is infinite-dimensional: clever algorithms (Walker, 2007; Papaspiliopoulos and Roberts, 2008) for exact sampling have been proposed for the case of mixture models, but it is not immediate to extend them to the tree structure. For example, the retrospective sampler introduced in Papaspiliopoulos and Roberts (2008) requires a Metropolis-Hastings step for the allocation variables, whose generalization to the hierarchical structure seems challenging. An alternative would be to approximate $P_{\mathbf{p}}$ by truncating the series representation given by (3.7). This leads to an efficient sampler in the exchangeable case (see e.g. Ishwaran and James (2001)), but in our setting this would require a truncation at each node: thus, the propagation of the associated error (and the consequent choice of the trucncation threshold) becomes a significantly harder issue.

A second approach, termed *marginal*, is given by integrating out the random probability measures $P_{\mathbf{p}}$ and sample directly new observations at each node. See Escobar and West (1995) for marginal algorithms in the exchangeable case. In our setting, for example, the distribution of m new observations $X_{\mathbf{p},n_{\mathbf{p}}+1}, \dots, X_{\mathbf{p},n_{\mathbf{p}}+m}$ is given by

$$\mathbb{P}\left(\cap_{i=1}^{m} \left\{X_{\mathbf{p},n_{\mathbf{p}}+i} \in A_i\right\}\right) = \int \prod_{i=1}^{m} P_{\mathbf{p}}(A_i) Q(P_{\mathbf{p}} \mid \mathcal{X}), \tag{3.15}$$

where $Q(\cdot \mid \mathcal{X})$ is the posterior distribution of $\{P_{\mathbf{p}}\,;\, \mathbf{p} \in \mathcal{T}\}$ induced by model (3.10).

Direct evaluation and exact sampling of (3.15) are infeasible, but the availability of the pEPPF in Theorem 15 allows to explicitly derive the full conditionals of the set $\{(\mathcal{X}, \mathbf{T})\,;\, \mathbf{p} \in \mathcal{T}\}$, using the notation of the previous Section. Therefore it is possible to devise a Gibbs sampler on the augmented space to sample new observations at different nodes. Assume for simplicity that we want to sample a single new observation at node $\mathbf{q}$, that may be placed everywhere on the tree: the algorithm can be straightforwardly extended to multiple new observations at different nodes. In this case, the algorithm requires to sample node specific labels $T_{\mathbf{q},i}$, associated to some observation X_i, and the new pair $\left(T_{\mathbf{p},\mathbf{n_p}+1}, X_{\mathbf{p},n_{\mathbf{p}}+1}\right)$. As regards the former, the conditional distribution is given by

$$\begin{aligned} \mathbb{P}\left(T_{\mathbf{q},i} = \text{ new} \mid \mathcal{X}, \mathbf{T}^{-i}\right) &\propto \frac{\theta_{\mathbf{q}} + \sigma_{\mathbf{q}} l_{\mathbf{q}\bullet}^{-i}}{\theta_{\mathbf{q}} + n_{\mathbf{q}} + l_{\mathbf{q}+1\bullet} - 1}, \\ \mathbb{P}\left(T_{\mathbf{q},i} = T^*_{\mathbf{q},r} \mid \mathcal{X}, \mathbf{T}^{-i}\right) &\propto \frac{q^{-i}_{\mathbf{q},X_i,T^*_{\mathbf{q},r}} - \sigma_{\mathbf{q}}}{\theta_{\mathbf{q}} + n_{\mathbf{q}} + l_{\mathbf{q}+1\bullet} - 1}, \end{aligned} \tag{3.16}$$

where $q_{\mathbf{q},t}$ is the number of customers at table t in node $\mathbf{q}$, while the superscript $(-i)$ refers to quantities computed after the removal of the label $T_{\mathbf{q},i}$. Notice that if the sampling of $T_{\mathbf{q},i}$ results in a new label, i.e. creating a new table, this implies that $l_{\mathbf{q}\bullet}$ is increased by one and a new label at the parent node $\underline{\mathbf{q}}$ must be sampled by its conditional distribution, as in (3.16) with $\underline{\mathbf{q}}$ in place of $\mathbf{q}$. This procedure must be performed recursively along $\mathcal{P}(\mathbf{q})$ until a label copies one of the dishes at the same level (second line in (3.16) or the the root is reached. Therefore, sampling a label in $\mathbf{q}$ may lead to creating new labels in the path from the root to $\mathbf{q}$, reflecting the hierarchical structure of model (3.10). As regards sampling the pair $\left(T_{\mathbf{p},\mathbf{n_p}+1}, X_{\mathbf{p},n_{\mathbf{p}}+1}\right)$, similarly to (3.16) the label $T_{\mathbf{p},\mathbf{n_p}+1}$ yields a conditional distribution

$$\begin{aligned} \mathbb{P}\left(T_{\mathbf{p},\mathbf{n_p}+1} = \text{ new} \mid \mathcal{X}, \mathbf{T}\right) &\propto \frac{\theta_{\mathbf{p}} + \sigma_{\mathbf{p}} l_{\mathbf{p}\bullet}}{\theta_{\mathbf{p}} + n_{\mathbf{p}} + l_{\mathbf{p}+1\bullet}}, \\ \mathbb{P}\left(T_{\mathbf{p},\mathbf{n_p}+1} = T^*_{\mathbf{p},r} \mid \mathcal{X}, \mathbf{T}\right) &\propto \frac{q_{\mathbf{p},X_{T^*_{\mathbf{p},r}},t} - \sigma_{\mathbf{p}}}{\theta_{\mathbf{p}} + n_{\mathbf{p}} + l_{\mathbf{p}+1\bullet}}, \end{aligned} \tag{3.17}$$

while for the value it holds

$$X_{\mathbf{p},n_{\mathbf{p}}+1} \mid T_{\mathbf{p},\mathbf{n_p}+1}, \mathcal{X}, \mathbf{T} \sim \begin{cases} \delta_{X_{\mathbf{q},T^*_{\mathbf{q},r}}} & \text{if } T_{\mathbf{p},\mathbf{n_p}+1} \text{ copied } T^*_{\mathbf{q},r}, \text{ with } \mathbf{q} \in \mathcal{P}(\mathbf{p}) \\ Q_0 \quad \text{otherwise} \end{cases} \tag{3.18}$$

Therefore the new value either copies one of the already existing observations either is sampled by the base measure of the root, becoming a completely new value. The algorithm then reads:

1. Initialize $\mathbf{T}$.
2. For every $t = 1, \dots, T$:
   - For every $\mathbf{q} \in \mathcal{T}$, sample $T_{\mathbf{q},i}$ for every $i = 1, \dots, n_{\mathbf{q}} + l_{\mathbf{q}+1\bullet}$ as in (3.16). Sample the possibly new labels in the previous levels.
   - Sample $T_{\mathbf{p},\mathbf{n_p}+1}$ as in (3.17). Sample the possibly new labels in the previous levels.
   - Sample $X_{\mathbf{p},n_{\mathbf{p}}+1}$ according to (3.18).

A nice byproduct of the above algorithm is that it provides the full clustering structure at each iteration. By Theorem 15 and especially (3.14), conditional on $\mathbf{T}$ the dependence among the random measure greatly simplifies: in particular $P_{\mathbf{p}}$ will depend only on $P_{\mathbf{q}}$, with $\mathbf{q} \in \mathcal{P}(\mathbf{p}) \backslash \{\mathbf{p}\}$. The latter fact makes posterior sampling of $P_{\mathbf{p}}$ much simpler, by sampling sequentially at each level starting from the root, so that a marginal algorithm can be also used if direct sampling of the random measures is required.

3.3.7 Distribution of the number of clusters

When studying discrete priors, the number of clusters, i.e. the number of distinct values in a sample of n observations, is often a crucial object. This happens either in species sampling problems, when the data are given by frequencies, or in mixture models, when the clustering structure is latent. In topic modelling it represents the number of topics used to describe all the documents in the corpus.

Given a tree $\mathcal{T}$ with d levels and denoting with K_n the number of clusters, with $n = \sum_{i=1}^{d} n_i$, we say that K_n behaves asymptotically as a deterministic sequence $\lambda(n)$ if

$$\lim_{n \to \infty} \frac{K_n}{\lambda(n)} = M$$

almost surely, where M is a positive and finite random variable. For simplicity we use the notation $K_n \approx \lambda(n)$. Let

$$\lambda_\sigma(n) = \begin{cases} n^\sigma & \text{if } \sigma > 0 \\ \log(n) & \text{if } \sigma = 0 \end{cases} \tag{3.19}$$

Notice that $\lambda_\sigma(n)$ describes the asymptotic behaviour of K_n arising from a single PY process with parameter σ, see De Blasi et al. (2015) for more details. We consider the regime in which $n = n_{\mathbf{p}}$, with $\mathbf{p} \in \mathcal{T}$, so that all the observations are collected at a single node.

Theorem 16. *Let $\mathcal{T}$ be a tree and $\boldsymbol{p} \in \mathcal{T}$. If $n = n_{\boldsymbol{p}}$ we have*

$$K_n \approx \left(\prod_{\boldsymbol{q} \in \mathcal{P}(\boldsymbol{p})} \lambda_{\sigma_{\boldsymbol{q}}} \right) (n)$$

almost surely as $n \to \infty$*, where* $\lambda_{\sigma_1}\lambda_{\sigma_2}(n) = \lambda_{\sigma_1}\left(\lambda_{\sigma_2}(n)\right)$.

The asymptotic behaviour of K_n is given by combining the ones of the nodes forming a path from the root to $\mathbf{p}$; the asymptotic rate becomes lower as we go along the tree, as expected. If the $\sigma_{\mathbf{p}}$'s are all strictly positive or all equal to 0 (i.e. the nodes have Dirichlet process law) the behaviour becomes particularly simple, as shown in the next example.

Example 9. *Assume* $\sigma_{\boldsymbol{q}} = 0$ *for every* $\boldsymbol{q} \in \mathcal{P}(\boldsymbol{p})$*. Thus, under the setting of Theorem 16 we have*

$$K_n \approx \underbrace{\log \ldots \log}_{|\boldsymbol{p}|+1 \ times} n,$$

almost surely as $n \to \infty$*. Assume instead* $\sigma_{\boldsymbol{q}} > 0$ *for every* $\boldsymbol{q} \in \mathcal{P}(\boldsymbol{p})$*, then*

$$K_n \approx n^{\prod_{q \in \mathcal{P}(p)} \sigma_q},$$

almost surely as $n \to \infty$*.*

This behaviour is reminiscent to the exchangeable case: K_n diverges almost surely, with a rate that depends on the position in the tree. We consider now an alternative regime, where $m \geq 1$ observations are collected at each level $i = 1, \ldots, d$, and the number of levels diverge. This is somewhat complementary to the first case, in the sense that the sample is spread over the whole tree.

Theorem 17. *Let* $\{\mathcal{T}_d\}_d$ *be a sequence of trees such that* $\mathcal{T}_d$ *has d levels and the restriction of* $\mathcal{T}_{d+1}$ *to the first d levels is equal to* $\mathcal{T}_d$*, for every d. Moreover, assume there exists* $\bar{\theta}$ *and* $\bar{\sigma} < 1$ *such that* $\theta_{\boldsymbol{p}} \leq \bar{\theta}$ *and* $\sigma_{\boldsymbol{p}} \leq \bar{\sigma}$*, for every* $\boldsymbol{p} \in \mathcal{T}$*. Then, if* $m \geq 1$ *observations are collected at each level different from* 0*, we have*

$$\limsup_{d \to \infty} K_n < \infty$$

almost surely, where $n = n(d) = md$*.*

The intuition is that the observations become more and more correlated (see Proposition 20), so that in the limit with infinite-levels the probability of a completely new value (i.e. dish) becomes negligible.

Theorems 16 and 17 show that the geometry of the tree is crucial for the clustering properties of the model and should be chosen wisely. Indeed, a change in the level for a specific node $\mathbf{p}$ leads to a substantial change in its rate of divergence; moreover, each node separately would yield an infinite number of clusters, so it is really the tree structure that leads to a finite amount in the second regime. Notice that the techniques of Theorems 16 and 17 could be applied to different regimes than the two considered, without additional difficulties.

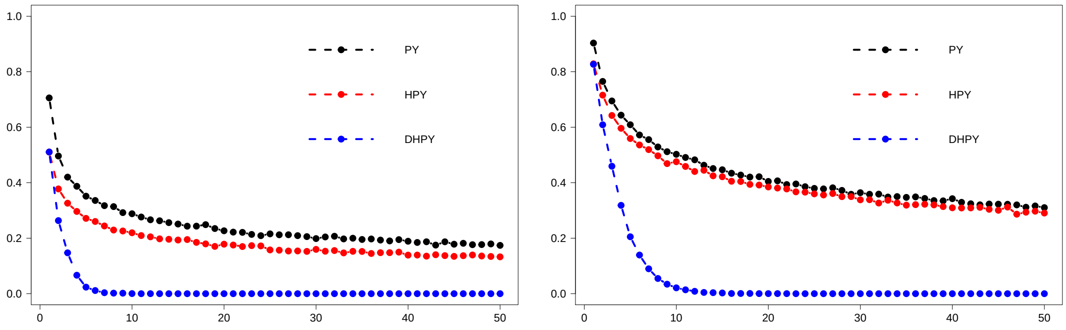
In order to appreciate this variety, we consider as an example three different structures:

(i) a single-node tree (i.e. exchangeable case), distributed as a Pitman Yor process (PY) with parameters σ and θ;

(ii) a hierarchical process with a single group, in which the two nodes are distributed as PY with parameters σ and θ. All the observations are collected at the leaf and the model is denoted with HPY;

P_0
P_0
P_1
P_0
P_1
P_2
$\cdots$

Figure 3.9: Three different structures: single-node tree (left), hierarchical process (center) and sequence of nodes (right).



(a) $\theta = 1, \sigma = 0.7$

(b) $\theta = 10, \sigma = 0.7$

Figure 3.10: Proportion of new values for each subset of 10 elements out of 500, with $m = 10$ and averaging over 1000 samples, for PY, HPY and DHPY.

(iii) a sequence of nodes distributed as a Normalized Stable Process (NSP) with parameters σ and θ. It is a special tree with a single branch. At each node $m = 10$ observations are collected and the model is denoted with DHPY.

The three specifications are shown in Figure 3.9. According to Theorems 16 and 17 the number of clusters in the first two settings should diverge with rates n^σ and n^{σ^2} respectively, with $\sigma \in (0, 1)$, while in the third scenario a finite amount of clusters should be observed. Figure 3.10 shows the average proportion of new values for each batch of 10 observations: as expected, PY and HPY have a similar decay, but with different rate (single-node tree has the highest number of clusters), while for DHPY the proportion of new values drops close to zero after few batches. Therefore, our proposal is able to encompass a large variety of prior clustering behaviours and the structure of the tree should be chosen with care, that should be tuned according to the problem at hand, as showed in the next Section.

3.3.8 Application

We consider a topic modelling application, in which model (3.10) is convolved with a kernel. Consider a vocabulary of V words and let $\mathbb{X}$ to be the space of probability distributions over $\{1, \ldots, V\}$: therefore, each value (i.e. topic) sampled at node $\mathbf{p}$, say $X_\mathbf{p}$, is a vector of V elements, denoted with $X_\mathbf{p}(w)$, with $w = 1, \ldots, V$. Each node $\mathbf{p}$ corresponds to a document, whose words $\{Y_{\mathbf{p},i}\}_{i=1}^{n_\mathbf{p}}$ are assumed to be exchangeable with multinomial kernel. In formula it reads

$$\mathbb{P}\left(Y_{\mathbf{p},i} = w \mid X_\mathbf{p}\right) = X_\mathbf{p}(w), \tag{3.20}$$

with $w = 1, \ldots, V$. The baseline distribution of the root, denoted with Q, is a Dirichlet distribution with V elements and common parameter α: it means that a priori there is no preference among the words.

Specification (3.20) implies that each document is a mixture of topics, that are shared within the whole corpus with different relevance among the documents. See Blei et al. (2003); Teh et al. (2006) for more details.

The goal of the analysis may be either explorative, that is studying which topics arise in the corpus and how they are distributed, or predictive, to sample new words associated to the corpus. The work Blei and Lafferty (2006) is probably the most connected to our proposal in the context of topic modelling, since it is a dynamic extension of the well known latent Dirichlet allocation (Blei et al., 2003). However, being a parametric model, it requires to specify the number of topics, that in our nonparametric framework is automatically chosen through the data. Instead Caron et al. (2007, 2017) propose a time-varying model based on Pólya urns, whose invariant distribution is given by the Dirichlet or the Pitman-Yor process, with an elegant formulation. A weakness of the above constructions, shared with Wang et al. (2017), is that they are defined to accomodate temporal dependence and it is not easy to introduce additional structures, e.g. the field of the document. Within our framework, instead, the tree can be suitably defined to describe appropriately any structure of the corpus. In the following we show how encoding information about the corpus architecture may make inference and prediction more robust, in particular with a high percentage of missing data.

Alice in Wonderland

A book can be considered as a sequence of chapters, that have a precise order. Moreover, it is reasonable to assume that later chapters regard mostly topics from the previous part of the books.

Considering the first three chapters of Alice in Wonderland (by Lewis Carroll) we show how it is possible to model them with our proposal. As shown in the left part of Figure 3.11, their relationship can be described through a very specific tree, with a single branch: notice that observations (i.e. words) are collected at all nodes apart from the root. For comparison we consider also a hierarchical process, depicted in the left part of Figure 3.11: the three chapters are still dependent, through the common root P_0, but the sequentiality of the chapters is not included.

For both models each node is endowed with a Pitman-Yor distribution with node-specific random parameters θ and σ, with Gamma and uniform priors respectively. Standard stop words (e.g. conjuctions) have been eliminated and only the roots of the words are taken into consideration, through a so-called stemming procedure (in particular the Porter algorithm, see e.g. Jivani et al. (2011)). This leaves around 5000 distinct words, after eliminating the ones appearing less than 4 times. The parameter α of the baseline distribution is set equal to $50/V$, to avoid a negligible prior variance.

In order to compare the performances, we hold out an increasing portion of words in chapter 2 and measure how well the two models replace the missing data. Figure 3.12 depicts the perplexity, that measures the discrepancy between the held-out words and the predictive distributions (see Teh et al. (2006)), for the two models: it is clear that the tree structure behaves better and has a good reconstruction even with a high proportion of missing words. This shows that incorporating the structures of the data in the model architecture may better predictive performances.

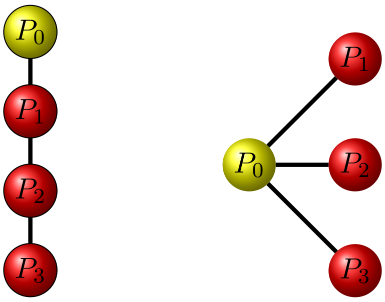


Figure 3.11: Two structures to model the first three chapters of Alice in Wonderland: tree (left) and hierarchical process (right).

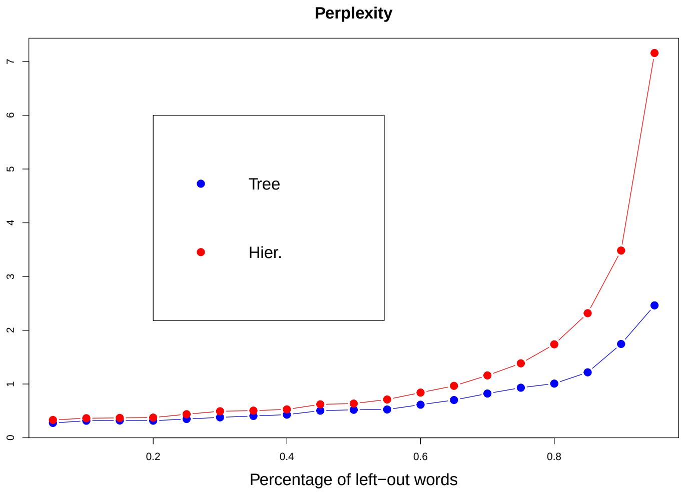


Figure 3.12: Perplexity associated to hierarchical and tree structures with an increasing proportion of missing words from the second chapter of Alice in Wonderland. Results are averaged over 20 runs.

A1 Proofs of Section 3.2

Proof of Proposition 11

Proof. Consider two partially exchangeable sequences X and Y whose elements take value in $\mathbb{R}$. By de Finetti's representation theorem, there exist two random probability measures P_1 and P_2 such that

$$(X_i, Y_j) \mid P_1, P_2 \overset{\text{i.i.d.}}{\sim} P_1 \times P_2.$$

Note that $\mathrm{Cov}(X_i, Y_j) = \mathbb{E}\left[\mathrm{Cov}(X_i, Y_j \mid P_1, P_2)\right] + \mathrm{Cov}\left(\mathbb{E}[X_i \mid P_1], \mathbb{E}[Y_j \mid P_2]\right)$, where the first term equals 0, so that

$$\mathrm{Cov}(X_i, Y_j) = \mathrm{Cov}\left(\int x\, P_1(dx), \int x\, P_2(dx)\right),$$

and analogously

$$\mathrm{Cov}(X_i, X_{i'}) = \mathrm{Cov}\left(\int x\, P_1(dx), \int x\, P_1(dx)\right) = \mathrm{var}\left(\int x\, P_1(dx)\right).$$

Lastly assume that $P_1 \overset{d}{=} P_2$, where $\overset{d}{=}$ indicates equality in distribution. By the Cauchy-Schwartz inequality

$$-\mathrm{Var}\left(\int x\, P_1(dx)\right) \leq \mathrm{Cov}\left(\int x\, P_1(dx), \int x\, P_2(dx)\right) \leq \mathrm{Var}\left(\int x\, P_1(dx)\right),$$

which, in terms of the observables, can be equivalently rewritten as

$$-\mathrm{Cov}(X_i, X_{i'}) \leq \mathrm{Cov}(X_i, Y_j) \leq \mathrm{Cov}(X_i, X_{i'}).$$

□

Proof of Proposition 12

Proof. By definition of covariance we have

$$\mathrm{Cov}(X_i, Y_j) = \mathrm{Cov}\left(\sum_{j\geq 1} J_j\theta_j, \sum_{k\geq 1} W_k\phi_k\right) = \sum_{j\geq 1}\sum_{k\geq 1} \mathrm{Cov}\left(J_j\theta_j, W_k\phi_k\right).$$

For arbitrary j and k we have

$$\mathbb{E}\left[J_jW_k\theta_j\phi_k\right[= \mathbb{E}[J_jW_k]\mathbb{E}[\theta_j\phi_k] \geq \mathbb{E}[J_jW_k]\mathbb{E}[\theta_j]\mathbb{E}[\phi_k],$$

since $\mathrm{Cov}(\theta_j, \phi_k) \geq 0$. Denoting $c = \mathbb{E}[\theta_j] = \mathbb{E}[\phi_k]$, we get

$$\mathrm{Cov}\left(J_j\theta_j, W_k\phi_k\right) \geq c^2\mathrm{Cov}(J_j, W_k).$$

Finally, since P_1 and P_2 are random probability measures it holds

$$\mathrm{Cov}(X_i, Y_j) \geq c^2\mathrm{Cov}\left(\sum_{j\geq 1} J_j, \sum_{k\geq 1} W_k\right) = 0,$$

which completes the proof. □

Proof of Proposition 13

Proof. Recall that

$$\beta := \sum_{k\geq 1}\mathbb{E}(\bar{J}_k^2) = \sum_{k\geq 1}\mathbb{E}(\bar{W}_k^2) \qquad \gamma := \sum_{k\geq 1}\mathbb{E}(\bar{J}_k\bar{W}_k).$$

Since

$$\mathbb{E}(\bar{J}_k\bar{W}_k) \leq \sqrt{\mathbb{E}(\bar{J}_k^2)\mathbb{E}(\bar{W}_k^2)} = \mathbb{E}(\bar{J}_k^2)$$

it follows that $\gamma \leq \beta$. Moreover, the equality holds if and only if $\bar{J}_k \overset{a.s}{=} a_k + \bar{W}_k$, for any k, with $a_k \in \mathbb{R}$. However the equality of marginal distributions implies $a_k = 0$. □

Proof of Proposition 14

Proof. Recall that

$$\mathrm{Cov}(X_i, Y_j) = \mathrm{Cov}\left(\sum_{k\geq 1} \bar{J}_k\theta_k, \sum_{h\geq 1} \bar{W}_h\phi_h\right) = \sum_{k\geq 1}\sum_{h\geq 1} \mathrm{Cov}\left(\bar{J}_k\theta_k, \bar{W}_h\phi_h\right).$$

and for arbitrary k and h, we have

$$\begin{aligned}\mathbb{E}[\bar{J}_k\bar{W}_h\theta_k\phi_h] =& \mathbb{E}[\bar{J}_k\bar{W}_h]\mathbb{E}[\theta_k\phi_h]\\ =& \mathbb{E}[\bar{J}_k\bar{W}_h]\left\{\mathbb{E}[\theta_k\phi_k]\mathbb{1}_{\{k=h\}} + \mathbb{E}[\theta_k]\mathbb{E}[\phi_h]\mathbb{1}_{\{k\neq h\}}\right\},\end{aligned}$$

while

$$\mathbb{E}[\bar{J}_k\theta_k] = \mathbb{E}[\bar{J}_k]\mathbb{E}[\theta_k].$$

Thus, setting $c = E[\theta_k] = E[\phi_h]$, we have

$$\mathrm{Cov}(X_i, Y_j) = \sum_{k\geq 1} \mathbb{E}[\bar{J}_k\bar{W}_h]\mathbb{E}[\theta_k\phi_k] - c^2\sum_{k\geq 1}\mathbb{E}[\bar{J}_k]\mathbb{E}[\bar{W}_k] + +c^2\sum_{k\geq 1}\sum_{h\neq k}\mathrm{Cov}\left(\bar{J}_k, \bar{W}_h\right)$$

where

$$\begin{aligned}\sum_{k\geq 1}\sum_{h\neq k}\mathrm{Cov}\left(\bar{J}_k, \bar{W}_h\right) =& \mathrm{Cov}\left(\sum_{k\geq 1}\bar{J}_k, \sum_{h\geq 1}\bar{W}_h\right) - \sum_{k\geq 1}\mathrm{Cov}\left(\bar{J}_k\bar{W}_k\right)\\ =& -\sum_{k\geq 1}\mathbb{E}[\bar{J}_k\bar{W}_h] + \sum_{k\geq 1}\mathbb{E}[\bar{J}_k]\mathbb{E}[\bar{W}_k]\end{aligned}$$

Putting everything together we obtain

$$\mathrm{Cov}(X_i, Y_j) = \sum_{k\geq 1}\mathbb{E}[\bar{J}_k\bar{W}_k]\mathrm{Cov}(\theta_k, \phi_k).$$

Moreover

$$\mathrm{Var}(X_i) = \mathrm{Var}(Y_j) = \int\int x\, G_0(\mathrm{d}x, \mathrm{d}y) = \mathrm{var}(\theta_k)$$

Thus, $\mathrm{Corr}(X_i, Y_j) = \gamma\,\rho_0$ proving the second statement in Proposition 4. Finally, applying the same procedure marginally, we get

$$\mathrm{Cov}(X_i, X_i') = \sum_{k\geq 1}\mathbb{E}[\bar{J}_k^2]\,\mathrm{Var}(\theta_k)$$

which proves the first statement in Proposition 14. □

Proof of Corollary 5

Proof. The result immediately follows from Propositions 13 and 14. □

Proof of Proposition 15

Proof. Let β be the probability of a tie. By definitionwe get

$$\begin{aligned}\mathbb{P}\left(X_1 \in A, X_2 \in B\right) =& \mathbb{P}(X_1 \in A, X_2 \in B \mid X_1 = X_2)\beta + \\ &+ \mathbb{P}(X_1 \in A, X_2 \in B \mid X_1 \neq X_2)(1-\beta),\end{aligned}$$

which, by independence of the atoms, equals

$$\begin{aligned}\mathbb{P}\left(X_1 \in A, X_2 \in B\right) =& \mathbb{P}(X_1 \in A \in B)\beta + \\ &+ \mathbb{P}(X_1 \in A)\mathbb{P}(X_2 \in B)(1-\beta).\end{aligned}$$

Analogously, we have

$$\begin{aligned}\mathbb{P}\left(X_1 \in A, Y_1 \in B\right) =& \mathbb{P}(X_1 \in A, Y_1 \in B \mid X_1 \text{ and } Y_1 \text{ form an hyper-tie })\gamma + \\ &+ \mathbb{P}(X_1 \in A, Y_1 \in B \mid X_1 \text{ and } Y_1 \text{ do not form an hyper-tie })(1-\gamma),\end{aligned}$$

where γ is the probability of a hyper-tie, which equals

$$\begin{aligned}\mathbb{P}\left(X_1 \in A, Y_1 \in B\right) =& \mathbb{P}((X_1, Y_1) \in A \times B \mid X_1 \text{ and } Y_1 \text{ form an hyper-tie })\gamma + \\ &+ \mathbb{P}(X_1 \in A)\mathbb{P}(Y_1 \in B)(1-\gamma).\end{aligned}$$

□

Proof of Proposition 16

Proof. The first point follows from the Lévy-Khintchine representation of the Laplace functional of a CRV. As for (ii), one has

$$\begin{aligned}\mathbb{E}\left(\exp\{-\lambda_1\mu_1(A) - \lambda_2\mu_2(B)\}\right) =& \mathbb{E}\left(\exp\{-\lambda_1\tilde{\mu}_1(A \times \mathbb{X}) - \lambda_2\tilde{\mu}_2(\mathbb{X} \times B)\}\right) \\ =& \mathbb{E}\Big(\exp\{-\lambda_1\tilde{\mu}_1(A \times B^c) - \lambda_1\tilde{\mu}_1(A \times B) + \\ & - \lambda_2\tilde{\mu}_2(A^c \times B) - \lambda_2\tilde{\mu}_2(A \times B)\}\Big).\end{aligned}$$

By independence of evaluations on disjoint sets, $\tilde{\mu}_1(C)$ and $\tilde{\mu}_2(D)$ are independent if $C \cap D = \emptyset$, so that the right hand side reads

$$\begin{aligned}\mathbb{E}\Big(\exp\{-\lambda_1\mu_1(A) - \lambda_2\mu_2(B)\}\Big) =& \mathbb{E}\left(\exp\{-\lambda_1\tilde{\mu}_1(A \times B^c)\}\right)\mathbb{E}\left(\exp\{-\lambda_2\tilde{\mu}_2(A^c \times B)\}\right) \times \\ & \times \mathbb{E}\left(\exp\{-\lambda_1\tilde{\mu}_1(A \times B) - \lambda_2\tilde{\mu}_2(A \times B)\}\right).\end{aligned}$$

The result follows upon upon using the expressions of the marginal and joint Laplace exponents of $\tilde{\mu}_1$ and $\tilde{\mu}_2$. Since from the joint Lévy intensity it is possible to recover the joint Laplace exponent, (iii) is also proved. □

Proof of Proposition 17

We want to show that

$$\mathbb{P}\left(X \in A, Y \in B\right) = Q_0(A)Q_0(B)\left(1-\delta\right) + G_0(A \times B)\delta,$$

where

$$\delta := -\int_{\mathbb{R}_+^2} \left\{ \frac{\partial^2}{\partial u_1 \partial u_2} \psi_b(u_1, u_2) \right\} e^{-\psi_b(u_1,u_2)} \, \mathrm{d}u_1 \mathrm{d}u_2.$$

is the probability of a pseudo-tie. We start with three Lemmas.

Lemma 27. *If ψ_b is the joint Laplace exponent of a CRV, then*

$$\int_{\mathbb{R}_+^2} \left\{ \frac{\partial}{\partial u_1} \psi_b(u_1, u_2) \right\} \left\{ \frac{\partial}{\partial u_2} \psi_b(u_1, u_2) \right\} e^{-\psi_b(u_1,u_2)} \, \mathrm{d}u_1 \mathrm{d}u_2 = 1 - \delta.$$

Proof. Integrating by parts

$$\begin{aligned}
&\int_0^\infty \left\{ \frac{\partial}{\partial u_1} \psi_b(u_1, u_2) \right\} \left\{ \frac{\partial}{\partial u_2} \psi_b(u_1, u_2) \right\} e^{-\psi_b(u_1,u_2)} \, \mathrm{d}u_1 \\
&= -\int_0^\infty \left\{ \frac{\partial}{\partial u_2} \psi_b(u_1, u_2) \right\} \left\{ \frac{\partial}{\partial u_1} e^{-\psi_b(u_1,u_2)} \right\} \mathrm{d}u_1 \\
&= \left[\left[-\left\{ \frac{\partial}{\partial u_2} \psi_b(u_1, u_2) \right\} e^{-\psi_b(u_1,u_2)} \right]_0^\infty + \int_0^\infty \left\{ \frac{\partial^2}{\partial u_1 \partial u_2} \psi_b(u_1, u_2) \right\} e^{-\psi_b(u_1,u_2)} \, \mathrm{d}u_1 \right] \\
&= \left[\left\{ \frac{\partial}{\partial u_2} \psi_b(0, u_2) \right\} e^{-\psi_b(0,u_2)} + \int_0^\infty \left\{ \frac{\partial^2}{\partial u_1 \partial u_2} \psi_b(u_1, u_2) \right\} e^{-\psi_b(u_1,u_2)} \, \mathrm{d}u_1 \right].
\end{aligned}$$

Note that $\int_0^\infty \left\{ \frac{\mathrm{d}}{\mathrm{d}u_2} \psi_b(0, u_2) \right\} e^{-\psi_b(0,u_2)} \, \mathrm{d}u_2 = 1$, by the fundamental theorem of calculus. Thus the result follows immediately. □

Lemma 28. *We have*

$$\int_{\mathbb{R}_+^2} \mathbb{E} \left(e^{-u_1 \tilde{\mu}_1(\mathbb{X}\times\mathbb{X}) - u_2 \tilde{\mu}_2(\mathbb{X}\times\mathbb{X})} \tilde{\mu}_1(C) \tilde{\mu}_2(C) \right) \mathrm{d}u_1 \mathrm{d}u_2 = G_0(C)^2 (1 - \delta) + G_0(C)\delta.$$

Proof. By independence of evaluations on disjoint sets it follows that

$$\begin{aligned}
&\int_{\mathbb{R}_+^2} \mathbb{E} \left(e^{-u_1 \tilde{\mu}_1(\mathbb{X}\times\mathbb{X}) - u_2 \tilde{\mu}_2(\mathbb{X}\times\mathbb{X})} \tilde{\mu}_1(C) \tilde{\mu}_2(C) \right) \mathrm{d}u_1 \mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \mathbb{E} \left(e^{-u_1 \tilde{\mu}_1(C) - u_2 \tilde{\mu}_2(C) - u_1 \tilde{\mu}_1(C^c) - u_2 \tilde{\mu}_2(C^c)} \tilde{\mu}_1(C) \tilde{\mu}_2(C) \right) \mathrm{d}u_1 \mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \mathbb{E} \left(e^{-u_1 \tilde{\mu}_1(C) - u_2 \tilde{\mu}_2(C)} \tilde{\mu}_1(C) \tilde{\mu}_2(C) \right\} \mathbb{E} \left(e^{-u_1 \tilde{\mu}_1(C^c) - u_2 \tilde{\mu}_2(C^c)} \right) \mathrm{d}u_1 \mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \mathbb{E} \left(\frac{\partial}{\partial u_1} \frac{\partial}{\partial u_2} e^{-u_1 \tilde{\mu}_1(C) - u_2 \tilde{\mu}_2(C)} \right) \mathbb{E} \left(e^{-u_1 \tilde{\mu}_1(C^c) - u_2 \tilde{\mu}_2(C^c)} \right) \mathrm{d}u_1 \mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \frac{\partial}{\partial u_1} \frac{\partial}{\partial u_2} \left[\mathbb{E} \left(e^{-u_1 \tilde{\mu}_1(C) - u_2 \tilde{\mu}_2(C)} \right) \right] \mathbb{E} \left(e^{-u_1 \tilde{\mu}_1(C^c) - u_2 \tilde{\mu}_2(C^c)} \right) \mathrm{d}u_1 \mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \frac{\partial}{\partial u_1} \frac{\partial}{\partial u_2} \left\{ e^{-G_0(C)\psi_b(u_1,u_2)} \right\} e^{-G_0(C^c)\psi_b(u_1,u_2)} \, \mathrm{d}u_1 \mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \frac{\partial}{\partial u_1} \left\{ -G_0(C) \frac{\partial}{\partial u_2} \psi_b(u_1, u_2) e^{-G_0(C)\psi_b(u_1,u_2)} \right\} e^{-G_0(C^c)\psi_b(u_1,u_2)} \, \mathrm{d}u_1 \mathrm{d}u_2.
\end{aligned}$$

Performing the derivative with respect to u_1, the latter expression can be written as follows

$$\begin{aligned}
&= \int_{\mathbb{R}_+^2} \left\{ G_0(C)^2 \frac{\partial}{\partial u_1} \psi_b(u_1,u_2) \frac{\partial}{\partial u_2} \psi_b(u_1,u_2) \right\} e^{-G_0(C)\psi_b(u_1,u_2)} e^{-G_0(C^c)\psi_b(u_1,u_2)} \,\mathrm{d}u_1 \mathrm{d}u_2 + \\
&+ \int_{\mathbb{R}_+^2} \left\{ -G_0(C) \frac{\partial}{\partial u_1 \partial u_2} \psi_b(u_1,u_2) \right\} e^{-G_0(C)\psi_b(u_1,u_2)} e^{-G_0(C^c)\psi_b(u_1,u_2)} \,\mathrm{d}u_1 \mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \left\{ G_0(C)^2 \frac{\partial}{\partial u_1} \psi_b(u_1,u_2) \frac{\partial}{\partial u_2} \psi_b(u_1,u_2) \right\} e^{-\psi_b(u_1,u_2)} \,\mathrm{d}u_1 \mathrm{d}u_2 + \\
&+ \int_{\mathbb{R}_+^2} \left\{ -G_0(C) \frac{\partial}{\partial u_1 \partial u_2} \psi_b(u_1,u_2) \right\} e^{\psi_b(u_1,u_2)} \,\mathrm{d}u_1 \mathrm{d}u_2
\end{aligned}$$

By Lemma 27 we then obtain

$$\int_{\mathbb{R}_+^2} \mathbb{E}\left(e^{-u_1\tilde{\mu}_1(\mathbb{X}\times\mathbb{X}) - u_2\tilde{\mu}_2(\mathbb{X}\times\mathbb{X})} \tilde{\mu}_1(C)\tilde{\mu}_2(C) \right) \mathrm{d}u_1 \mathrm{d}u_2 = G_0(C)^2\,(1-\delta) + G_0(C)\delta,$$

as desired. □

Lemma 29. *Let C, D be such that $C \cap D = \emptyset$. Then*

$$\int_{\mathbb{R}_+^2} \mathbb{E}\left(e^{-u_1\tilde{\mu}_1(\mathbb{X}\times\mathbb{X}) - u_2\tilde{\mu}_2(\mathbb{X}\times\mathbb{X})} \tilde{\mu}_1(C)\tilde{\mu}_2(D) \right) \mathrm{d}u_1 \mathrm{d}u_2 = G_0(C)G_0(D)\,(1-\delta)$$

Proof. Let $Y = (C \cup D)^c$. Since C and D are disjoint, by independence of evaluations on disjoint sets it holds

$$\begin{aligned}
&\int_{\mathbb{R}_+^2} \mathbb{E}\left(e^{-u_1\tilde{\mu}_1(\mathbb{X}\times\mathbb{X}) - u_2\tilde{\mu}_2(\mathbb{X}\times\mathbb{X})} \tilde{\mu}_1(C)\tilde{\mu}_2(D) \right) \mathrm{d}u_1 \mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \mathbb{E}\left(e^{-u_1\tilde{\mu}_1(C\cup D) - u_2\tilde{\mu}_2(C\cup D)} \tilde{\mu}_1(C)\tilde{\mu}_2(D) \right\} \mathbb{E}\left\{ e^{-u_1\tilde{\mu}_1(Y) - u_2\tilde{\mu}_2(Y)} \right) \mathrm{d}u_1 \mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \mathbb{E}\left(e^{-u_1\tilde{\mu}_1(C) - u_2\tilde{\mu}_2(C)} \tilde{\mu}_1(C) \right) \mathbb{E}\left(e^{-u_1\tilde{\mu}_1(D) - u_2\tilde{\mu}_2(D)} \tilde{\mu}_2(D) \right) \times \\
&\qquad \times \mathbb{E}\left(e^{-u_1\tilde{\mu}_1(Y) - u_2\tilde{\mu}_2(Y)} \right) \mathrm{d}u_1 \mathrm{d}u_2 \\
&= \int_{\mathbb{R}_+^2} \frac{\partial}{\partial u_1} \left\{ e^{-G_0(C)\psi_b(u_1,u_2)} \right\} \frac{\partial}{\partial u_2} \left\{ e^{-G_0(D)\psi_b(u_1,u_2)} \right\} e^{-G_0(Y)\psi_b(u_1,u_2)} \,\mathrm{d}u_1 \mathrm{d}u_2 \\
&= G_0(C)G_0(D) \int_{\mathbb{R}_+^2} \left\{ \frac{\partial}{\partial u_1} \psi_b(u_1,u_2) \frac{\partial}{\partial u_2} \psi_b(u_1,u_2) \right\} e^{-\psi_b(u_1,u_2)} \,\mathrm{d}u_1 \mathrm{d}u_2
\end{aligned}$$

The result follows by applying Lemma 27. □

Proof of Proposition 17. We have

$$\mathbb{P}\left(X\in A, Y\in B\right)=\mathbb{E}\left[\frac{\mu_1(A)}{\mu_1(\mathbb{X})}\frac{\mu_2(B)}{\mu_2(\mathbb{X})}\right[=\mathbb{E}\left[\frac{\tilde{\mu}_1(A\times\mathbb{X})}{\tilde{\mu}_1(\mathbb{X}\times\mathbb{X})}\frac{\tilde{\mu}_2(\mathbb{X}\times B)}{\tilde{\mu}_2(\mathbb{X}\times\mathbb{X})}\right]=$$

$$=\int_{\mathbb{R}_+^2}\mathbb{E}\left(e^{-u_1\tilde{\mu}_1(\mathbb{X}\times\mathbb{X})-u_2\tilde{\mu}_2(\mathbb{X}\times\mathbb{X})}\tilde{\mu}_1(A\times\mathbb{X})\tilde{\mu}_2(\mathbb{X}\times B)\right)\,\mathrm{d}u_1\mathrm{d}u_2=$$

$$=\int_{\mathbb{R}_+^2}\mathbb{E}\Bigg(e^{-u_1\tilde{\mu}_1(\mathbb{X}\times\mathbb{X})-u_2\tilde{\mu}_2(\mathbb{X}\times\mathbb{X})}\Big\{\tilde{\mu}_1(A\times B)\tilde{\mu}_2(A\times B)+\tilde{\mu}_1(A\times B)\tilde{\mu}_2(A^c\times B)+$$

$$+\tilde{\mu}_1(A\times B^c)\tilde{\mu}_2(A\times B)+\tilde{\mu}_1(A\times B^c)\tilde{\mu}_2(A^c\times B)\Big\}\Bigg)\,\mathrm{d}u_1\mathrm{d}u_2$$

We compute each integral separately applying Lemmas 28 and 29 and obtain

$$\begin{aligned}\mathbb{P}\left(X\in A, Y\in B\right)&=G_0(A\times\mathbb{X})G_0(\mathbb{X}\times B)\left(1-\delta\right)+G_0(A\times B)\delta\\&=Q_0(A)Q_0(B)\left(1-\delta\right)+G_0(A\times B)\delta,\end{aligned}\tag{21}$$

as desired. Then the probability of a tie in the product space is given exactly by δ, denoted γ in the manuscript. The probability of a tie is given by the particular case $\psi_b(u_1,u_2)=\psi(u_1+u_2)$, since

$$-\int_{\mathbb{R}_+^2}\left\{\frac{\partial^2}{\partial u_1\partial u_2}\psi_b(u_1+u_2)\right\}e^{-\psi_b(u_1+u_2)}\,\mathrm{d}u_1\mathrm{d}u_2=-\int_0^\infty\int_0^u\mathrm{d}v\left\{\frac{\partial^2}{\partial u^2}\psi_b(u)\right\}e^{-\psi_b(u)}\,\mathrm{d}u,$$

with the change of variables $u=u_1+u_2$ and $v=u_1$. □

Proof of Proposition 18

Proof. Since

$$\mathbb{E}\left(P_1(A)P_2(B)\right)=\mathbb{P}\left(X\in A, Y\in B\right),$$

by (21) we have

$$\mathbb{E}\left[P_1(A)P_2(B)\right]=G_0(A\times\mathbb{X})G_0(\mathbb{X}\times B)\left(1-\gamma\right)+G_0(A\times B)\gamma.$$

Finally,

$$\begin{aligned}\mathrm{Cov}\left(P_1(A),P_2(B)\right)&=G_0(A\times\mathbb{X})G_0(\mathbb{X}\times B)\left(1-\gamma\right)+G_0(A\times B)\gamma-G_0(A\times\mathbb{X})G_0(\mathbb{X}\times B)\\&=\gamma\left\{G_0(A\times B)-G_0(A\times\mathbb{X})G_0(\mathbb{X}\times B)\right\}.\end{aligned}$$

From this one also obtains

$$\begin{aligned}\mathrm{Var}\left(P_1(A)\right)=\mathrm{Cov}\left(P_1(A),P_1(A)\right)&=\beta\left\{Q_0(A)-Q_0(A)^2\right\}\\&=\beta Q_0(A)\left\{1-Q_0(A)\right\},\end{aligned}$$

and the desired result follows. □

Proof of Theorem 13

Proof. We need to compute the conditional Laplace functional of $(\tilde{\mu}_1, \tilde{\mu}_2)$, i.e.

$$\mathbb{E}\left(e^{-\int_{\mathbb{X}^2} h_1(x)\,\tilde{\mu}_1(\mathrm{d}x) - \int_{\mathbb{X}^2} h_2(x)\,\tilde{\mu}_2(\mathrm{d}x)} \mid (X_i)_{i=1}^n, (Y_j)_{j=1}^m\right),$$

with $h_i : \mathbb{X}^2 \to \mathbb{R}^+$ measurable functions. Define $A_j = A_{j,\epsilon} = \{x \in \mathbb{X} \mid d(x, X_i^*) < \epsilon\}$ and $B_j = B_{j,\epsilon} = \left\{x \in \mathbb{X} \mid d(x, Y_j^*) < \epsilon\right\}$, with $1 \le i \le k$ and $1 \le j \le c$, such that $A_i \cap A_j = \emptyset$ and $B_i \cap B_j = \emptyset$ for any $i \neq j$. Moreover, denote

$$A_{k+1} = \left(\cup_{i=1}^k A_i\right)^c, \quad B_{c+1} = \left(\cup_{i=1}^c B_i\right)^c.$$

Thus our goal becomes to compute

$$\begin{aligned}
&\mathbb{E}\left[e^{-\int_{\mathbb{X}^2} h_1(x)\,\tilde{\mu}_1(\mathrm{d}x) - \int_{\mathbb{X}^2} h_2(x)\,\tilde{\mu}_2(\mathrm{d}x)} \mid (X_i)_{i=1}^n, (Y_j)_{j=1}^m\right] \\
&= \lim_{\epsilon \to 0} \mathbb{E}\left[e^{-\int_{\mathbb{X}^2} h_1(x)\,\tilde{\mu}_1(\mathrm{d}x) - \int_{\mathbb{X}^2} h_2(x)\,\tilde{\mu}_2(\mathrm{d}x)} \mid \underline{X}_n^* \in \times_{j=1}^k A_j, \underline{Y}_m^* \in \times_{j=1}^c B_j\right] \\
&= \lim_{\epsilon \to 0} \frac{\mathbb{E}\left[e^{-\int_{\mathbb{X}^2} h_1(x)\,\tilde{\mu}_1(\mathrm{d}x) - \int_{\mathbb{X}^2} h_2(x)\,\tilde{\mu}_2(\mathrm{d}x)} \prod_{j=1}^k P_1(A_j)^{n_j} \prod_{j=1}^c P_2(B_j)^{m_j}\right]}{\mathbb{E}\left[\prod_{j=1}^k P_1(A_j)^{n_j} \prod_{j=1}^c P_2(B_j)^{m_j}\right]}.
\end{aligned} \tag{22}$$

We start to evaluate

$$\begin{aligned}
&\mathbb{E}\left[P_1(A_1)^{n_1} \dots P_1(A_k)^{n_k} P_2(B_1)^{m_1} P_2(B_c)^{m_c}\right] = \\
&= \mathbb{E}\left[\frac{\mu_1(A_1)^{n_1} \dots \mu_1(A_k)^{n_k} \mu_2(B_1)^{m_1} \mu_2(B_c)^{m_c}}{\mu_1(\mathbb{X})^n \mu_2(\mathbb{X})^m}\right] \\
&= \mathbb{E}\left[\frac{\tilde{\mu}_1(A_1 \times \mathbb{X})^{n_1} \dots \tilde{\mu}_1(A_k \times \mathbb{X})^{n_k} \tilde{\mu}_2(\mathbb{X} \times B_1)^{m_1} \tilde{\mu}_2(\mathbb{X} \times B_c)^{m_c}}{\tilde{\mu}_1(\mathbb{X} \times \mathbb{X})^n \tilde{\mu}_2(\mathbb{X} \times \mathbb{X})^m}\right) = \mathcal{I}.
\end{aligned}$$

By Netwon's binomial

$$\begin{aligned}
\tilde{\mu}_1(A_h \times \mathbb{X}) &= \sum_{i_1^h + \dots i_{c+1}^h = n_h} \binom{n_h}{i_1^h, \dots, i_{c+1}^h} \prod_{r=1}^{c+1} \tilde{\mu}_1^{i_r^h}(A_h \times B_r), \quad h = 1, \dots, k, \\
\tilde{\mu}_2(\mathbb{X} \times B_r) &= \sum_{j_1^r + \dots j_{k+1}^r = m_r} \binom{m_r}{j_1^r, \dots, j_{k+1}^r} \prod_{h=1}^{k+1} \tilde{\mu}_2^{j_h^r}(A_h \times B_r), \quad r = 1, \dots, c.
\end{aligned}$$

For ease of notation denote

$$\begin{aligned}
\sum_{\boldsymbol{i},\boldsymbol{j}} \binom{\boldsymbol{n}}{\boldsymbol{i}} \binom{\boldsymbol{m}}{\boldsymbol{j}} = &\sum_{i_1^1 + \dots i_{c+1}^1 = n_1} \binom{n_1}{i_1^1, \dots, i_{c+1}^1} \cdots \sum_{i_1^{c+1} + \dots i_{c+1}^{k+1} = n_{k+1}} \binom{n_{k+1}}{i_1^{k+1}, \dots, i_{c+1}^{k+1}} \times \\
&\times \sum_{j_1^1 + \dots j_{k+1}^1 = m_1} \binom{m_1}{j_1^1, \dots, j_{k+1}^1} \cdots \sum_{j_1^{k+1} + \dots j_{k+1}^{k+1} = m_{k+1}} \binom{m_{k+1}}{j_1^{k+1}, \dots, j_{k+1}^{k+1}}.
\end{aligned}$$

Thus

$$\mathcal{I} = \sum_{\boldsymbol{i},\boldsymbol{j}} \binom{\boldsymbol{n}}{\boldsymbol{i}} \binom{\boldsymbol{m}}{\boldsymbol{j}} \mathcal{I}_{\boldsymbol{i},\boldsymbol{j}},$$

with

$$\mathcal{I}_{\boldsymbol{i},\boldsymbol{j}} = \mathbb{E}\Bigg[\frac{\prod_{h=1}^{k} \prod_{r=1}^{c} \tilde{\mu}_1^{i_r^h}(A_h \times B_r) \tilde{\mu}_2^{j_h^r}(A_h \times B_r)}{\tilde{\mu}_1(\mathbb{X} \times \mathbb{X})^n} \times \\ \times \frac{\prod_{h=1}^{k} \tilde{\mu}_1^{i_{c+1}^h}(A_h \times B_{c+1}) \prod_{r=1}^{c} \tilde{\mu}_2^{j_{k+1}^r}(A_{k+1} \times B_r)}{\tilde{\mu}_2(\mathbb{X} \times \mathbb{X})^m} \Bigg]$$

Letting $\tilde{\mu}_1 := \tilde{\mu}_1(\mathbb{X} \times \mathbb{X})$ and $\tilde{\mu}_2 := \tilde{\mu}_2(\mathbb{X} \times \mathbb{X})$, we have

$$\frac{1}{\tilde{\mu}_1(\mathbb{X} \times \mathbb{X})^n \tilde{\mu}_2(\mathbb{X} \times \mathbb{X})^m} = \frac{1}{\Gamma(n)\Gamma(m)} \int_{\mathbb{R}_+^2} u_1^{n-1} u_2^{m-1} e^{-u_1\tilde{\mu}_1 - u_2\tilde{\mu}_2} \, \mathrm{d}\underline{u},$$

with $\underline{u} = (u_1, u_2)$. Thus, by Fubini's Theorem

$$\mathcal{I}_{\mathbf{i},\mathbf{j}} = \int_{\mathbb{R}_+^2} \frac{u_1^{n-1} u_2^{m-1}}{\Gamma(n)\Gamma(m)} \mathbb{E}\Bigg[e^{-u_1\tilde{\mu}_1 - u_2\tilde{\mu}_2} \left\{ \prod_{h=1}^{k} \prod_{r=1}^{c} \tilde{\mu}_1^{i_r^h}(A_h \times B_r) \tilde{\mu}_2^{j_h^r}(A_h \times B_r) \right\} \times \\ \times \prod_{h=1}^{k} \tilde{\mu}_1^{i_{c+1}^h}(A_h \times B_{c+1}) \prod_{r=1}^{c} \tilde{\mu}_2^{j_{k+1}^r}(A_{k+1} \times B_r) \Bigg] \mathrm{d}\underline{u} = \\ = \int_{\mathbb{R}_+^2} \frac{u_1^{n-1} u_2^{m-1}}{\Gamma(n)\Gamma(m)} \rho_{\mathbf{i},\mathbf{j}}(\underline{u}) \, \mathrm{d}\underline{u}.$$

By independence of evaluations on disjoint sets we have

$$\rho_{\mathbf{i},\mathbf{j}}(\underline{u}) = \mathbb{E}\Bigg[\left\{ \prod_{h=1}^{k} \prod_{r=1}^{c} e^{-u_1\tilde{\mu}_1(A_h \times B_r) - u_2\tilde{\mu}_2(A_h \times B_r)} \tilde{\mu}_1^{i_r^h}(A_h \times B_r) \tilde{\mu}_2^{j_h^r}(A_h \times B_r) \right\} \times \\ \times \left\{ \prod_{h=1}^{k} e^{-u_1\tilde{\mu}_1(A_h \times B_{c+1}) - u_2\tilde{\mu}_2(A_h \times B_{c+1})} \tilde{\mu}_1^{i_{c+1}^h}(A_h \times B_{c+1}) \right\} \times \\ \times \left\{ \prod_{r=1}^{c} e^{-u_1\tilde{\mu}_1(A_{k+1} \times B_r) - u_2\tilde{\mu}_2(A_{k+1} \times B_r)} \mu_2^{j_{k+1}^r}(A_{k+1} \times B_r) \right\} \Bigg]$$

This can be equivalently written as

$$\prod_{h=1}^{k} \prod_{r=1}^{c} \mathbb{E}\left[e^{-u_1\tilde{\mu}_1(A_h \times B_r) - u_2\tilde{\mu}_2(A_h \times B_r)} \tilde{\mu}_1^{i_r^h}(A_h \times B_r) \tilde{\mu}_2^{j_h^r}(A_h \times B_r) \right] \times \\ \times \prod_{h=1}^{k} \mathbb{E}\left[e^{-u_1\tilde{\mu}_1(A_h \times B_{c+1}) - u_2\tilde{\mu}_2(A_h \times B_{c+1})} \tilde{\mu}_1^{i_{c+1}^h}(A_h \times B_{c+1}) \right] \times \\ \times \prod_{r=1}^{c} \mathbb{E}\left[e^{-u_1\tilde{\mu}_1(A_{k+1} \times B_r) - u_2\tilde{\mu}_2(A_{k+1} \times B_r)} \tilde{\mu}_2^{j_{k+1}^r}(A_{k+1} \times B_r) \right].$$

Considering each element separately we have

$$\begin{aligned}\mathbb{E}\left[e^{-u_1\tilde{\mu}_1(A_h\times B_r)-u_2\tilde{\mu}_2(A_h\times B_r)}\tilde{\mu}_1^i(A_h\times B_r)\tilde{\mu}_2^j(A_h\times B_r)\right]&\\ &=\mathbb{E}\left[(-1)^{i+j}\frac{\partial^{i+j}}{\partial u_1^i\partial u_2^j}e^{-u_1\tilde{\mu}_1(A_h\times B_r)-u_2\tilde{\mu}_2(A_h\times B_r)}\right]\\ &=(-1)^{i+j}\frac{\partial^{i+j}}{\partial u_1^i\partial u_2^j}\mathbb{E}\left[e^{-u_1\tilde{\mu}_1(A_h\times B_r)-u_2\tilde{\mu}_2(A_h\times B_r)}\right]\\ &=(-1)^{i+j}\frac{\partial^{i+j}}{\partial u_1^i\partial u_2^j}\left\{e^{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(ds)G_0(x)}\right\}.\end{aligned}$$

Recall that we are interested in the limit as $\epsilon\to 0$, so that

$$\begin{aligned}\frac{\partial^{i+j}}{\partial u_1^i\partial u_2^j}\left\{e^{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}\right\}\sim e^{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}\times&\\ \times\frac{\partial^{i+j}}{\partial u_1^i\partial u_2^j}\left\{\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right\},&\end{aligned}\tag{23}$$

where we say $f\sim g$ if $\lim_{\epsilon\to 0}f(x)/g(x)=1$. By simple algebra we get

$$\begin{aligned}\frac{\partial^{i+j}}{\partial u_1^i\partial u_2^j}\left\{e^{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}\right\}=\frac{\partial^{i+j-1}}{\partial u_1^{i-1}\partial u_2^j}\left\{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}e^{-u_1s_1-u_2s_2}\times\right.&\\ \left.\times\,s_1\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)e^{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}\right\}&\\ =\frac{\partial^{i+j-2}}{\partial u_1^{i-2}\partial u_2^j}\left\{\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}e^{-u_1s_1-u_2s_2}s_1^2\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)e^{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}\right.&\\ \left.+\left(\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}e^{-u_1s_1-u_2s_2}s_1\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right)^2e^{-\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}(1-e^{-u_1s_1-u_2s_2})\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}\right\},&\end{aligned}$$

and

$$\lim_{\epsilon\to 0}\frac{\left(\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}e^{-u_1s_1-u_2s_2}s_1\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)\right)^2}{\int_{A_h\times B_r}\int_{\mathbb{R}_+^2}e^{-u_1s_1-u_2s_2}s_1^2\,\rho(\mathrm{d}s)G_0(\mathrm{d}x)}=0.$$

By applying this argument repeatedly we obtain (23). Thus, letting $\rho(\underline{u}) = \sum_{\boldsymbol{i},\boldsymbol{j}} \binom{\boldsymbol{n}}{\boldsymbol{i}}\binom{\boldsymbol{m}}{\boldsymbol{j}} \rho_{\boldsymbol{i},\boldsymbol{j}}(\underline{u})$, by aggregating the terms we have

$$\begin{aligned}
\rho(\underline{u}) \sim & \sum_{\boldsymbol{i},\boldsymbol{j}} \binom{\boldsymbol{n}}{\boldsymbol{i}}\binom{\boldsymbol{m}}{\boldsymbol{j}} (-1)^{n+m} e^{-\psi_b(u)} \times \\
& \times \prod_{h=1}^{k}\prod_{r=1}^{c} \left\{ \frac{\partial^{i_r^h + j_h^r}}{\partial u_1^{i_r^h} \partial u_2^{j_h^r}} \int_{A_h \times B_r} \int_{\mathbb{R}_+^2} (1 - e^{-u_1 s_1 - u_2 s_2})\, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\} \times \\
& \times \prod_{h=1}^{k} \left\{ \frac{\partial^{i_{c+1}^h}}{\partial u_1^{i_{c+1}^h}} \int_{A_h \times B_{c+1}} \int_{\mathbb{R}_+^2} (1 - e^{-u_1 s_1 - u_2 s_2})\, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\} \times \\
& \times \prod_{r=1}^{c} \left\{ \frac{\partial^{j_{k+1}^r}}{\partial u_2^{i_{k+1}^r}} \int_{A_{k+1} \times B_r} \int_{\mathbb{R}_+^2} (1 - e^{-u_1 s_1 - u_2 s_2})\, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\} \\
= & \sum_{\boldsymbol{i},\boldsymbol{j}} \binom{\boldsymbol{n}}{\boldsymbol{i}}\binom{\boldsymbol{m}}{\boldsymbol{j}} (-1)^{n+m} V(\boldsymbol{i},\boldsymbol{j}).
\end{aligned}$$

The following three Lemmas characterize the set of indices $(\boldsymbol{i},\boldsymbol{j})$ that are relevant once the limit is taken.

Lemma 30. *Consider* $(\boldsymbol{i},\boldsymbol{j})$ *such that* $0 < i_r^h, i_l^h < n_h$*, with* $r > l$ *and* $1 \leq h \leq k$*. Then* $\exists(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}})$ *such that* $\lim_{\epsilon\to 0} V(\boldsymbol{i},\boldsymbol{j})/V(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}}) \to 0$.

Proof. For ease of notation set $\boldsymbol{i}^h = (i_1^h, \dots, i_{c+1}^h)$. Then

- If $r = c+1$, set $\tilde{\boldsymbol{i}}^h = (i_1^h, \dots, i_l^h + i_{c+1}^h, \dots, 0)$.
- If $j_h^r = 0$, set $\tilde{i}^h = (i_1^h, \dots, i_l^h + i_r^h, \dots, 0, \dots)$.
- If $j_h^l = 0$, set $\tilde{\boldsymbol{i}}^h = (i_1^h, \dots, 0, \dots, i_r^h + i_l^h, \dots)$.
- If $j_h^l > 0$ and $j_h^r > 0$, set $\tilde{\boldsymbol{j}}^r = (j_1^r, \dots, 0, \dots, j_{k+1}^r + j_h^r)$ and $\tilde{\boldsymbol{i}}^h = (i_1^h, \dots, i_l^h + i_r^h, \dots, 0, \dots)$.

For example in the last case we have

$$\lim_{\epsilon\to 0} \frac{\mathrm{var}(\boldsymbol{i},\boldsymbol{j})}{\mathrm{var}(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}})} = \lim_{\epsilon\to 0} \frac{\int_{A_h \times B_r} \int_{\mathbb{R}_+^2} e^{-u_1 s_1 - u_2 s_2} s_1^{i_r^h} s_2^{j_h^r}\, \rho(\mathrm{d}s) G_0(\mathrm{d}x)}{\int_{A_{c+1} \times B_r} \int_{\mathbb{R}_+^2} e^{-u_1 s_1 - u_2 s_2} s_2^{j_h^r + j_{c+1}^r}\, \rho(\mathrm{d}s) G_0(\mathrm{d}x)} = 0,$$

as desired. □

Thus, Lemma 34 guarantees that $\boldsymbol{i}^h$ has exactly one element different from 0, that is equal to n_h.

Lemma 31. *Consider* $(\boldsymbol{i},\boldsymbol{j})$ *such that* $i_r^h = n_h$ *and* $j_h^r = 0$*. Then there exists* $(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}})$ *such that* $\lim_{\epsilon\to 0} V(\boldsymbol{i},\boldsymbol{j})/V(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}}) \to 0$.

Proof. Set $(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}})$ equal to $(\mathbf{i},\mathbf{j})$, apart from $\tilde{i}_r^h = 0$ and $\tilde{i}_{c+1}^h = n_h$. □

Lemma 32. *Consider* $(\boldsymbol{i},\boldsymbol{j})$ *such that* $i_{c+1}^h = n_h$ *and* $j_h^r > 0$*. Then there exists* $(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}})$ *such that* $\lim_{\epsilon\to 0} V(\boldsymbol{i},\boldsymbol{j})/V(\tilde{\boldsymbol{i}},\tilde{\boldsymbol{j}}) \to 0$.

of Lemma S2.6. Set $(\tilde{\boldsymbol{i}}, \tilde{\boldsymbol{j}})$ equal to $(\boldsymbol{i}, \boldsymbol{j})$, apart from $\tilde{j}^r_h = 0$ and $\tilde{j}^r_{k+1} = m_r$. □

The three lemmas imply that each relevant $(\boldsymbol{i}, \boldsymbol{j})$ corresponds to an admissible latent structure, i.e.

$$\rho(u) \sim \sum_{\mathbf{p}\in\mathcal{P}} (-1)^{n+m} e^{-\psi_b(\underline{u})} \prod_{(i,j)\in\Delta_{\mathbf{p}}} \left\{ \frac{\partial^{n_i+m_j}}{\partial u_1^{n_i} \partial u_2^{m_j}} \int_{A_i\times B_j} \int_{\mathbb{R}^2_+} (1 - e^{-u_1 s_1 - u_2 s_2})\, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\} \times$$
$$\times \prod_{(i,j)\in\Delta^1_{\mathbf{p}}} \left\{ \frac{\partial^{n_i}}{\partial u_1^{n_i}} \int_{A_i\times B_{c+1}} \int_{\mathbb{R}^2_+} (1 - e^{-u_1 s_1 - u_2 s_2})\, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\} \times$$
$$\times \prod_{(i,j)\in\Delta^2_{\mathbf{p}}} \left\{ \frac{\partial^{m_j}}{\partial u_2^{m_j}} \int_{A_{k+1}\times B_j} \int_{\mathbb{R}^2_+} (1 - e^{-u_1 s_1 - u_2 s_2})\, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\}.$$

Evaluating the derivatives we have

$$\rho(\underline{u}) \sim \sum_{\mathbf{p}\in\mathcal{P}} e^{-\psi_b(\underline{u})} \prod_{(i,j)\in\Delta_{\mathbf{p}}} \left\{ \int_{A_i\times B_j} \int_{\mathbb{R}^2_+} e^{-u_1 s_1 - u_2 s_2} s_1^{n_i} s_2^{m_j}\, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\} \times$$
$$\times \prod_{(i,j)\in\Delta^1_{\mathbf{p}}} \left\{ \int_{A_i\times B_{c+1}} \int_{\mathbb{R}^2_+} e^{-u_1 s_1 - u_2 s_2} s_1^{n_i}\, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\} \times$$
$$\times \prod_{(i,j)\in\Delta^2_{\mathbf{p}}} \left\{ \int_{A_{k+1}\times B_j} \int_{\mathbb{R}^2_+} e^{-u_1 s_1 - u_2 s_2} s_2^{m_j}\, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\}.$$

Finally, we get

$$\mathcal{I} \sim \sum_{\mathbf{p}\in\mathcal{P}} \int_{\mathbb{R}^2_+} \frac{u_1^{n-1} u_2^{m-1}}{\Gamma(n)\Gamma(m)} e^{-\psi_b(\underline{u})} \prod_{(i,j)\in\Delta_{\mathbf{p}}} \left\{ \int_{A_i\times B_j} \int_{\mathbb{R}^2_+} e^{-u_1 s_1 - u_2 s_2} s_1^{n_i} s_2^{m_j}\, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\} \times$$
$$\times \prod_{(i,j)\in\Delta^1_{\mathbf{p}}} \left\{ \int_{A_i\times B_{c+1}} \int_{\mathbb{R}^2_+} e^{-u_1 s_1 - u_2 s_2} s_1^{n_i}\, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\} \times$$
$$\times \prod_{(i,j)\in\Delta^2_{\mathbf{p}}} \left\{ \int_{A_{k+1}\times B_j} \int_{\mathbb{R}^2_+} e^{-u_1 s_1 - u_2 s_2} s_2^{m_j}\, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\} \mathrm{d}u.$$

Evaluating the numerator of (22) the same reasoning yields a formula asymptotic to

$$\sum_{\mathbf{p}\in\mathcal{P}} \int_{\mathbb{R}^2_+} \frac{u_1^{n-1} u_2^{m-1}}{\Gamma(n)\Gamma(m)} e^{-\psi_h(\underline{u})} \prod_{(i,j)\in\Delta_{\mathbf{p}}} \left\{ \int_{A_i\times B_j} \int_{\mathbb{R}^2_+} e^{-(h_1(x)+u_1)s_1 - (h_2(x)+u_2)s_2} s_1^{n_i} s_2^{m_j}\, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\}$$
$$\prod_{(i,j)\in\Delta^1_{\mathbf{p}}} \left\{ \int_{A_i\times B_{c+1}} \int_{\mathbb{R}^2_+} e^{-(h_1(x)+u_1)s_1 - (h_2(x)+u_2)s_2} s_1^{n_i}\, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\}$$
$$\prod_{(i,j)\in\Delta^2_{\mathbf{p}}} \left\{ \int_{A_{k+1}\times B_j} \int_{\mathbb{R}^2_+} e^{-(h_1(x)+u_1)s_1 - (h_2(x)+u_2)s_2} s_2^{m_j}\, \rho(\mathrm{d}s) G_0(\mathrm{d}x) \right\} \mathrm{d}u.$$

where $\psi_h(\underline{u}) = \int_{\mathbb{X}^2} \int_{\mathbb{R}_+^2} \left(1 - e^{-(h_1(x)+u_1)s_1-(h_2(x)+u_2)s_2}\right) \rho(\mathrm{d}s)G_0(\mathrm{d}x)$. Note that

$$
\begin{aligned}
1 - e^{-(h_1(x)+u_1)s_1-(h_2(x)+u_2)s_2} &= e^{-u_1s_1-u_2s_2}\left[e^{u_1s_1+u_2s_2} - 1 + 1 - e^{-h_1(x)s_1-h_2(x)s_2}\right] \\
&= \left[1 - e^{-u_1s_1-u_2s_2}\right] + \left[1 - e^{-h_1(x)s_1-h_2(x)s_2}\right],
\end{aligned}
$$

so that

$$
\begin{aligned}
e^{-\psi_h(\underline{u})} &= e^{-\psi_b(\underline{u})} e^{-\int_{\mathbb{X}^2}\int_{\mathbb{R}_+^2}\left[1-e^{-h_1(x)s_1-h_2(x)s_2}\right]\rho(\mathrm{d}s)G_0(\mathrm{d}x)} \\
&= e^{-\psi_b(\underline{u})} \mathbb{E}\left[e^{-\int_{\mathbb{X}^2} h_1(x)\,\hat{\mu}_1(\mathrm{d}x) - \int_{\mathbb{X}^2} h_2(x)\,\hat{\mu}_2(\mathrm{d}x)}\right].
\end{aligned}
$$

Furthermore

$$
G_0(A_h \times B_r) = \epsilon \frac{G_0(A_h \times B_r)}{\epsilon} \sim \epsilon g_{h,r}, \quad 1 \le i \le c, 1 \le j \le k,
$$

and

$$
G_0(A_h \times \mathrm{d}x) \sim \epsilon g_{h,c+1} Q_{X_h^*}(\mathrm{d}x), \quad G_0(\mathrm{d}x \times B_r) \sim \epsilon g_{k+1,r} P_{Y_r^*}(\mathrm{d}x).
$$

Thus, evaluating the limit in (22) we get

$$
\begin{aligned}
&\mathbb{E}\left[e^{-\int_{\mathbb{X}^2} h_1(x)\,\tilde{\mu}_1(\mathrm{d}x) - \int_{\mathbb{X}^2} h_2(x)\,\tilde{\mu}_2(\mathrm{d}x)} \mid (X_i)_{i\ge1}^n, (Y_j)_{j\ge1}^m\right] = \\
&\times \sum_{p\in\mathcal{P}} \int_{\mathbb{R}_+^2} \mathbb{E}\left[e^{-\int_{\mathbb{X}^2} h_1(x)\,\hat{\mu}_1(\mathrm{d}x) - \int_{\mathbb{X}^2} h_2(x)\,\hat{\mu}_2(\mathrm{d}x)}\right] \times \\
&\times \prod_{(i,j)\in\Delta_p} \int_{\mathbb{R}_+^2} e^{-h_1(X_i^*,Y_j^*)s_1-h_2(X_i^*,Y_j^*)s_2} \frac{s_1^{n_i} s_2^{m_j} e^{-u_1s_1-u_2s_2}\rho(\mathrm{d}s)}{\tau_{n_i,m_j}(\underline{u})} \times \\
&\times \prod_{(i,j)\in\Delta_p^1} \int_{\mathbb{X}} \int_{\mathbb{R}_+^2} e^{-h_1(X_i^*,x_2)s_1-h_2(X_i^*,x_2)s_2} \frac{s_1^{n_i} e^{-u_1s_1-u_2s_2}\rho(\mathrm{d}s)}{\tau_{n_i,0(\underline{u})}} Q_{X_i^*}(\mathrm{d}x_2) \times \\
&\times \prod_{(i,j)\in\Delta_p^2} \int_{\mathbb{X}} \int_{\mathbb{R}_+^2} e^{-h_1(x_1,Y_2^*)s_1-h_2(x_1,Y_2^*)s_2} \frac{s_2^{m_j} e^{-u_1s_1-u_2s_2}\rho(\mathrm{d}s)}{\tau_{0,m_j(\underline{u})}} P_{Y_j^*}(\mathrm{d}x_1) \times \\
&\times \left(\frac{\int_{\mathbb{R}_+^2} u_1^{n-1}u_2^{m-1} \prod_{(i,j)\in p} g_{i,j}\tau_{n_i,m_j}(\underline{u})e^{-\psi_b(\underline{u})}\,\mathrm{d}\underline{u}}{\sum_{\mathbf{q}\in\mathcal{P}} \int_{\mathbb{R}_+^2} u_1^{n-1}u_2^{m-1} \prod_{(i,j)\in\mathbf{q}} g_{i,j}\tau_{n_i,m_j}(\underline{u})e^{-\psi_b(\underline{u})}\mathrm{d}\underline{u}}\right) \times \\
&\times \frac{u_1^{n-1}u_2^{m-1} \prod_{(i,j)\in p} \tau_{n_i,m_j}(\underline{u})e^{-\psi_b(\underline{u})}\,\mathrm{d}u}{\int_{\mathbb{R}_+^2} u_1^{n-1}u_2^{m-1} \prod_{(i,j)\in p} \tau_{n_i,m_j}(\underline{u})e^{-\psi_b(\underline{u})}\,\mathrm{d}\underline{u}},
\end{aligned}
$$

as desired. □

Proof of Corollary 6

Proof. We use the shorthand notation $\mu_1(f) = \int_{\mathbb{X}} f(x)\,\tilde{\mu}_1(\mathrm{d}x)$ for any measurable function $f : \mathbb{X} \to \mathbb{R}$ such that $\mu_1(|f|) < \infty$. Letting U be the set of latent variables of Theorem 13, i.e.

$U = (p, U_1, U_2, Z^x, Z^y)$ for any $y_1, \dots, y_n \in (0,1)$ and $A_1, \dots, A_n \in \mathcal{X}^2$ we get

$$\begin{aligned}&\mathbb{P}\left(\tilde{P}_1(A_1) \le y_1, \dots, \tilde{P}_1(A_n) \le y_n \mid U, (X_i)_{i=1}^n, (Y_j)_{j=1}^m\right)\\ =&\mathbb{P}\left(\tilde{\mu}_1(\mathbb{1}_{A_1} - y_1) \le 0, \dots, \tilde{\mu}_1(\mathbb{1}_{A_n} - y_n) \le 0 \mid U, (X_i)_{i=1}^n, (Y_j)_{j=1}^m\right].\end{aligned}$$

The result follows since the finite dimensional distributions of $\tilde{P}_1$ given U, $(X_i)_{i=1}^n$, and $(Y_j)_{j=1}^m$ coincide with the ones of the normalized posterior distribution of $\tilde{\mu}_1$, given U, $(X_i)_{i=1}^n$, and $(Y_j)_{j=1}^m$. □

Proof of Theorem 14

Proof. Set $\underline{H} = (p, U_1, U_2)$ with domain D. Then

$$\begin{aligned}\mathbb{P}(X_{n+1} \in \mathrm{d}x \mid (X_i)_{i=1}^n, (Y_j)_{j=1}^m) &= \mathbb{E}[P_1(\mathrm{d}x) \mid (X_i)_{i=1}^n, (Y_j)_{j=1}^m]\\ &= \int_D \mathbb{E}[P_1(\mathrm{d}x) \mid \underline{H} = \underline{h}, (X_i)_{i=1}^n, (Y_j)_{j=1}^m]\, F(\mathrm{d}\underline{v}),\end{aligned}$$

where $F(\cdot)$ is the posterior distribution of $\underline{H}$, with $\underline{h} = (p, u_1, u_2)$. Recalling the notation in Corollary 6 we have

$$\begin{aligned}\mathbb{E}[P_1(\mathrm{d}x) \mid \underline{H} = \underline{h}, (X_i)_{i=1}^n, (Y_j)_{j=1}^m] &= \mathbb{E}\left[\frac{\hat{\mu}_1(\mathrm{d}x \times \mathbb{X})}{R}\right] + \mathbb{E}\left[\frac{\sum_{(i,j)\in\Delta_p} J^1_{i,j}\delta_{X_i^*}}{R}\right] +\\ &+ \mathbb{E}\left[\frac{\sum_{(i,j)\in\Delta_p^1} J^1_{i,c+1}\delta_{X_i^*}}{R}\right] + \mathbb{E}\left[\frac{\sum_{(i,j)\in\Delta_p^2} J^1_{k+1,j}\delta_{Z_j^y}}{R}\right] = \sum_{k=1}^4 I_k,\end{aligned}$$

where $R = T_1 + \sum_{(i,j)\in\Delta_p} J^1_{i,j} + \sum_{(i,j)\in\Delta_p^1} J^1_{i,c+1} + \sum_{(i,j)\in\Delta_p^2} J^1_{k+1,j}$.
Set $S = \sum_{(i,j)\in\Delta_p} J^1_{i,j} + \sum_{(i,j)\in\Delta_p^1} J^1_{i,c+1} + \sum_{(i,j)\in\Delta_p^2} J^1_{k+1,j}$ and exploit the conditional independence between J^1_{ij} and $\hat{\mu}_1$ to obtain

$$\begin{aligned}I_1 &= \int_{\mathbb{R}_+} \mathbb{E}\left[e^{-vS}\right] \mathbb{E}\left[\hat{\mu}_1(\mathrm{d}x \times \mathbb{X}) e^{-vT_1}\right] \mathrm{d}v\\ &= \theta P_0(\mathrm{d}x) \int_{\mathbb{R}_+} \left(\prod_{(i,j)\in p} \frac{\tau_{n_i,m_j}(u_1+v, u_2)}{\tau_{n_i,m_j}(u_1, u_2)}\right) \tau_{1,0}(u_1+v, u_2) e^{-\psi_b^u(v,0)}\, \mathrm{d}v,\end{aligned}$$

where $\psi_b^{\underline{u}}(\lambda_1, \lambda_2)$ is the Laplace exponent of $(\hat{\mu}_1, \hat{\mu}_2)$ in Theorem 13. Observing that $\psi_b^{\underline{u}}(v, 0) + \psi(u_1, u_2) = \psi(u_1 + v, u_2)$ and denoting with $L(\cdot)$ the distribution of $\mathbf{p}$, we obtain

$$\begin{aligned}
\xi_0 &= \int_D I_1 \, F(\mathrm{d}\underline{u}) \\
&= \theta Q_0(\mathrm{d}x) \int \int_{\mathbb{R}_+^3} \Bigg\{ u_1^{n-1} u_2^{m-1} \left(\prod_{(i,j) \in p} \tau_{n_i, m_j}(u_1 + v, u_2) \right) \tau_{1,0}(u_1 + v, u_2) \times \\
&\times e^{-\psi(u_1+v, u_2)} \, \mathrm{d}u_1 \mathrm{d}u_2 \mathrm{d}v L(\mathrm{d}p) \Bigg\} \\
&= \frac{\theta Q_0(\mathrm{d}x)}{n} \int \int_{\mathbb{R}_+^2} u_1^n u_2^{m-1} \left(\prod_{(i,j) \in p} \tau_{n_i, m_j}(u_1, u_2) \right) \tau_{1,0}(u_1, u_2) e^{-\psi(u_1, u_2)} \, \mathrm{d}u_1 \mathrm{d}u_2 L(\mathrm{d}p) \\
&= \frac{\theta Q_0(\mathrm{d}x)}{n} \int_D u_1 \tau_{1,0}(u_1, u_2) \, F(\mathrm{d}\underline{u}),
\end{aligned}$$

where the second equality follows from the change of variables $(w, z) = (u_1 + v, u_1)$. The proof for the remaining weights follows along the same lines and leads to

$$\xi_i^x = \frac{1}{n} \int_D u_1 \left[\frac{\tau_{n_i+1, m_j}(u_1, u_2)}{\tau_{n_i, m_j}(u_1, u_2)} + \frac{\tau_{n_i+1, 0}(u_1, u_2)}{\tau_{n_i, 0}(u_1, u_2)} \right] F(\mathrm{d}\underline{u})$$

and

$$\xi_i^y = \frac{1}{n} \int_D u_1 \frac{\tau_{1, m_j}(u_1, u_2)}{\tau_{0, m_j}(u_1, u_2)} F(\mathrm{d}\underline{u}).$$

The weights for Y_{m+1} can be computed in an analogous fashion. □

A2 Additional material for Section 3.2

A toy example of borrowing of information

Classical borrowing of information across samples is typically associated to positive correlation across observations in different populations and, as a consequence, it induces shrinkage of the predictions. Let us consider the toy situation in which observations coming from two different populations have been collected and a normal model is assumed

$$\begin{aligned}
X_i \mid \mu_x &\overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_x, 1) \qquad \text{for } i = 1, \dots, n \\
Y_j \mid \mu_y &\overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu_y, 1) \qquad \text{for } j = 1, \dots, m
\end{aligned}$$

To obtain a working model, one has to specify a certain prior over μ_x and μ_y. The main typical strategies one may employ are the following:

- Modeling μ_x and μ_y as independent, which ultimately means that we do not consider the information coming from one population to be relevant for inference on the other.

- Modeling μ_x and μ_y as dependent, which induces borrowing of information. This typically reflects the idea that, if the observed values of $Y_1, \dots, Y_m$ are on average higher than our prior guess on μ_y, then we should upwards revise our belief on μ_x and our prediction for X_1.

To clarify this last point, we compare a typical strategy used to perform borrowing of information, which is provided by the following hierarchy

$$\begin{aligned}\mu_x \mid \mu_0 &\sim \mathrm{N}(\mu_0,\, 1)\\ \mu_y \mid \mu_0 &\sim \mathrm{N}(\mu_0,\, 1)\\ \mu_0 &\sim \mathrm{N}(\nu,\, 1)\end{aligned} \tag{24}$$

with the case of independent priors, namely

$$\begin{gathered}\mu_x \sim \mathrm{N}(\nu,\, 2) \qquad\qquad \mu_y \sim \mathrm{N}(\nu,\, 2)\\ \mu_x \perp \mu_y\end{gathered} \tag{25}$$

where the variance is chosen to match the marginal distributions of the hierarchical specification. We assume that only the sample $(Y_1, \dots, Y_m)$ has been observed and we discuss its impact on the posterior distribution of μ_x and on the predictive distribution of X_1 under the two specifications. Under independence in (25), one obviously has

$$p(\mu_x \mid (Y_j)_{j=1}^m) = \mathrm{N}(\nu,\, 2)$$

while under model (24) the new distribution of μ_x is

$$\begin{aligned}p(\mu_x \mid (Y_j)_{j=1}^m) &\propto \int_{\mathbb{R}} p(\mu_x \mid \mu_0)\, p(\mu_0 \mid (Y_j)_{j=1}^m) \mathrm{d}\mu_0\\ &= \mathrm{N}\left(\frac{1}{2m+1}\nu + \frac{2m}{2m+1}\frac{\nu+\bar{y}}{2},\, 1 + \frac{m+1}{2m+1}\right),\end{aligned}$$

where $\bar{y}$ denotes the empirical average of $Y_1, \dots, Y_m$, and

$$\mathbb{E}[X_1 \mid (Y_j)_{j=1}^m] = \mathbb{E}[\mu_x \mid (Y_j)_{j=1}^m] = \nu + \frac{m}{2m+1}(\bar{y} - \nu)$$

Therefore, when $\bar{y} > \nu$ the borrowing results in an increase of the estimate for μ_x and of the prediction for X_1, while if $\bar{y} < \nu$ the borrowing of information induces the opposite effect. The shrinking behaviour is ultimately a consequence of the fact that the hierarchical prior in (24) induces positive correlation across X_i and Y_j. However, what we show in the main paper is that classical shrinkage of the estimates is not the only way to borrow information within partially exchangeable populations, neither necessarily the best one.

Example of correlation between FuRBI priors on Borel set

Consider a pair of n-FuRBI priors with equal jumps (see Example 4 in the main document), where the baseline distribution G_0 is given by a bivariate normal with zero mean, unit variances and correlation $\rho \in \{-0.99, -0.5, 0, 0.5, 0.99\}$. In Figure 13 we depict the correlations on sets of the form $(-\infty, x]$, with $x \in [-5, 5]$ and for each value of the correlation. Notice that such correlation may be of particular interest in survival settings, where the distribution function is often the main focus.

When $\rho = 0$, the correlation is equal to 0 as expected, since $G_0(A \times A) = Q_0(A)^2$ and the numerator of the formula in Proposition 8 vanishes. For values of ρ different from 0, the

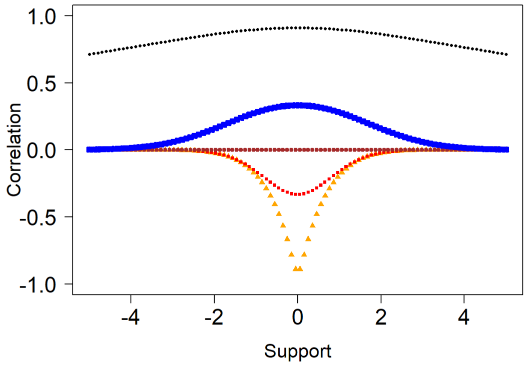


Figure 13: Correlation on Borel sets of the form $(-\infty, x]$, with $x \in [-5, 5]$. The four lines, from bottom to top, correspond to $\rho \in \{-0.99, -0.5, 0, 0.5, 0.99\}$.

correlation is symmetric around 0, due to the symmetry of the Gaussian distribution, and different signs indicate opposite behaviours: therefore, $\rho < 0$ implies negative correlation on such Borel sets.

However, note that a different sign does not mean a completely specular behaviour: for instance the correlation with $\rho = 0.99$ is higher in absolute value than the one with $\rho = -0.99$. This is due to the fact that it is somewhat impossible to have strictly negative correlation on all Borel sets. Intuitively, if the two priors have high negative correlation on $(-\infty, 0]$, it means that one of them has much larger mass on $(-\infty, 0]$ and the other on $(0, +\infty)$: therefore, both priors will have a high mass on $(-\infty, a]$, with a large positive number, so that the correlation can not attain again large negative values.

Finally, if $\rho \to 1$, then the correlation converges to the constant function 1, that is the value obtained with equal atoms: indeed, the two priors will have equal jumps and linearly dependent atoms (see Corollary 1).

Algorithms for posterior inference

In this section we address the issue of sampling from the posterior distribution. In discrete nonparametric models, we need to distinguish whether the random probability measures are directly applied to the data or rather convoluted with a suitable kernel (known as *mixture* model, see Section 3.2.6).

Nevertheless, from a computational perspective, if the first problem is solved the second one can be tackled in a similar way: it is indeed easy to propose a Gibbs sampler that alternates sampling of suitable latent variables and the posterior distribution given data originated by the random probability measure.

Therefore, in the following three sections, we assume to collect observations from

$$(X_i, Y_j) \mid (P_1, P_2) \overset{\text{i.i.d.}}{\sim} P_1 \times P_2 \qquad (P_1, P_2) \sim Q \tag{26}$$

Marginal posterior samplers

The first approach is to directly simulate the trajectories of (P_1, P_2) from its posterior, giving rise to so–called conditional algorithms. See, e.g, Ishwaran and James (2001); Walker (2007); Papaspiliopoulos and Roberts (2008); Arbel and Prünster (2017). Conditional samplers for the n-FuRBI priors can be found in Sections S3.2-3 below.

Alternatively, and this is the route followed in this section, one can use marginal algorithms, that integrate out the random measures and sample sequentially from the predictive distributions (see, for instance, Neal, 2000).

Given $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$ and using the results in Theorem 2, we can sample iteratively new observations from P_1 as follows

1. Compute weights ξ_0, $\{\xi_i^x\}$ and $\{\xi_j^y\}$ from $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$.
2. Draw X_{n+1} from $m(\mathrm{d}x) = \xi_0 Q_0(\mathrm{d}x) + \sum_{i=1}^k \xi_i^x \delta_{X_i^*}(\mathrm{d}x) + \sum_{j=1}^c \xi_j^y P_{Y_j^*}(\mathrm{d}x)$.

The algorithm is straightforward, but relies on the computation of the weights at point (a): this is not optimal, since in general the explicit evaluation can be demanding. Nonetheless, Theorem 1 and Corollary 2 show that, conditionally on a suitable set of latent variables, the posterior representation simplifies greatly. Indeed, given $((X_i)_{i=1}^n, (Y_j)_{j=1}^m, U_1, U_2, p)$, the predictive distribution of the first sample is

$$\begin{aligned} m(\mathrm{d}x) \propto \theta \tau_{1,0}(U_1, U_2) Q_0(\mathrm{d}x) &+ \sum_{(i,j)\in\Delta_p} \frac{\tau_{n_i+1,m_j}(U_1,U_2)}{\tau_{n_i,m_j}(U_1,U_2)} \delta_{X_i^*}(\mathrm{d}x) \\ + \sum_{(i,j)\in\Delta_p^1} \frac{\tau_{n_i+1,0}(U_1,U_2)}{\tau_{n_i,0}(U_1,U_2)} \delta_{X_i^*}(\mathrm{d}x) &+ \sum_{(i,j)\in\Delta_p^2} \frac{\tau_{1,m_j}(U_1,U_2)}{\tau_{0,m_j}(U_1,U_2)} P_{Y_j^*}(\mathrm{d}x). \end{aligned} \tag{27}$$

Those new weights, whose derivation can be found in Section S1.4, are easier to compute, as the next example shows.

Example 10 (Inverse Gaussian n-FuRBI with equal jumps)**.** *In this case*

$$\tau_{n,m}(u_1, u_2) = \int_{\mathbb{R}} s^{n+m} e^{-(u_1+u_2)s} \rho(\mathrm{d}s) := \tau_{n+m}(u_1 + u_2),$$

where $\rho(\mathrm{d}s)$ is the common marginal jump intensity. If the Lévy intensity is $v(\mathrm{d}s, \mathrm{d}x) = e^{-s/2}/(s^{3/2}\sqrt{2\pi})\mathrm{d}s\, \alpha(\mathrm{d}x)$ the resulting normalized CRM corresponds to the normalized inverse Gaussian process (Lijoi et al., 2005). We obtain $\tau_j(u) = 2^{j-1}\Gamma(j - 1/2)/(\sqrt{\pi}(2u + 1)^{j-1/2})$, where $u = u_1 + u_2$. Thus, conditionally on the latent variables, we have

$$\begin{aligned} m(\mathrm{d}x) \propto \theta Q_0(\mathrm{d}x) &+ \frac{2}{\sqrt{2U+1}} \sum_{(i,j)\in\Delta_p} \left(n_i + m_j - \frac{1}{2}\right) \delta_{X_i^*}(\mathrm{d}x) \\ + \frac{2}{\sqrt{2U+1}} \sum_{(i,j)\in\Delta_p^1} \left(n_i - \frac{1}{2}\right) \delta_{X_i^*}(\mathrm{d}x) &+ \frac{2}{\sqrt{2U+1}} \sum_{(i,j)\in\Delta_p^2} \left(m_j - \frac{1}{2}\right) P_{Y_j^*}(\mathrm{d}x), \end{aligned}$$

where $U = U_1 + U_2$. Sampling from this mixture is straightforward.

Thus we can derive a second marginal algorithm.

1. Draw (U_1, U_2, p) from their conditional distributions specified in Section 5.
2. Draw X_{n+1} from $m(\mathrm{d}x)$ in (27).

However, even the full conditional distribution of p may not always be available in closed form, and it may be computationally intensive to evaluate, since it may have a very large support. When this is the case, we may encode the latent clustering structure in a more convenient way introducing two arrays of latent variables $\mathcal{C}_x = (c_{i,x})_{i\geq 1}$ and $\mathcal{C}_y = (c_{j,y})_{j\geq 1}$ such that $c_{i,x} = c_{i',x}$ denotes a tie between X_i and $X_{i'}$, $c_{j,y} = c_{j',y}$ denotes a tie between Y_j and $Y_{j'}$, while $c_{i,x} = c_{j,y}$ denotes a hyper-tie between X_i and Y_j. Moreover, we reorder the unique values in $\underline{X}_n^*$ and

$\underline{Y}^*_m$, so that $X^*_c = X_i$ if and only if $c_{i,x} = c$ and $Y^*_c = Y_j$ if and only if $c_{j,y} = c$. Therefore, $\mathbb{P}(c_{n+1,x} = c \mid \mathcal{C}_x, \mathcal{C}_y, \underline{X}^*_n, \underline{Y}^*_m)$ is

$$\begin{cases} \mathbb{P}(X_{n+1} = X^*_c \mid \mathcal{C}_x, \mathcal{C}_y, \underline{X}^*_n, \underline{Y}^*_m), & \text{for } c \in \mathcal{C}_x \\ \int \mathbb{P}(X_{n+1} = x \mid \mathcal{C}_y, \underline{Y}^*_m)\, p_{Y^*_c}(x)\mathrm{d}x, & \text{for } c \in \mathcal{C}_y \setminus \mathcal{C}_x \\ \int \mathbb{P}(X_{n+1} = x)\, q_0(x)\mathrm{d}x, & \text{otherwise} \end{cases}$$

Finally, the distribution of p, given $\mathcal{C}_x$ and $\mathcal{C}_y$, is degenerate. Moreover, the posterior distribution of (U_1, U_2) given p is equal to the posterior distribution of (U_1, U_2) given $\mathcal{C}_x$ and $\mathcal{C}_y$. Therefore, we may build a marginal algorithm sampling $\mathcal{C}_x$ and $\mathcal{C}_y$ instead of p, without modifying the full conditional distribution for U_1 and U_2. The final marginal algorithm boils down to

1. Draw (U_1, U_2) and $c_{n+1,x}$

2. Sample X_{n+1} from $m(\mathrm{d}x) = \begin{cases} \delta_{X^*_{c_{n+1,x}}}(\mathrm{d}x), & \text{if } c_{n+1,x} \in \mathcal{C}_x \\ P_{Y^*_{c_{n+1,x}}}(\mathrm{d}x), & \text{if } c_{n+1,x} \in \mathcal{C}_y \setminus \mathcal{C}_x \\ Q_0(\mathrm{d}x), & \text{otherwise} \end{cases}$

The advantage of such approach is twofold. First, we do not need to sample directly the full conditional distribution of p. Second, when the algorithm is applied to mixture models, as in section 6, sampling the unique values, instead of single observations, improves the mixing of the algorithm (cfr. Neal, 2000).

Conditional posterior sampler based on the law of the CRV

To develop a conditional algorithm, we can sample from the distribution of (μ_1, μ_2) and then normalize each draw to get an approximate realization of the random probabilities. Here we develop a general conditional sampler based on this approach that can be tailored to specific choices of the intensity in the prior.

By Theorem 13, we know that a posteriori $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2)$ is the sum of two components, that we call μ_{obs} and $\hat{\mu}$ and are such that

$$\mu_{obs} = \sum_{(i,j)\in\Delta_p} J_{i,j}\delta_{\left(X^*_i, Y^*_j\right)} + \sum_{(i,j)\in\Delta^1_p} J_{i,c+1}\delta_{\left(X^*_i, Z^x_i\right)} + \sum_{(i,j)\in\Delta^2_p} J_{k+1,j}\delta_{\left(Z^y_j, Y^*_j\right)}.$$

where $J_{i,j} = (J^1_{i,j}, J^2_{i,j})$, and

$$\hat{\mu} = \left(\sum_{h=1}^{+\infty} S^1_h \delta_{(V_h, W_h)}, \sum_{h=1}^{+\infty} S^2_h \delta_{(V_h, W_h)}\right)$$

is a CRV with Lévy intensity $e^{-U_1 s_1 - U_2 s_2}\rho(\mathrm{d}s_1, \mathrm{d}s_2)G_0(\mathrm{d}x)$. Denote the marginal and joint tail integrals of $\hat{\mu}$ as

$$N_1(s) = \int_s^{+\infty}\int_0^{+\infty} e^{-U_1 s_1 - U_2 s_2}\rho(\mathrm{d}u_1, \mathrm{d}u_2), \quad N_2(s) = \int_0^{+\infty}\int_s^{+\infty} e^{-U_1 s_1 - U_2 s_2}\rho(\mathrm{d}u_1, \mathrm{d}u_2)$$

and

$$N(s_1, s_2) = \int_{s_1}^{+\infty} \int_{s_2}^{+\infty} e^{-U_1 s_1 - U_2 s_2} \rho(\mathrm{d}u_1, \mathrm{d}u_2).$$

Lastly, define the correspondent Lévy copula as $F(x, y) = N(N_1^{-1}(x), N_2^{-1}(y))$. If $F(x, y)$ is continuous on $[0, +\infty]^2$, the iterative conditional sampler based on the Ferguson and Klass algorithm (Ferguson and Klass, 1972) reads

(a) Generate μ_{obs} as follows

 (a1) Generate $(U_1, U_2, \mathbf{p})$ from the distributions specified in Section 5;

 (a2) Generate $J_{i,j} = (J_{i,j}^1, J_{i,j}^2)$ from the distributions specified in Theorem 13;

 (a3) Generate Z_i^x and Z_j^y from the distributions specified in Section 5.

(b) Generate an approximation of $\hat{\mu}$, given by $\left(\sum\limits_{h=1}^{M} S_h^1 \delta_{(V_h, W_h)}, \sum\limits_{h=1}^{M} S_h^2 \delta_{(V_h, W_h)} \right)$ as follows

 (b1) Generate $\xi_1^x, \dots, \xi_M^x$ from a Poisson Process with unit rate;

 (b2) Generate $\xi_1^y, \dots, \xi_M^y$ from $\xi_h^y \sim \frac{\partial}{\partial x} F(x, \xi) \Big|_{x = \xi_h^x}$

 (b3) Determine (S_h^1, S_h^2) solving

$$\xi_h^x = N_1(S_h^1) \qquad \xi_h^y = N_2(S_h^2)$$

 (b4) Generate (V_h, W_h) from G_0.

(c) Obtain a draw from P_1 as follows

$$P_1 \approx \frac{\sum\limits_{h=1}^{M} S_h^1 \delta_{V_h} + \sum_{(i,j) \in \Delta_p} J_{i,j}^1 \delta_{X_i^*} + \sum_{(i,j) \in \Delta_p^1} J_{i,c+1}^1 \delta_{X_i^*} + \sum_{(i,j) \in \Delta_p^2} J_{k+1,j}^1 \delta_{Z_j^y}}{\sum\limits_{h=1}^{M} S_h^1 + \sum_{(i,j) \in \Delta_p} J_{i,j}^1 + \sum_{(i,j) \in \Delta_p^1} J_{i,c+1}^1 + \sum_{(i,j) \in \Delta_p^2} J_{k+1,j}^1}.$$

An analogous approximation can be computed for $\tilde{p}_2$.

Conditional posterior sampler for gamma process with equal jumps

Alternatively, a second strategy for conditional algorithms is to sample approximate draws from the posterior distribution of the random probabilities (P_1, P_2). We provide an example for gamma FuRBI CRMs with equal jumps.

In the case of a process with equal jumps, we know from the definition that the measures in the product space are $p_1 = p_2 = p$. Therefore, posterior inference can be conducted without loss of generality on

$$p = \sum_{k \geq 1} \bar{W}_k \delta_{(\theta_k, \phi_k)}, \quad \text{with } (\theta_k, \phi_k) \overset{\text{i.i.d.}}{\sim} G_0(\cdot),$$

where $\{\bar{W}_k\}_k$ are the weights of a Dirichlet process, which can defined through the popular stick-breaking construction (Sethuraman, 1994). In this context, Ishwaran and James (2001) developed a conditional algorithm for hierarchical mixture models, called *blocked Gibbs* sampler, based on the approximation

$$p \approx \sum_{k=1}^{N} \bar{W}_k \delta_{(\theta_k,\phi_k)}, \quad \text{for large } N.$$

Exploiting the appealing analytical properties of the Dirichlet process, it is possible to devise simple formulae for the posterior distribution of the N jumps and N locations: see Section 5 of Ishwaran and James (2001) for more details.

Sampling from mixture models using marginal algorithms

Consider the mixture model defined in Section 3.2.6. Starting from the algorithms studied in the previous paragraphs, we devise a Gibbs sampler for drawing from the posterior distribution of $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$.

Denoting by $\boldsymbol{X}^t = (X_1^t, \dots, X_n^t)$ and $\boldsymbol{Y}^t = (Y_1^t, \dots, Y_n^t)$ the vectors sampled at step t, the algorithm reads

1. Initialize at random $\boldsymbol{X}^0$ and $\boldsymbol{Y}^0$.
2. For any $t \geq 1$ do:

   (b.1) Draw $(U_1, U_2, \mathbf{p})$ given $\boldsymbol{X}^{t-1}$ and $\boldsymbol{Y}^{t-1}$, from the distributions specified in Theorem 1.

   (b.2) Draw $\boldsymbol{X}_n$, given $(U_1, U_2, \mathbf{p})$ as follows: for any i sample X_i^t from

$$\begin{aligned} q(\mathrm{d}x \mid \boldsymbol{X}_{-i}^t) = q_{i,0}(U_1,U_2)P_0(\mathrm{d}x) &+ \sum_{(i,j)\in\Delta_{\boldsymbol{p}}} q_{i,j}(U_1,U_2)\delta_{X_i^*} \\ &+ \sum_{(i,j)\in\Delta_{\boldsymbol{p}}^1} q_{i,j}^1(U_1,U_2)\delta_{X_i^*}(\mathrm{d}x) + \sum_{(i,j)\in\Delta_{\boldsymbol{p}}} q_{i,j}^2(U_1,U_2)P_{Y_j^*}(\mathrm{d}x), \end{aligned}$$

   where $\boldsymbol{X}_{-i}^t = \left(X_1^t, \dots, X_{i-1}^t, X_{i+1}^{t-1}, \dots X_n^{t-1}\right)$, with unique values $\left(X_1^*, \dots, X_k^*\right)$ and multiplicities $(n_1, \dots, n_k)$. Analogously, $\left(Y_1^*, \dots, Y_c^*\right)$ denotes the unique values in $\boldsymbol{Y}^{t-1}$ with multiplicities $(m_1, \dots, m_c)$. The mixing proportions are given by

$$\begin{aligned} q_{i,0}(U_1,U_2) &\propto \theta\tau_{1,0}(U_1,U_2)\int_{\mathbb{X}} f(W_i \mid x)P_0(\mathrm{d}x), \\ q_{i,j}(U_1,U_2) &\propto \frac{\tau_{n_i+1,m_j}(U_1,U_2)}{\tau_{n_i,m_j}(U_1,U_2)} f(W_i \mid X_i^*), \\ q_{i,j}^1(U_1,U_2) &\propto \frac{\tau_{n_i+1,0}(U_1,U_2)}{\tau_{n_i,0}(U_1,U_2)} f(W_i \mid X_i^*), \\ q_{i,j}^2(U_1,U_2) &\propto \frac{\tau_{1,m_j}(U_1,U_2)}{\tau_{0,m_j}(U_1,U_2)} \int_{\mathbb{X}} f(W_i \mid x)P_{Y_j^*}(\mathrm{d}x) \end{aligned}$$

(c) Sample $\boldsymbol{Y}^t$ similarly to point (b).

Once a sample of $(X_i)_{i=1}^n$ and $(Y_j)_{j=1}^m$ is available, sampling new observations X_{n+1} and Y_{n+1} proceeds as explained in Section $S3.1$.

Additional simulation studies

Additional simulation scenarios

We consider the same setting of the numerical illustrations, with different data generating distributions. Formally we have

$$W_i \overset{\text{i.i.d.}}{\sim} p(\cdot - 10), \quad V_j \overset{\text{i.i.d.}}{\sim} p(\cdot - v),$$

where $v \in [-16, 16]$ and $p(\cdot)$ is a density function. In the main manuscript we let $p(\cdot) = N(\cdot \mid 0, 1)$, while here we consider three different choices

$$p_1(\cdot) = \text{Exp}(\cdot \mid 1), \quad p_2(\cdot) = 0.5N(\cdot \mid 5, 1) + 0.5N(\cdot \mid -5, 1), \quad p_3(\cdot) = t(\cdot \mid 3),$$

where $t(\cdot \mid q)$ denotes the density of a Student's t distribution with q degrees of freedom. We let $i = 1, \dots, 20$, $j = 1, \dots, 100$ and consider the same nonparametric models of Section 6.2, with Gaussian kernel. Therefore, the prior specification is misspecified in the first and third case, with different tail behaviours of the kernel with respect to the true data generating mechanism. This implies a more complex behaviour of the latent clustering structure: indeed the posterior distribution places positive mass to more than one clusters, in order to accommodate for the misspecification. The mean integrated error for the three cases is depicted in Figure 14, for different values of v. The interpretation is similar to the one discussed in Section 6.2: the FuRBI specification yields an advantage especially when v is far from 0, corresponding to the prior mean, and from 10, when the means of the two groups coincide. Indeed, in the first case the borrowing provides little information, while in the second one exchangeability holds.

The second setting, corresponding to the two-components mixture, apparently seems more problematic for the FuRBI model, which yields a less distinct advantage. Clearly, when v is close to zero the exchangeable and hierarchical models are favoured, since the two true distributions share one of the modes. Moreover, the availability of only 20 observations for the first group makes it more difficult to both detect the presence of two clusters and tune appropriately the correlation. Indeed, the left part of figure 15 depicts the error when 50 observations for the first group are collected: as expected, the performances of the FuRBI approach significantly improve.

Finally, the right part of figure 15 shows the error when the two distributions are different: the first group is endowed with a Student's t density, while the second one is exponentially distributed. Notice that the two groups are now very far in distributional sense, especially in terms of tail behaviour. The plot indicates an interesting trade-off: when v is far from the prior mean (i.e. 0) the FuRBI approach allows to alleviate the prior misspecification, otherwise borrowing information from very different distributions may be detrimental.

Logit stick-breaking prior and borrowing of information

Figure 16 is based on the same data of Section 3.2.6. See Rigon and Durante (2021) for the model and the associated algorithm. Once again, including a flexible dependence on the atoms allows to a better borrowing and thus density estimation.

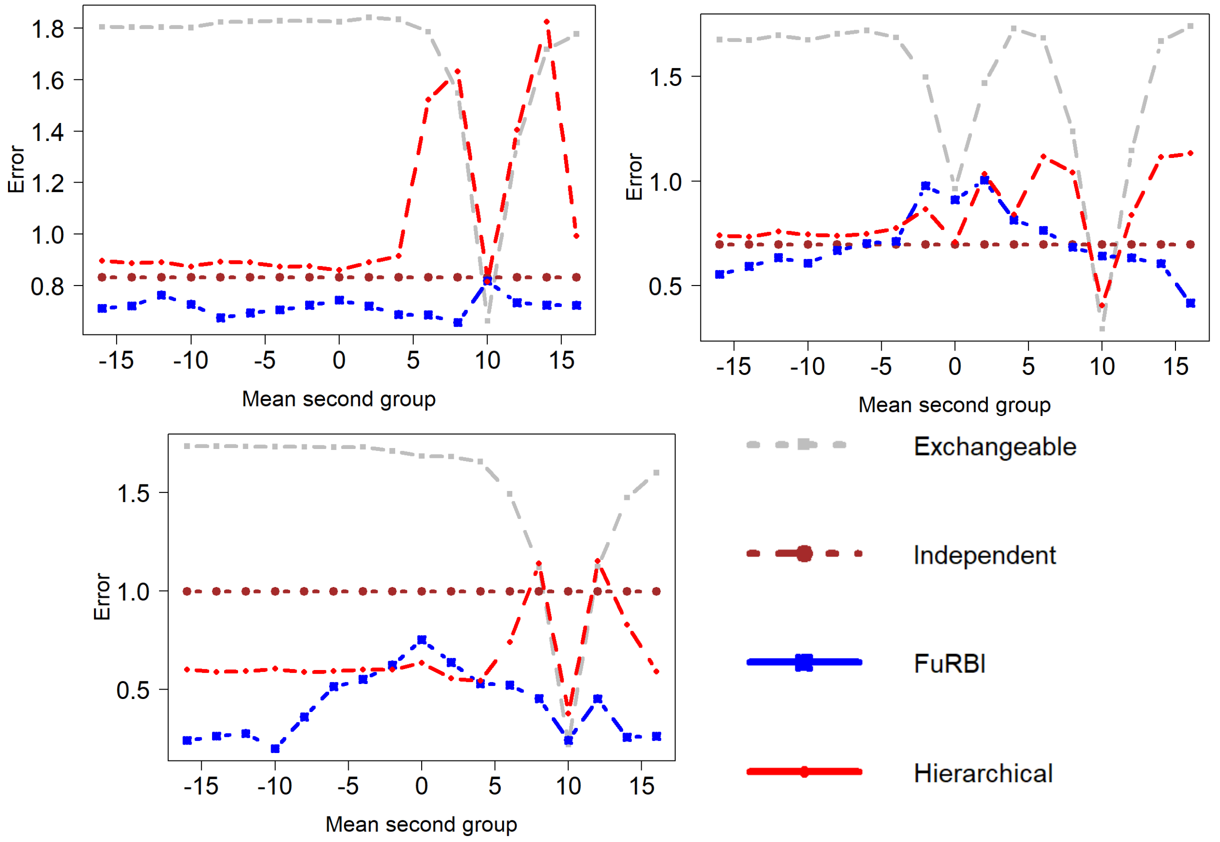


Figure 14: Mean integrated error (computed on a grid and as the median over 50 different samples) for the four models, as the true mean of the second group varies. Rotating clockwise from the top left panel: data generated from shifted exponential, mixtures of two Gaussians and shifted Student's t distributions.

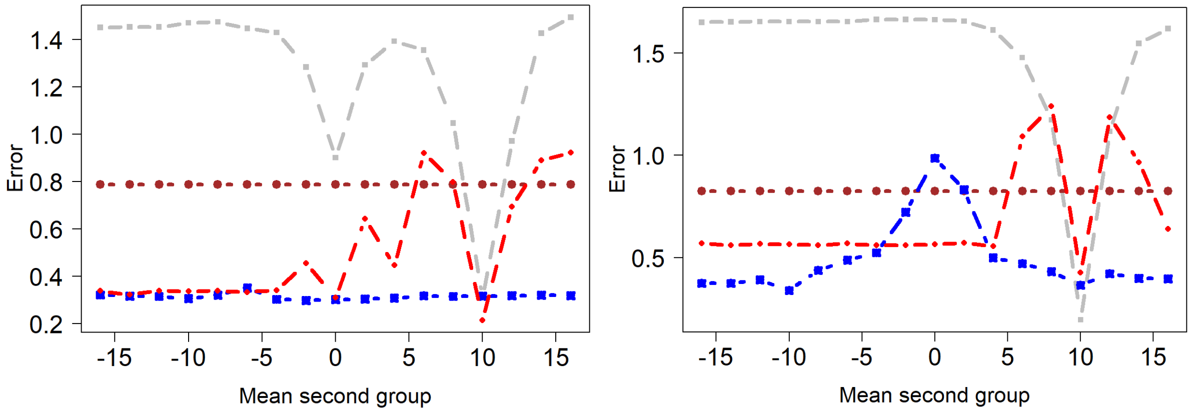


Figure 15: Mean integrated error (computed on a grid and as the median over 50 different samples) for the four models, as the true mean of the second group varies. Left: data generated from mixtures of two Gaussians (50 observations for the first group). Right: data generated from shifted Student's t (first group) and shifted exponential (second group) distributions.

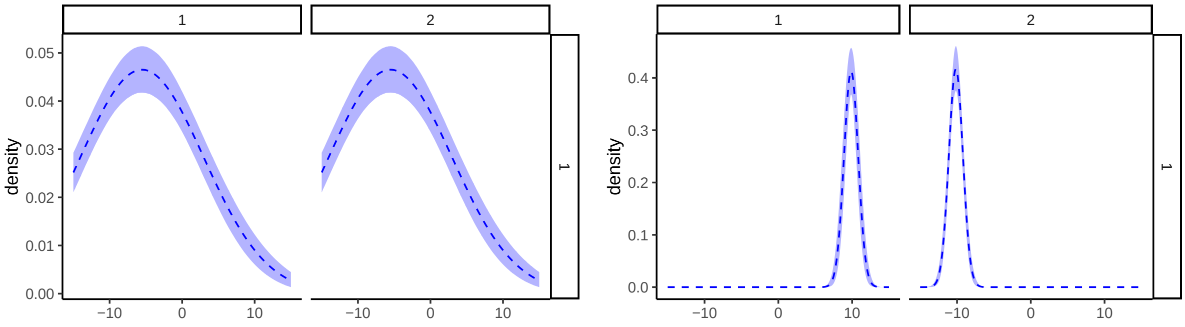
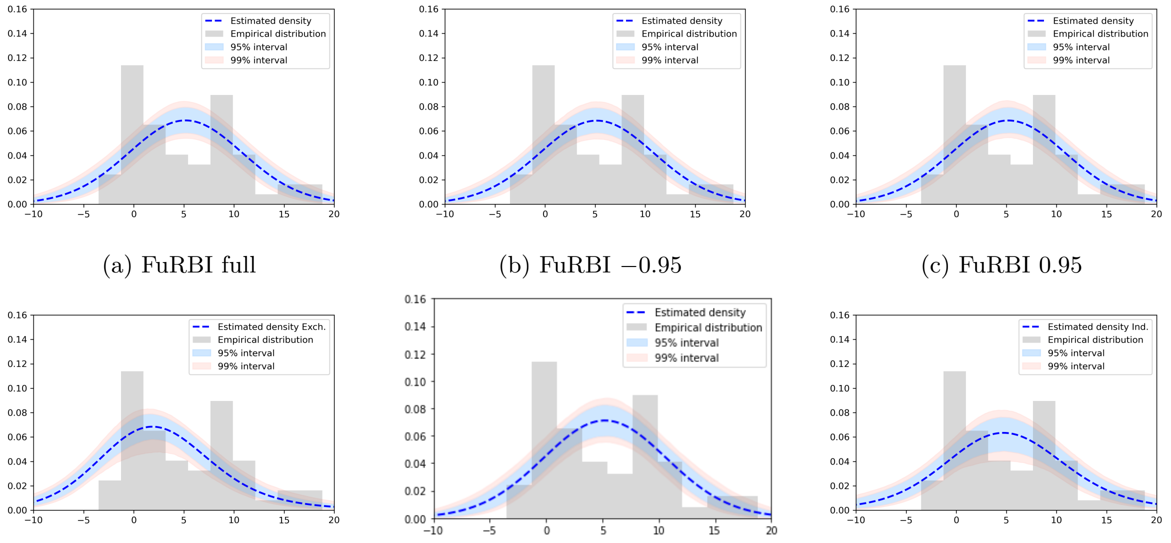


Figure 16: Left panel: density estimates for the logit stick-breaking model with only dependent weights, and thus, $\rho_0 = 1$. Right panel: density estimates for the logit stick-breaking model with dependent weights and atoms. Shaded areas denote 95% credible intervals. Data are simulated according to $W_i \overset{\text{i.i.d.}}{\sim} N(\cdot \mid 10, 1)$, for $i = 1, \ldots, 20$ (for sample n.1), and $V_j \overset{\text{i.i.d.}}{\sim} N(\cdot \mid -10, 1)$, for $j = 1, \ldots, 100$ (for sample n.2).



(a) FuRBI full

(b) FuRBI −0.95

(c) FuRBI 0.95

(d) Exchangeable model

(e) GM-dependent model

(f) Independent model

Figure 17: Density estimates for bonds returns.

Predicting stocks and bonds returns: additional results

Density estimation for bond returns

Sensitivity analysis

Figure 17 shows the results obtained with different specifications of the hyperparameters, which are

- Specification n.1: $\lambda_j = 0.1$, $\alpha_j = 3$, and $\beta_j = 3$, $j = 1, 2$,
- Specification n.2: $\lambda_j = 0.1$, $\alpha_j = 1.5$, and $\beta_j = 4.5$, $j = 1, 2$,
- Specification n.3: $\lambda_j = 0.01$, $\alpha_j = 0.1$, and $\beta_j = 0.2$, $j = 1, 2$.

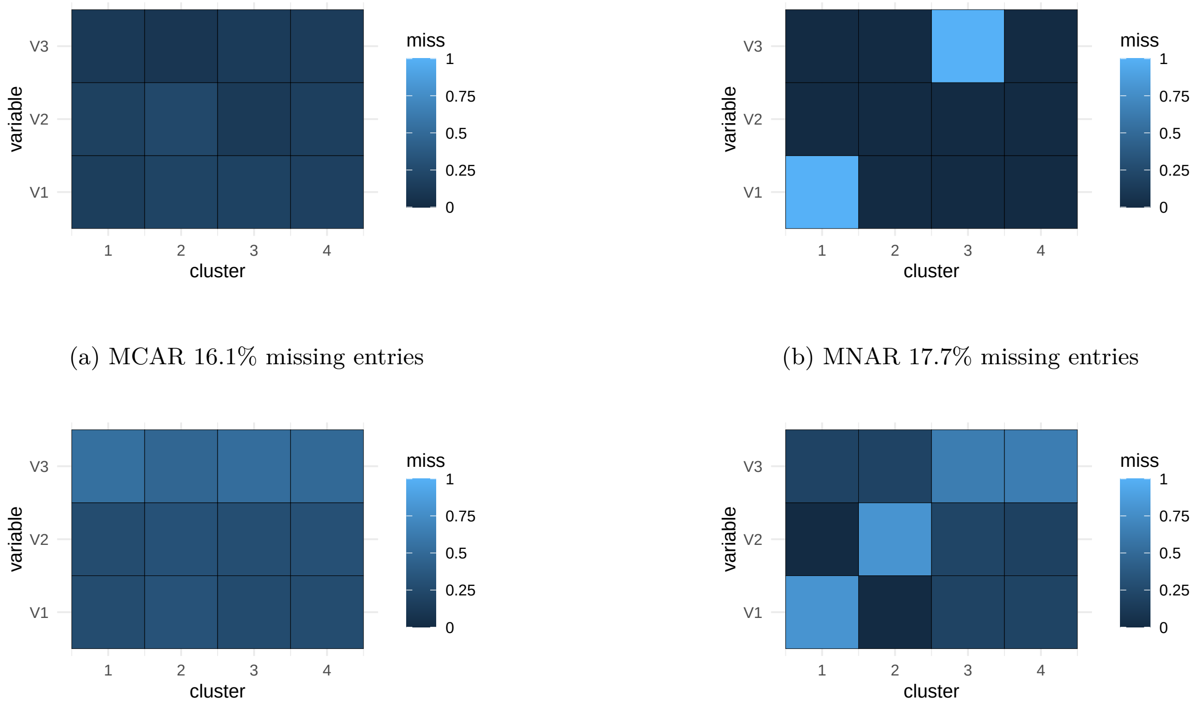
Clustering multivariate data with missing entries: additional details

Choosing the hyperparameters

Assume $P = 3$, as in the simulation study of Section 6.4: the general case follows accordingly. In this case $I = \{\emptyset, (1), (2), (3), (1,2), (2,3), (1,3), (1,2,3)\}$. In order to specify the prior, assumptions on the missing generating mechanism should be made. The missing completely at random (MCAR) assumption implies that each observation $W_i^{(x)}$, for $x \in I$, is the result of randomly eliminating entries from an (unobserved) complete observation W_i. For instance, $W_i^{(1)} = (w_{2,i}, w_{3,i})$ is obtained from a latent $W_i = (w_{1,i}, w_{2,i}, w_{3,i})$ after eliminating the first entry. Under this assumption the latent complete observations W_i are exchangeable, because the original value of W_i is independent from the mechanism that generates the missing values. Thus, there exists q such that $W_i \mid q \overset{\text{i.i.d.}}{\sim} q$ and q_x is the projection of q onto coordinates different than x, e.g. $q_{(1)}(\cdot,\cdot) \overset{a.s}{=} \int q(\mathrm{d}x_1, \cdot, \cdot)$. This implies that the weights of q_x should be almost surely the same for every x. Instead, if the missing mechanism is not completely at random, q_x can not be described as the projection of a unique q. Indeed the missing mechanism may be informative, leading to sample-specific features. Therefore, the choice of an additive n-FuRBIs allows q_x to have sample-specific components when needed.

As for the baseline distribution G_0 on $\tilde{\boldsymbol{\mu}}$, suppose that an hyper-tie is sampled between an observation $(w_{2,i}, w_{3,i})$ from sample "(1)" and one observation $(w_{1,i}, w_{3,i})$ from sample "(2)", thus assigning the two observations to the same cluster. G_0 is then used to sample the corresponding locations: (X_2^*, X_3^*) and (Y_1^*, Y_3^*). Since we want to interpret the hyper-tie between incomplete observations as a tie between complete observations, we must have $X_3^* = Y_3^*$, while X_2^* and Y_1^* are sampled jointly with a certain correlation $\rho_{1,2}$ and depending on X_3^* through correlations $\rho_{1,3}$ and $\rho_{2,3}$. Therefore, since coordinates corresponding to the same original variable should be assigned the same value, G_0 is actually degenerate on a $P = 3$ dimensional space. In the simulation and real data application G_0 is a 3-variate normal, whose correlation matrix ρ_0 depends on correlation parameters ρ_{12}, ρ_{23}, ρ_{13} on which a truncated uniform hyperprior is used, where the truncation ensures that the matrix is almost-surely positive-definite. Since the data are centered, the mean of G_0 is instead fixed equal to a vector of all 0. Moreover, an independent Gamma$(3,3)$ prior is assigned to the three variances $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$. Finally, the concentration parameter θ is set equal to 0.1 in order to favor sparsity, i.e., a lower number of clusters.

Simulating scenarios: missing data distribution



(a) MCAR 16.1% missing entries

(b) MNAR 17.7% missing entries

(c) MCAR 35.9% missing entries

(d) MNAR 34% complete observations

Figure 18: Percentages of missing entries of each variable-cluster pair.

A3 Proofs of Section 3.3

Preliminary results on Pitman-Yor and σ-stable processes

Let $\mathbb{X}$ the sampling space and denote with $\mathbb{M}_{\mathbb{X}}$ the set of boundedly finite measures on $(\mathbb{X}, \mathcal{X})$; we refer to Daley and Vere-Jones (2007) for technical details. If Q_0 is a probability measure on $\mathbb{X}$ and $\sigma \in (0, 1)$, we define a random variable μ_σ that takes value in $\mathbb{M}_{\mathbb{X}}$ as

$$E\left[e^{-u\mu_\sigma(A)}\right] = e^{-Q_0(A)u^\sigma}, \quad u > 0. \tag{28}$$

We say that μ_σ is endowed with the law of a σ-stable process, denoted with $\mathbb{P}_\sigma$, which is an example of Completely Random Measure (Kingman, 1967). The latter can be normalized under suitable conditions Regazzini et al. (2003); James et al. (2009) to obtain random probability measures.

Consider the following model

$$X_i \mid P \overset{\text{i.i.d.}}{\sim} P, \quad P \sim \text{PY}(\sigma, \theta, Q_0), \tag{29}$$

with $\sigma \in [0, 1)$, $\theta > 0$ and Q_0 arbitrary probability measure on $\mathbb{X}$. Let $\mathbb{P}_{\sigma,\theta}$ be absolutely continuous with respect to $\mathbb{P}_\sigma$, with Radon-Nikodym derivative

$$\frac{\mathbb{P}_{\sigma,\theta}}{\mathbb{P}_\sigma}(m) = \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)} m^{-\theta}(\mathbb{X}).$$

The resulting random measure $\mu_{\sigma,\theta}$ can be shown (Pitman and Yor, 1997) to be such that

$$P(\cdot) \overset{\mathrm{d}}{=} \frac{\mu_{\sigma,\theta}(\cdot)}{\mu_{\sigma,\theta}(\mathbb{X})}, \tag{30}$$

with P as in (29). Therefore, the PY process can be represented through a σ-stable process, with a suitable change of measure. For ease of notation, as the in main document, we denote

$$\gamma = \frac{1-\sigma}{\theta+1}. \tag{31}$$

We start with a well-known result, to show the mathematical techniques employed throught the paper. The computations in this proof follow the ones in Section 2 of James et al. (2006).

Lemma 33. *Consider model* (29)*. Then it holds* $E[P(A)] = Q_0(A)$ *for every* $A \in \mathcal{X}$*.*

Proof. By representation (30) we write

$$\mathbb{E}[P(A)] = \mathbb{E}\left[\frac{\mu_{\sigma,\theta}(A)}{\mu_{\sigma,\theta}(\mathbb{X})}\right] = \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)}\mathbb{E}\left[\frac{\mu_\sigma(A)}{\mu_\sigma^{1+\theta}(\mathbb{X})}\right],$$

where μ_σ is a σ-stable CRM. Notice that

$$\frac{1}{\mu_\sigma^{1+\theta}(\mathbb{X})} = \int_0^\infty \frac{u^\theta}{\Gamma(\theta+1)} e^{-u\mu_\sigma(\mathbb{X})}\,\mathrm{d}u,$$

so that by Fubini Theorem we get

$$\begin{aligned}\mathbb{E}[P(A)] &= \frac{\sigma}{\theta\Gamma(\theta/\sigma)}\int_0^\infty u^\theta \mathbb{E}\left[\mu_\sigma(A)e^{-u\mu_\sigma(\mathbb{X})}\right]\,\mathrm{d}u \\ &= \frac{\sigma}{\theta\Gamma(\theta/\sigma)}\int_0^\infty u^\theta \mathbb{E}\left[\mu_\sigma(A)e^{-u\mu_\sigma(A)}\right]\mathbb{E}\left[e^{-u\mu_\sigma(A^c)}\right]\,\mathrm{d}u,\end{aligned}$$

by independence over evaluation on disjoint sets (which holds since μ_σ is a completelt random measure). By (28) we have $\mathbb{E}\left[e^{-u\mu_\sigma(A^c)}\right] = e^{-Q_0(A^c)u^\sigma}$ and

$$\mathbb{E}\left[\mu_\sigma(A)e^{-u\mu_\sigma(A)}\right] = -\frac{\mathrm{d}}{\mathrm{d}u}\mathbb{E}\left[e^{-u\mu_\sigma(A)}\right] = Q_0(A)\sigma u^{\sigma-1}e^{-Q_0(A)u^\sigma},$$

which implies

$$\begin{aligned}\mathbb{E}[P(A)] &= Q_0(A)\frac{\sigma^2}{\theta\Gamma(\theta/\sigma)}\int_0^\infty u^{\theta+\sigma-1}e^{-u^\sigma}\,\mathrm{d}u \\ &= Q_0(A)\frac{\sigma\Gamma(\theta/\sigma+1)}{\theta\Gamma(\theta/\sigma)}\int_0^\infty \frac{\sigma}{\Gamma(\theta/\sigma+1}u^{\theta+\sigma-1}e^{-u^\sigma}\,\mathrm{d}u \\ &= Q_0(A),\end{aligned}$$

as desired. □

Now we give three preliminary lemmas.

Lemma 34. *Consider model* (29)*. If $A \in \mathcal{X}$ then it holds*

$$\mathbb{E}\left[P^2(A)\right] = (1-\gamma)Q_0^2(A) + \gamma Q_0(A)$$

Proof. By representation (30), proceeding as in the proof of Lemma 33 we get

$$\begin{aligned}\mathbb{E}\left[P^2(A)\right] &= \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)}\mathbb{E}\left[\frac{\mu_\sigma^2(A)}{\mu_\sigma^{2+\theta}(\mathbb{X})}\right] = \frac{\sigma\Gamma(\theta)}{\Gamma(\theta+2)\Gamma(\theta/\sigma)}\int_0^\infty u^{\theta+1}\mathbb{E}\left[\mu_\sigma^2(A)e^{-u\mu_\sigma(\mathbb{X})}\right]\mathrm{d}u \\ &= \frac{\sigma\Gamma(\theta)}{\Gamma(\theta+2)\Gamma(\theta/\sigma)}\int_0^\infty u^{\theta+1}\mathbb{E}\left[\mu_\sigma^2(A)e^{-u\mu_\sigma(A)}\right]\mathbb{E}\left[e^{-u\mu_\sigma(A^c)}\right]\mathrm{d}u.\end{aligned}$$

By applying again (28) we have

$$\begin{aligned}\mathbb{E}\left[\mu_\sigma^2(A)e^{-u\mu_\sigma(A)}\right] &= \frac{\mathrm{d}^2}{\mathrm{d}u^2}\mathbb{E}\left[e^{-u\mu_\sigma(A)}\right] = -\sigma Q_0(A)\frac{\mathrm{d}}{\mathrm{d}u}\left\{u^{\sigma-1}e^{-Q_0(A)u^\sigma}\right\} \\ &= \left[\sigma^2 Q_0^2(A)u^{2\sigma-2} - \sigma(\sigma-1)Q_0(A)u^{\sigma-2}\right]e^{-Q_0(A)u^\sigma},\end{aligned}$$

which implies

$$\begin{aligned}\mathbb{E}\left[P^2(A)\right] &= \frac{\sigma\Gamma(\theta)}{\Gamma(\theta+2)\Gamma(\theta/\sigma)}\int_0^\infty \left[\sigma^2 Q_0^2(A)u^{2\sigma+\theta-1} - \sigma(\sigma-1)Q_0(A)u^{\sigma+\theta-1}\right]e^{-u^\sigma}\mathrm{d}u \\ &= Q_0^2(A)\frac{\theta+\sigma}{\theta+1} + Q_0(A)\frac{1-\sigma}{\theta+1} = (1-\gamma)Q_0^2(A) + \gamma Q_0(A),\end{aligned}$$

as desired. □

Lemma 35. *Consider model* (29)*. If $A, B \in \mathcal{X}$ are disjoint, then it holds*

$$\mathbb{E}\left[P(A)P(B)\right] = (1-\gamma)Q_0(A)Q_0(B).$$

Proof. By representation (30), proceeding as in the proof of Lemma 33 we get

$$\begin{aligned}&\mathbb{E}\left[P(A)P(B)\right] = \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)}\mathbb{E}\left[\frac{\mu_\sigma(A)\mu_\sigma(B)}{\mu_\sigma^{2+\theta}(\mathbb{X})}\right] = \frac{\sigma\Gamma(\theta)}{\Gamma(\theta+2)\Gamma(\theta/\sigma)}\int_0^\infty u^{\theta+1}\mathbb{E}\left[\mu_\sigma(A)\mu_\sigma(B)e^{-u\mu_\sigma(\mathbb{X})}\right]\mathrm{d}u \\ &= \frac{\sigma\Gamma(\theta)}{\Gamma(\theta+2)\Gamma(\theta/\sigma)}\int_0^\infty u^{\theta+1}\mathbb{E}\left[\mu_\sigma(A)e^{-u\mu_\sigma(A)}\right]\mathbb{E}\left[\mu_\sigma(B)e^{-u\mu_\sigma(B)}\right]\mathbb{E}\left[e^{-u\mu_\sigma((A\cup B)^c)}\right]\mathrm{d}u \\ &= Q_0A)Q_0(B)\frac{\sigma^3\Gamma(\theta)}{\Gamma(\theta+2)\Gamma(\theta/\sigma)}\int_0^\infty u^{\theta+2\sigma-1}e^{-u^\sigma}\mathrm{d}u \\ &= (1-\gamma)Q_0(A)Q_0(B),\end{aligned}$$

as desired. □

Lemma 36. *Consider model* (29)*. Then*

$$\mathbb{E}\left[P(A)P(B)\right] = \gamma Q_0(A\cap B) + (1-\gamma)Q_0(A)Q_0(B),$$

for every $A, B \in \mathcal{X}$.

Proof. By definition we have

$$\begin{aligned}\mathbb{E}\left[P(A)P(B)\right] &= \mathbb{E}\Big[\left(P(A\cap B)+P(A\backslash B)\right)\left(P(A\cap B)+P(B\backslash A)\right)\Big]\\ &= \mathbb{E}\Big[P^2(A\cap B)\Big]+\mathbb{E}\left[P(A\cap B)P(B\backslash A)\right]\\ &+\mathbb{E}\left[P(A\cap B)P(A\backslash B)\right]+\mathbb{E}\left[P(A\backslash B)P(B\backslash A)\right].\end{aligned}$$

Applying Lemmas 34 and 35 we get

$$\begin{aligned}\mathbb{E}\left[P(A)P(B)\right] &= \gamma Q_0(A\cap B)+(1-\gamma)Q_0(A\cap B\left[Q_0(A\cap B)+Q_0(B\backslash A)\right]\\ &+(1-\gamma)Q_0(A\backslash B)\left[Q_0(A\cap B)+Q_0(B\backslash A)\right]\\ &= \gamma Q_0(A\cap B)+(1-\gamma)Q_0(A)Q_0(B),\end{aligned}$$

as desired. □

Proof of Propositions 19 and 20 and Corollary 7

We need three preliminary lemmas.

Lemma 37. *Let $\mathcal{T}$ be a tree. For every $\boldsymbol{p}\in\mathcal{T}$ and $\boldsymbol{q}\in\mathcal{P}(\boldsymbol{p})$ it holds*

$$\mathbb{E}\left[P_{\boldsymbol{p}}(A)\mid P_{\boldsymbol{q}}\right]=P_{\boldsymbol{q}}(A),$$

for every $A\in\mathcal{X}$.

Proof. For $\mathbf{q}=\mathbf{p}$ the result holds by construction, while for $\mathbf{q}=\underline{p}$ it holds by Lemma 33. Thus we prove the result by induction on the elements of $\mathcal{P}$, assuming

$$\mathbb{E}\left[P_{\mathbf{p}}(A)\mid P_{\mathbf{g}}\right]=P_{\mathbf{g}}(A),$$

with $\mathbf{g}\in\mathcal{P}(\mathbf{p})$ and $|\mathbf{g}|\leq|\mathbf{p}|$. By the Double Expectation Theorem we have

$$\mathbb{E}\Big[P_{\mathbf{p}}(A)\mid P_{\underline{\mathbf{g}}}\Big]=\mathbb{E}\Big[E\left[P_{\mathbf{p}}(A)\mid P_{\mathbf{g}}\right]\mid P_{\underline{\mathbf{g}}}\Big]=\mathbb{E}\Big[P_{\mathbf{g}}(A)\mid P_{\underline{\mathbf{g}}}\Big]=P_{\underline{\mathbf{g}}}(A),$$

by Lemma 33. □

Lemma 38. *Let $\mathcal{T}$ be a tree. Let $\boldsymbol{p},\boldsymbol{q}\in\mathcal{T}$ with MRCA $\boldsymbol{m}$. Then it holds:*

$$\mathbb{E}\left[P_{\boldsymbol{p}}(A)P_{\boldsymbol{q}}(B)\right]=\mathbb{E}\left[P_{\boldsymbol{m}}(A)P_{\boldsymbol{m}}(B)\right],$$

for every $A, B\in\mathcal{X}$.

Proof. By the Double Expectation Theorem and Lemma 37 we get immediately

$$\begin{aligned}\mathbb{E}\left[P_{\mathbf{p}}(A)P_{\mathbf{q}}(B)\right] &= \mathbb{E}\Big[E\left[P_{\mathbf{p}}(A)P_{\mathbf{q}}(B)\mid P_{\mathbf{m}}\right]\Big]\\ &= \mathbb{E}\Big[E\left[P_{\mathbf{q}}(B)\mid P_{\mathbf{m}}\right]\mathbb{E}\left[P_{\mathbf{p}}(B)\mid P_{\mathbf{m}}\right]\Big]=\mathbb{E}\left[P_{\mathbf{m}}(A)P_{\mathbf{m}}(B)\right],\end{aligned}$$

since $\mathbf{m}\in\mathcal{P}(\mathbf{p})\cap\mathcal{P}(\mathbf{q})$. □

Lemma 39. *Let $\mathcal{T}$ be a tree and $\boldsymbol{p} \in \mathcal{T}$. Then it holds:*

$$\mathbb{E}\left[P_{\boldsymbol{p}}(A)P_{\boldsymbol{p}}(B)\right] = \prod_{\boldsymbol{l}\in\mathcal{P}(\boldsymbol{p})}(1-\gamma_{\boldsymbol{l}})Q_0(A)Q_0(B) + \left(1 - \prod_{\boldsymbol{l}\in\mathcal{P}(\boldsymbol{p})}(1-\gamma_{\boldsymbol{l}})\right)Q_0(A\cap B),$$

for every $A, B \in \mathcal{X}$.

Proof. By the Double Expectation Theorem we have

$$\mathbb{E}\left[P_{\mathbf{p}}(A)P_{\mathbf{p}}(B)\right] = \mathbb{E}\left[\mathbb{E}\left[P_{\mathbf{p}}(A)P_{\mathbf{p}}(B) \mid P_{\underline{\mathbf{p}}}\right]\right].$$

Since $P_{\mathbf{p}} \mid P_{\underline{\mathbf{p}}} \sim \mathrm{PY}\left(\sigma_{\mathbf{p}}, \theta_{\mathbf{p}}, P_{\underline{\mathbf{p}}}\right)$, we get

$$\begin{aligned}\mathbb{E}\left[P_{\mathbf{p}}(A)P_{\mathbf{p}}(B)\right] &= \gamma_{\mathbf{p}}\mathbb{E}\left[P_{\underline{\mathbf{p}}}(A\cap B)\right] + (1-\gamma_{\mathbf{p}})\mathbb{E}\left[P_{\underline{\mathbf{p}}}(A)P_{\underline{\mathbf{p}}}(B)\right]\\ &= \gamma_{\mathbf{p}}Q_0(A\cap B) + (1-\gamma_{\mathbf{p}})\mathbb{E}\left[P_{\underline{\mathbf{p}}}(A)P_{\underline{\mathbf{p}}}(B)\right]\end{aligned}$$

by Lemma 36 and the first point of Proposition 19. Thus we need to solve the recursion

$$\begin{cases}R_{\mathbf{p}} = \gamma_{\mathbf{p}}Q_0(A\cap B) + (1-\gamma_{\mathbf{p}})R_{\underline{\mathbf{p}}},\\ R_{\underline{0}} = Q_0(A)Q_0(B)\end{cases}$$

whose solution is given exactly by

$$R_{\mathbf{p}} = \prod_{\mathbf{l}\in\mathcal{P}(\mathbf{p})}(1-\gamma_{\mathbf{l}})Q_0(A)Q_0(B) + \left(1 - \prod_{\mathbf{l}\in\mathcal{P}(\mathbf{p})}(1-\gamma_{\mathbf{l}})\right)Q_0(A\cap B).$$

□

Proof of Proposition 19. As regards the first point, by the Double Expectation Theorem we have

$$\mathbb{E}\left[P_{\mathbf{p}}(A)\right] = \mathbb{E}\left[E\left[P_{\mathbf{p}}(A) \mid P_0\right]\right] = \mathbb{E}\left[P_0(A)\right] = Q_0(A),$$

by Lemma 37. As regards the second point, through Lemma 39, we obtain

$$\mathbb{E}\left[P_{\mathbf{p}}(A)P_{\mathbf{p}}(A)\right] = \prod_{\mathbf{l}\in\mathcal{P}(\mathbf{p})}(1-\gamma_{\mathbf{l}})Q_0(A)Q_0(B) + \left(1 - \prod_{\mathbf{l}\in\mathcal{P}(\mathbf{p})}(1-\gamma_{\mathbf{l}})\right)Q_0(A\cap B),$$

$$\mathbb{E}\left[P_{\mathbf{q}}(A)P_{\mathbf{q}}(A)\right] = \prod_{\mathbf{l}\in\mathcal{P}(\mathbf{q})}(1-\gamma_{\mathbf{l}})Q_0(A)Q_0(B) + \left(1 - \prod_{\mathbf{l}\in\mathcal{P}(\mathbf{q})}(1-\gamma_{\mathbf{l}})\right)Q_0(A\cap B).$$

Instead, by Lemma 38 we have

$$\begin{aligned}\mathbb{E}\left[P_{\mathbf{p}}(A)P_{\mathbf{q}}(A)\right] &= \mathbb{E}\left[P_{\mathbf{m}}(A)P_{\mathbf{m}}(A)\right]\\ &= \prod_{\mathbf{l}\in\mathcal{P}(\mathbf{m})}(1-\gamma_{\mathbf{l}})Q_0(A)Q_0(B) + \left(1-\prod_{\mathbf{l}\in\mathcal{P}(\mathbf{m})}(1-\gamma_{\mathbf{l}})\right)Q_0(A\cap B).\end{aligned}$$

Then it holds

$$\begin{aligned}\mathrm{Cov}\left(P_{\mathbf{p}}(A),P_{\mathbf{q}}(A)\right) &= \mathbb{E}\left[P_{\mathbf{p}}(A)P_{\mathbf{q}}(A)\right]-\mathbb{E}\left[P_{\mathbf{p}}(A)\right]\mathbb{E}\left[P_{\mathbf{q}}(A)\right]\\ &= \left(1-\prod_{\mathbf{l}\in\mathcal{P}(\mathbf{m})}(1-\gamma_{\mathbf{l}})\right)\left[Q_0(A)-Q_0(A)^2\right],\end{aligned}$$

and

$$\mathrm{Var}\left(P_{\mathbf{p}}(A)\right) = \mathbb{E}\left[P_{\mathbf{p}}(A)P_{\mathbf{p}}(A)\right]-\mathbb{E}^2\left[P_{\mathbf{p}}(A)\right] = \left(1-\prod_{\mathbf{l}\in\mathcal{P}(\mathbf{p})}(1-\gamma_{\mathbf{l}})\right)\left[Q_0(A)-Q_0(A)^2\right],$$

$$\mathrm{Var}\left(P_{\mathbf{q}}(A)\right) = \mathbb{E}\left[P_{\mathbf{q}}(A)P_{\mathbf{q}}(A)\right]-E^2\left[P_{\mathbf{q}}(A)\right] = \left(1-\prod_{\mathbf{l}\in\mathcal{P}(\mathbf{q})}(1-\gamma_{\mathbf{l}})\right)\left[Q_0(A)-Q_0(A)^2\right],$$

from which the result follows. □

Proof of Proposition 20 and Corollary 7. The first point of Proposition 20 follows immediately by

$$\mathbb{P}(X_{\mathbf{p},i}\in A) = \mathbb{E}\left[P_{\mathbf{p}}(A)\right] = Q_0(A),$$

by Lemma 37. Similarly, Corollary 1 follows by noticing

$$\begin{aligned}\mathbb{P}\left(X_{\mathbf{p},i}\in A, X_{\mathbf{q},j}\in B\right) &= \mathbb{E}\left[P_{\mathbf{p}}(A)P_{\mathbf{q}}(B)\right] = E\left[P_{\mathbf{m}}(A)P_{\mathbf{m}}(B)\right]\\ &= \prod_{\mathbf{l}\in\mathcal{P}(\mathbf{m})}(1-\gamma_{\mathbf{l}})Q_0(A)Q_0(B) + \left(1-\prod_{\mathbf{l}\in\mathcal{P}(\mathbf{m})}(1-\gamma_{\mathbf{l}})\right)Q_0(A\cap B),\end{aligned}$$

by Lemmas 38 and 39. Thus the joint distribution of the vector $\left(X_{\mathbf{p},i},X_{\mathbf{q},j}\right)$ is given by

$$\mu(\mathrm{d}x_{\mathbf{p}},\mathrm{d}y_{\mathbf{q}}) = \prod_{\mathbf{l}\in\mathcal{P}(\mathbf{m})}(1-\gamma_{\mathbf{l}})Q_0(\mathrm{d}x_{\mathbf{p}})Q_0(\mathrm{d}x_{\mathbf{q}}) + \left(1-\prod_{\mathbf{l}\in\mathcal{P}(\mathbf{m})}(1-\gamma_{\mathbf{l}})\right)Q_0(\mathrm{d}x_{\mathbf{p}})\delta_{x_{\mathbf{q}}}(x_{\mathbf{p}}),$$

from which the second point of Proposition 20 immediately follows. □

Proof of Theorem 15

We start by reporting Lemma 1 in the supplementary material of Camerlenghi et al. (2019b) for the case of the σ-stable process, that will be useful in the following.

Lemma 40. *Let $\sigma \in (0,1)$. Define $\tau_q(u) = \frac{\sigma\Gamma(q-\sigma)}{\Gamma(1-\sigma)} u^{\sigma-q}$ and*

$$\xi_{n,i} = \sum_{\boldsymbol{q}} \frac{1}{i!} \binom{n}{q_1, \dots, q_i} \tau_{q_1}(u) \dots \tau_{q_i}(u), \quad (A1)$$

where the sum runs over all vectors $\boldsymbol{q} = (q_1, \dots, q_i)$ of positive integers such that $\sum_{j=1}^{i} q_j = n$. Then the following relation holds

$$(-1)^n \frac{\mathrm{d}^n}{\mathrm{d}u^n} e^{-cu^\sigma} = e^{-cu^\sigma} \sum_{i=1}^{n} c^i \xi_{n,i}, \quad (A2)$$

for every $c > 0$.

Then we prove a preliminary lemma.

Lemma 41. *Let $P \sim PY(\sigma, \theta, Q)$, with Q probability measure on $\mathbb{X}$. Consider a collection of disjoint sets $A_1, \dots, A_k$ and a vector $(n_1, \dots, n_k)$ of positive integers such that $\sum_{j=1}^{k} n_j = n$. Then we have*

$$\mathbb{E}\left[\prod_{j=1}^{k} P^{n_j}(A_j)\right] = \sum_{\boldsymbol{l}} \sum_{\boldsymbol{q}} \left[\prod_{j=1}^{k} Q(A_j)^{l_j}\right] \left[\prod_{j=1}^{k} \frac{1}{l_j!} \binom{n_j}{q_{j,1}, \dots, q_{j,l_j}}\right] \Phi_{l_\bullet}^{(n)}(\boldsymbol{q}_1, \dots, \boldsymbol{q}_k),$$

where

1. *$\sum_{\boldsymbol{l}} = \sum_{l_1=1}^{n_1} \cdots \sum_{l_k=1}^{n_k}$ and $l_\bullet = \sum_{j=1}^{k} l_j$;*
2. *$\sum_{\boldsymbol{q}} = \sum_{\boldsymbol{q}_1} \cdots \sum_{\boldsymbol{q}_k}$, where $\boldsymbol{q}_j = (\boldsymbol{q}_1, \dots, \boldsymbol{q}_k)$ is a vector of positive integers such that $\sum_{t=1}^{l_j} q_{j,t} = n_j$ and $\sum_{\boldsymbol{q}_j}$ is as in (A1).*

Proof. Let μ_σ be a σ-stable process with parameter σ. With the same reasoning of the proof of Lemma 33 we have

$$\begin{aligned}\mathbb{E}\left[\prod_{j=1}^{k} P^{n_j}(A_j)\right] &= \mathbb{E}\left[\frac{\prod_{j=1}^{k} \mu_\sigma^{n_j}(A_j)}{\mu_\sigma^{\theta+n}(\mathbb{X})}\right] \\ &= \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)\Gamma(\theta+n)} \int_0^\infty u^{\theta+n-1} \mathbb{E}\left[e^{-u\mu_\sigma(\mathbb{X}_-)}\right] \prod_{j=1}^{k} \mathbb{E}\left[e^{-u\mu_\sigma(A_j)} \mu_\sigma^{n_j}(A_j)\right] \mathrm{d}u,\end{aligned}$$

where $\mathbb{X}_- = (A_1 \cup \cdots \cup A_k)^c$. Moreover, by Lemma 40 we have

$$\begin{aligned}E\left[e^{-u\mu_\sigma(A_j)} \mu_\sigma^{n_j}(A_j)\right] &= \frac{\mathrm{d}u^{n_j}}{\mathrm{d}u^{n_j}} E\left[(-1)^{n_j} e^{-u\mu_\sigma(A_j)}\right] = (-1)^{n_j} \frac{\mathrm{d}u^{n_j}}{\mathrm{d}u^{n_j}} e^{-Q(A_j)u^\sigma} \\ &= e^{-Q(A_j)u^\sigma} \sum_{l_j=1}^{n_j} Q^{l_j}(A_j) \xi_{n_j,l_j}(u).\end{aligned}$$

Therefore, by definition (A1) of $\xi_{n,i}$

$$\begin{aligned}\mathbb{E}\left[\prod_{j=1}^{k} p^{n_j}(A_j)\right] &= \frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)\Gamma(\theta+n)}\int_0^\infty u^{n+\theta-1}e^{-u^\sigma}\prod_{j=1}^{k}\sum_{l_j=1}^{n_j} Q^{l_j}(A_j)\xi_{n_j,l_j}(u)\,\mathrm{d}u\\
&= \sum_{\mathbf{l}}\left[\prod_{j=1}^{k} Q(A_j)^{l_j}\right]\frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)\Gamma(\theta+n)}\int_0^\infty u^{n+\theta-1}e^{-u^\sigma}\prod_{j=1}^{k}\xi_{n_j,l_j}(u)\,\mathrm{d}u\\
&= \sum_{\mathbf{l}}\sum_{\mathbf{q}}\left[\prod_{j=1}^{k} Q(A_j)^{l_j}\right]\left[\prod_{j=1}^{k}\frac{1}{l_j!}\binom{n_j}{q_{j,1},\dots,q_{j,l_j}}\right]\frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)\Gamma(\theta+n)}\times\\
&\qquad\times\int_0^\infty u^{\theta+n-1}e^{-u^\sigma}\prod_{j=1}^{k}\prod_{t=1}^{l_j}\tau_{q_{j,t}}(u)\,\mathrm{d}u.\end{aligned}$$

By definition of $\tau_q(u)$ we have

$$\prod_{j=1}^{k}\prod_{t=1}^{l_j}\tau_{q_{j,t}}(u) = \frac{\sigma^{l_\bullet}}{\Gamma^{l_\bullet}(1-\sigma)}\prod_{j=1}^{k}\prod_{t=1}^{l_j}\Gamma(q_{j,t}-\sigma)u^{\sigma l_\bullet - n}$$

and $\int_0^\infty u^{\theta+\sigma l_\bullet-1}e^{-u^\sigma}\,\mathrm{d}u = \frac{\Gamma(l_\bullet+\theta/\sigma)}{\sigma}$, which implies

$$\begin{aligned}\frac{\sigma\Gamma(\theta)}{\Gamma(\theta/\sigma)\Gamma(\theta+n)}\int_0^\infty u^{\theta+n-1}e^{-u^\sigma}\prod_{j=1}^{k}\prod_{t=1}^{l_j}\tau_{q_{j,t}}(u)\,\mathrm{d}u &= \frac{\sigma^{l_\bullet}\Gamma(l_\bullet+\theta/\sigma)\Gamma(\theta)}{\Gamma(\theta/\sigma)\Gamma(\theta+n)}\prod_{j=1}^{k}\prod_{t=1}^{l_j}\frac{\Gamma(q_{j,t}-\sigma)}{\Gamma(1-\sigma)}\\
&= \frac{1}{(\theta+1)_{(n-1)}}\frac{\sigma^{l_\bullet}\Gamma(l_\bullet+\theta/\sigma)}{\theta\Gamma(\theta/\sigma)}\prod_{j=1}^{k}\prod_{t=1}^{l_j}(1-\sigma)_{q_{j,t}-1}\\
&= \frac{\prod_{i=1}^{l_\bullet-1}(\theta+i\sigma)}{(\theta+1)_{(n-1)}}\prod_{j=1}^{k}\prod_{t=1}^{l_j}(1-\sigma)_{q_{j,t}-1} = \Phi_{l_\bullet}^{(n)}(\mathbf{q}_1,\dots,\mathbf{q}_k),\end{aligned}$$

as desired. □

Proof of Theorem 15. Denoting $\mathbf{n_p} = (n_{\mathbf{p},1},\dots,n_{\mathbf{p},k})$, for any $x_1 \neq \dots \neq x_k$ we evaluate

$$\begin{aligned}M(\mathrm{d}x_1,\dots,\mathrm{d}x_k) &= \mathbb{E}\left[\prod_{i=1}^{d}\prod_{\mathbf{p}\in L_i}\prod_{j=1}^{k} P_{\mathbf{p}}^{n_{\mathbf{p},j}}(\mathrm{d}x_j)\right]\\
&= \lim_{\epsilon\to 0}\mathbb{E}\left[\prod_{i=1}^{d}\prod_{\mathbf{p}\in L_i}\prod_{j=1}^{k} P_{\mathbf{p}}^{n_{\mathbf{p},j}}(A_j)\right] = \lim_{\epsilon\to 0} M(A_1,\dots,A_k),\end{aligned}$$

where $A_j = A_{j,\epsilon} = B(x_j,\epsilon)$ is a ball of radius ϵ around x_j, with $\epsilon > 0$ small enough so that

$A_i \cap A_j = \emptyset$, for any $i \neq j$. By basic properties of conditional expectation, we get

$$M(A_1, \dots, A_k) = \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^{d} \prod_{\mathbf{p}\in L_i} \prod_{j=1}^{k} P_{\mathbf{p}}^{n_{\mathbf{p},j}}(A_j) \mid P_{\mathbf{g}}\,;\, \mathbf{g} \in L_{d-1}\right]\right]$$
$$= \mathbb{E}\left[\prod_{i=1}^{d-1} \prod_{\mathbf{p}\in L_i} \prod_{j=1}^{k} P_{\mathbf{p}}^{n_{\mathbf{p},j}}(A_j) \mathbb{E}\left[\prod_{\mathbf{p}\in L_d} \prod_{j=1}^{k} P_{\mathbf{p}}^{n_{\mathbf{p},j}}(A_j) \mid P_{\mathbf{g}}\,;\, \mathbf{g} \in L_{d-1}\right]\right].$$

Conditional to $\{P_{\mathbf{g}}\,;\, \mathbf{g} \in L_{d-1}\}$, the random measures $P_{\mathbf{p}}$, with $\mathbf{p} \in L_d$, are independent, so that

$$\mathbb{E}\left[\prod_{\mathbf{p}\in L_d} \prod_{j=1}^{k} P_{\mathbf{p}}^{n_{\mathbf{p},j}}(A_j) \mid P_{\mathbf{g}}\,;\, \mathbf{g} \in L_{d-1}\right] = \prod_{\mathbf{p}\in L_d} \mathbb{E}\left[\prod_{j=1}^{k} P_{\mathbf{p}}^{n_{\mathbf{p},j}}(A_j) \mid P_{\mathbf{g}}\,;\, \mathbf{g} \in L_{d-1}\right].$$

By Lemma 41, we have

$$\mathbb{E}\left[\prod_{j=1}^{k} P_{\mathbf{p}}^{n_{\mathbf{p},j}}(A_j) \mid P_{\mathbf{g}}\,;\, \mathbf{g} \in L_{d-1}\right] = \sum_{\mathbf{l}_{\mathbf{p}}} \sum_{\mathbf{q}_{\mathbf{p}}} \prod_{j=1}^{k} P_{\underline{\mathbf{p}}}^{l_{\mathbf{p},j}}(A_j) \times$$
$$\times \prod_{j=1}^{k} \frac{1}{l_{\mathbf{p},j}!} \binom{n_{\mathbf{p}}}{q_{\mathbf{p},j,1}, \dots, q_{\mathbf{p},j,l_{\mathbf{p},j}}} \Phi_{l_{\mathbf{p}\bullet},\mathbf{p}}^{(n_{\mathbf{p}})}(\mathbf{q}_{\mathbf{p},1}, \dots, \mathbf{q}_{\mathbf{p},k}).$$

By definition of $l_{\mathbf{p}+1}$ and readjusting the terms, thanks to the linearity of the expected value we are left with computing

$$\mathbb{E}\left[\prod_{i=1}^{d-2} \prod_{\mathbf{p}\in L_i} \prod_{j=1}^{k} P_{\mathbf{p}}^{n_{\mathbf{p},j}}(A_j) \prod_{\mathbf{p}\in L_{d-1}} \prod_{j=1}^{k} P_{\mathbf{p}}^{n_{\mathbf{p},j}+l_{\mathbf{p}+1,j}}(A_j)\right] =$$
$$\mathbb{E}\left[\prod_{i=1}^{d-2} \prod_{\mathbf{p}\in L_i} \prod_{j=1}^{k} P_{\mathbf{p}}^{n_{\mathbf{p},j}}(A_j) \mathbb{E}\left[\prod_{\mathbf{p}\in L_{d-1}} \prod_{j=1}^{k} P_{\mathbf{p}-1}^{n_{\mathbf{p},j}+l_{\mathbf{p}+1,j}}(A_j) \mid P_{\mathbf{g}}\,;\, \mathbf{g} \in L_{d-2}\right]\right],$$

therefore we apply repeatedly Lemma 41. In the end we obtain

$$M(A_1, \dots, A_k) = \sum_{\mathbf{l}} \sum_{\mathbf{q}} \frac{1}{\mathbf{l}!} \binom{\mathbf{n}}{\mathbf{q}} \mathbb{E}\left[\prod_{j=1}^{k} P_0^{l_{1,j}}(A_j)\right] \prod_{i=1}^{d} \prod_{\mathbf{p}\in L_i} \Phi_{l_{\mathbf{p}\bullet},i}^{(n_{\mathbf{p}}+l_{\mathbf{p}+1\bullet})}(\mathbf{q}_{\mathbf{p},1}, \dots, \mathbf{q}_{\mathbf{p},k}).$$

As $\epsilon \to 0$, the non-atomicity of Q_0 implies

$$\mathbb{E}\left[\prod_{j=1}^{k} P_0^{l_{1,j}}(A_j)\right] = \left(\prod_{j=1}^{k} Q_0(\mathrm{d}x_j)\right) \Phi_{k,0}^{(l_{1\bullet})}(l_{1,1}, \dots, l_{1,k}).$$

Finally, the result follows by noticing

$$\Pi_k^{(n)}(\mathbf{n}_{\mathbf{p}}\,;\, \mathbf{p} \in \mathcal{T}) = \int_{\mathbb{X}^k} M(\mathrm{d}x_1, \dots, \mathrm{d}x_k).$$

□

Proof of Theorem 16

Proof. If $|\mathbf{p}| = 1$, the result follows by Theorems 7 and 8 of Camerlenghi et al. (2019b). We prove the main result by induction on the number of levels. Assume that for every $\mathbf{q}$ with $|\mathbf{q}| = k - 1$ it holds

$$K_{\mathbf{q},n} \approx \left(\prod_{\mathbf{g} \in \mathcal{P}(\mathbf{q})} \lambda_{\sigma_{\mathbf{g}}} \right) (n) = \lambda_{\mathbf{q}}(n),$$

where we use the notation $K_{\mathbf{q},n}$ to emphasize that the observations are collected at $\mathbf{q}$. Assume now that $\mathbf{p}$ has level k. By the same reasoning of Remark 1 or of the proof of Theorem 7 in Camerlenghi et al. (2019b), it holds

$$K_{\mathbf{p},n} \overset{\text{a.s.}}{=} K_{\underline{\mathbf{p}},K'_{\mathbf{p},n}},$$

where $K'_{\mathbf{p},n}$ is the number of distinct values in $T_{\mathbf{p}} = (T_{\mathbf{p},1}, \dots, T_{\mathbf{p},n})$, with $T_{\mathbf{p},i} \mid Q_{\mathbf{p}} \sim Q_{\mathbf{p}}$ and $Q_{\mathbf{p}} \sim \mathrm{PY}(\sigma_{\mathbf{p}}, \theta_{\mathbf{p}}, Q)$, Q being a diffuse measure. Therefore we can write

$$\frac{K_{\mathbf{p},n}}{\lambda_{\underline{\mathbf{p}}}(\lambda_{\sigma_{\mathbf{p}}}(n))} \overset{\text{a.s.}}{=} \frac{K_{\underline{\mathbf{p}},K'_{\mathbf{p},n}}}{K_{\underline{\mathbf{p}},\lambda_{\sigma_{\mathbf{p}}}(n)}} \frac{K_{\underline{\mathbf{p}},\lambda_{\sigma_{\mathbf{p}}}(n)}}{\lambda_{\underline{p}}(\lambda_{\sigma_{\mathbf{p}}}(n))}.$$

The second product on the right hand side converges almost surely to a finite random variable, by induction hypothesis, while

$$\frac{K_{\underline{\mathbf{p}},K'_{\mathbf{p},n}}}{K_{\underline{\mathbf{p}},\lambda_{\sigma_{\mathbf{p}}}(n)}} = \frac{\lambda_{\underline{\mathbf{p}}}\left(K'_{\mathbf{p},n}\right)}{\lambda_{\underline{\mathbf{p}}}\left(\lambda_{\sigma_{\mathbf{p}}}(n)\right)} \frac{K_{\underline{\mathbf{p}},K'_{\mathbf{p},n}}/\lambda_{\underline{\mathbf{p}}}\left(K'_{\mathbf{p},n}\right)}{K_{\underline{\mathbf{p}},\lambda_{\sigma_{\mathbf{p}}}(n)}/\lambda_{\underline{\mathbf{p}}}\left(\lambda_{\sigma_{\mathbf{p}}}(n)\right)}$$

By definition, $K'_{\mathbf{p},n}/\lambda_{\sigma_{\mathbf{p}}}(n)$ converges almost surely to a finite random variable, so that by induction hypothesis the same happens for the ratio on the left hand side. Thus, we conclude

$$K_{\mathbf{p},n} \approx \lambda_{\underline{\mathbf{p}}}(\lambda_{\sigma_{\mathbf{p}}}(n)) = \left(\prod_{\mathbf{q} \in \mathcal{P}(\mathbf{p})} \lambda_{\sigma_{\mathbf{q}}} \right) (n),$$

as desired. □

Proof of Theorem 17

Proof. Let $d = |n/m|$ be the level at which the n-th observation is collected, where $|a|$ is the lowest integer bigger than a. Moreover, let $m_{\mathbf{q}}$ be the number of observations collected at node $\mathbf{q}$. By Theorem 15, conditional on the first $n-1$ observations and the auxiliary variables $\mathbf{T}$, the probability that the n-th observation is completely new is given by

$$\prod_{\mathbf{r} \in \mathcal{P}(\mathbf{p})} \frac{\Phi^{(n_{\mathbf{r}} + l_{\mathbf{r}+1\bullet} + 1)}_{\mathbf{r}, l_{\mathbf{r}\bullet} + 1}(\mathbf{q}_{\mathbf{r},1}, \dots, \mathbf{q}_{\mathbf{r},k}, 1)}{\Phi^{(n_{\mathbf{r}} + l_{\mathbf{r}+1\bullet})}_{\mathbf{r}, l_{\mathbf{r}\bullet}}(\mathbf{q}_{\mathbf{r},1}, \dots, \mathbf{q}_{\mathbf{r},k})} = \prod_{\mathbf{r} \in \mathcal{P}(\mathbf{p})} \frac{\theta_{\mathbf{r}} + \sigma_{\mathbf{r}} l_{\mathbf{r}\bullet}}{\theta_{\mathbf{r}} + m_{\mathbf{r}} + l_{\mathbf{r}+1\bullet}},$$

where $\mathbf{p}$ is the node at level d where the observation is collected. By hypothesis we have

$$\begin{aligned}\frac{\theta_{\mathbf{r}} + \sigma_{\mathbf{r}} l_{\mathbf{r}\bullet}}{\theta_{\mathbf{r}} + m_{\mathbf{r}} + l_{\mathbf{r}+1\bullet}} &\leq \frac{\bar{\theta} + \bar{\sigma}(m_{\mathbf{r}} + l_{\mathbf{r}+1\bullet})}{\bar{\theta} + m_{\mathbf{r}} + l_{\mathbf{r}+1\bullet}} \leq \frac{\bar{\theta} + \bar{\sigma} m(d - |\mathbf{r}| + 1)}{\bar{\theta} + m(d - |\mathbf{r}| + 1)} \\ &\leq \frac{\bar{\theta} + \bar{\sigma} m(d+1)}{\bar{\theta} + m(d+1)}.\end{aligned}$$

Since the last bound does not depend on the data and $\mathbf{T}$ we can write

$$P\left(K_n - K_{n-1} = 1\right) \leq \left(\frac{\bar{\theta} + \bar{\sigma} m(d+1)}{\bar{\theta} + m(d+1)}\right)^{d+1} \approx \bar{\sigma}^{d+1},$$

as $d \to \infty$. Therefore

$$\sum_{n=1}^{\infty} P\left(K_n - K_{n-1} = 1\right) < \infty$$

and the result follows by Borel-Cantelli Lemma. □

References

An, Q., Wang, C., Shterev, I., Wang, E., Carin, L., and Dunson, D. B. (2008). Hierarchical kernel stick-breaking process for multi-task image analysis. In *Proceedings of the 25th international conference on Machine learning*, pages 17–24.

Arbel, J. and Prünster, I. (2017). A moment-matching Ferguson & Klass algorithm. *Statistics and Computing*, 27(1):3–17.

Ascolani, F., Franzolini, B., Lijoi, A., and Prünster, I. (2023a). Full range borrowing of information priors. Technical report, Work in progress.

Ascolani, F., Lijoi, A., and Prünster, I. (2023b). Tree-structured data and Bayesian nonparametric modelling. Technical report, Work in progress.

Barrios, E., Lijoi, A., Nieto-Barajas, L. E., and Prünster, I. (2013). Modeling with normalized random measure mixture models. *Statistical Science*, 28(3):313–334.

Bhardwaj, G. and Dunsby, A. (2013). The business cycle and the correlation between stocks and commodities. *Journal of Investment Consulting*, 14(2):14–25.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Brandsma, H. and Knuver, J. (1989). Effects of school and classroom characteristics on pupil progress in language and arithmetic. *International Journal of Educational Research*, 13(7):777–788.

Brillinger, D. R. (2002). John W. Tukey: his life and professional contributions. *The Annals of Statistics*, 30(6):1535–1575.

Camerlenghi, F., Dunson, D. B., Lijoi, A., Prünster, I., and Rodriguez, A. (2019a). Latent nested nonparametric priors. *Bayesian Analysis*, 14(4):1303–1356.

Camerlenghi, F., Lijoi, A., Orbanz, P., and Prünster, I. (2019b). Distribution theory for hierarchical processes. *The Annals of Statistics*, 47(1):67–92.

Camerlenghi, F., Lijoi, A., and Prünster, I. (2018). Bayesian nonparametric inference beyond the Gibbs-type framework. *Scandinavian Journal of Statistics*, 45(4):1062–1091.

Caron, F., Davy, M., and Doucet, A. (2007). Generalized pólya urn for time-varying dirichlet process mixtures. *Proc. 23rd Conf. on Uncertainty in Artificial Intelligence*, page Vancouver.

Caron, F., Neiswanger, W., Wood, F., Doucet, A., and Davy, M. (2017). Generalized pólya urn for time varying pitman-yor processes. *Journal of Machine Learning Research*, 18(27).

Catalano, M., Lavenant, H., Lijoi, A., and Prünster, I. (2023). A Wasserstein index of dependence for random measures. *Journal of the American Statistical Association*, page under revision.

Catalano, M., Lijoi, A., and Prünster, I. (2021). Measuring dependence in the Wasserstein distance for Bayesian nonparametric models. *The Annals of Statistics*, 49:2916–2947.

Cifarelli, D. M. and Regazzini, E. (1978). Nonparametric statistical problems under partial exchangeability: The role of associative means. *Quaderni Istituto Matematica Finanziaria dell'Università di Torino Serie III*, 12:1–36.

Daley, D. J. and Vere-Jones, D. (2007). *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(2):212–229.

de Finetti, B. (1938). Sur la condition d'equivalence partielle. *Actualités Scientifiques et Industrielles*, 739:5–18, Translated In: Studies in Inductive and Probability, II. Jeffrey, R. (ed.) University of California Press: Berkeley 1980.

Dunson, D. and Park, J. (2008). Kernel stick-breaking processes. *Biometrika*, 95(2):307–323.

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236(5):119–127.

Epifani, I. and Lijoi, A. (2010). Nonparametric priors for vectors of survival functions. *Statistica Sinica*, 20(4):1455–1484.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Favaro, S., Lijoi, A., Nava, C., Nipoti, B., Pruenster, I., and Teh, Y. W. (2016). On the stick-breaking representation for homogeneous NRMIs. *Bayesian Analysis*, 11(3):697–724.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Ferguson, T. S. and Klass, M. J. (1972). A representation of independent increment processes without gaussian components. *The Annals of Mathematical Statistics*, 43(5):1634–1643.

Foti, N. J. and Williamson, S. A. (2013). A survey of non-exchangeable priors for Bayesian nonparametric models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(2):359–371.

Franzolini, B., Cremaschi, A., Boom, W. v. d., and De Iorio, M. (2023). Bayesian clustering of multiple zero-inflated outcomes. *Philosophical Transactions of the Royal Society A, in press.*

Gao, L. L., Bien, J., and Witten, D. (2020). Are clusterings of multiple data views independent? *Biostatistics*, 21(4):692–708.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, Department of Statistics, Stanford University.

Gnedin, A. and Iksanov, A. (2020). On nested infinite occupancy scheme in random environment. *Probability Theory and Related Fields*, 177(3-4):855–890.

Gong, M., Liu, P., Sciurba, F. C., Stojanov, P., Tao, D., Tseng, G. C., Zhang, K., and Batmanghelich, K. (2021). Unpaired data empowers association tests. *Bioinformatics*, 37(6):785–792.

Griffin, J. E., Kolossiatis, M., and Steel, M. F. (2013). Comparing distributions by using dependent normalized random-measure mixtures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):499–529.

Griffin, J. E. and Leisen, F. (2017). Compound random measures and their use in Bayesian non-parametrics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2(79):525–545.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

James, L. F., Lijoi, A., and Prünster, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scandinavian Journal of Statistics*, 33(1):105–120.

James, L. F., Lijoi, A., and Prünster, I. (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1):76–97.

James, L. F., Lijoi, A., and Prünster, I. (2010). On the posterior distribution of classes of random means. *Bernoulli*, 16(1):155–180.

Jivani, A. G. et al. (2011). A comparative study of stemming algorithms. *Int. J. Comp. Tech. Appl*, 2(6):1930–1938.

Johnson, M., Griffiths, T., and Goldwater, S. (2006). Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in neural information processing systems*, 19.

Kingman, J. (1967). Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78.

Kingman, J. (1993). *Poisson Processes.* Clarendon Press, Oxford.

Lee, A. M., Sæther, B.-E., and Engen, S. (2020). Spatial covariation of competing species in a fluctuating environment. *Ecology*, 101(1):e02901.

Lijoi, A., Mena, R. H., and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *Journal of the American Statistical Association*, 100(472):1278–1291.

Lijoi, A. and Nipoti, B. (2014). A class of hazard rate mixtures for combining survival data from different experiments. *Journal of the American Statistical Association*, 109(506):802–814.

Lijoi, A., Nipoti, B., and Prünster, I. (2014a). Bayesian inference with dependent normalized completely random measures. *Bernoulli*, 20(3):1260–1291.

Lijoi, A., Nipoti, B., and Prünster, I. (2014b). Dependent mixture models: clustering and borrowing information. *Computational Statistics & Data Analysis*, 71:417–433.

Lijoi, A. and Prünster, I. (2010). Models beyond the Dirichlet process. In *Bayesian nonparametrics (Hjort, N.L., Holmes, C.C., Müller, P., Walker, S.G. Eds.)*, pages 80–136. Cambridge University Press, Cambridge.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics*, 12(1):351–357.

Lorenz, D. J., Levy, S., and Datta, S. (2018). Inferring marginal association with paired and unpaired clustered data. *Statistical methods in medical research*, 27(6):1806–1817.

MacEachern, S. N. (1999). Dependent nonparametric processes. pages Alexandria, VA: American Statistical Association.

MacEachern, S. N. (2000). Dependent dirichlet processes. *Technical Report,*, page The Ohio State University.

Müller, P., Quintana, F., and Rosner, G. (2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):735–749.

Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). *Bayesian nonparametric data analysis.* Springer.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9(2):249–265.

Nieto-Barajas, L. E. (2021). A class of dependent dirichlet processes via latent multinomial processes. *Statistics*, 55(5):1169–1179.

Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169–186.

Petralia, F., Rao, V., and Dunson, D. B. (2012). Repulsive mixtures. In *Advances in Neural Information Processing Systems - NIPS.*

Pitman, J. (2006). Combinatorial stochastic processes. *Lecture Notes in Math, Springer.*

Pitman, J. and Yor, M. (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *Annals of Probability*, 25:855–900.

Quinlan, J. J., Quintana, F. A., and Page, G. L. (2017). Parsimonious hierarchical modeling using repulsive distributions. *arXiv preprint arXiv:1701.04457*.

Quintana, F., Müller, P., Jara, A., and MacEachern, S. (2022). The dependent Dirichlet process and related models. *Statistical Science*, 37.

Regazzini, E., Lijoi, A., and Prünster, I. (2003). Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics*, 31(2):560–585.

Rigon, T. and Durante, D. (2021). Logit stick-breaking priors for bayesian density regression. *Journal of Statistical Planning and inference*, 211:131–142.

Riva-Palacio, A. and Leisen, F. (2021). Compound vectors of subordinators and their associated positive Lévy copulas. *Journal of Multivariate Analysis*, 183:104728.

Rodriguez, A. and Dunson, D. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, 6(1):145–178.

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154.

Sankaran, K. and Holmes, S. P. (2019). Latent variable modeling for the microbiome. *Biostatistics*, 20(4):599–614.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica sinica*, 4(2):639–650.

Snijders, T. and Bosker, R. (2012). Multilevel analysis. *Netherlands: SAGE Publications.*

Teh, Y. W. (2006). A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 985–992.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Comm. Statist. Simulation Comput.*, 36(1-3):45–54.

Wang, P., Zhang, P., Zhou, C., Li, Z., and Yang, H. (2017). Hierarchical evolving dirichlet processes for modeling nonlinear evolutionary traces in temporal data. *Data Mining and Knowledge Discovery*, 31:32–64.

Wang, Z., Mao, J., and Ma, L. (2021). Microbiome compositional analysis with logistic-tree normal models. *arXiv preprint arXiv:2106.15051*.

Wood, F., Gasthaus, J., Archambeau, C., James, L., and Teh, Y. W. (2011). The sequence memoizer. *Communications of the ACM*, 54(2):91–98.

Wood, F. and Teh, Y. W. (2009). A hierarchical nonparametric bayesian approach to statistical language model domain adaptation. In *Artificial Intelligence and Statistics*, pages 607–614. PMLR.

Chapter 4

Gibbs samplers for parametric hierarchical models

4.1 Introduction

Gibbs samplers Casella and george (1992) are a family of Markov Chain Monte Carlo (MCMC) algorithms Brooks et al. (2011) commonly used in various scientific fields. In the context of Bayesian Statistics, they are routinely employed to draw samples from posterior distributions of unknown parameters conditional to the observed data Green et al. (2015); Martin et al. (2023). Like most MCMC methods, they are guaranteed to converge to the correct posterior distribution as the number of iterations tends to infinity under mild assumptions (Roberts and Sahu, 1994). However, understanding how quickly this convergence occurs, for example by quantifying the so-called mixing time of the Markov chain generated by the algorithm, is in general a hard task. In this paper we address this question for Gibbs samplers targeting certain classes of high-dimensional Bayesian hierarchical models. Analysing convergence properties, such as mixing times, is the key technical step needed to rigorously quantify the computational cost of MCMC algorithms.

Hierarchical models

Our motivating example is given by classical Bayesian hierarchical models of the form

$$\begin{aligned} Y_j \mid \theta_j &\sim f(\cdot \mid \theta_j) \quad j = 1, \dots, J, \\ \theta_j \mid \psi &\overset{\text{iid}}{\sim} p(\cdot \mid \psi) \quad j = 1, \dots, J, \\ \psi &\sim p_0(\cdot)\,. \end{aligned} \tag{4.1}$$

Here the observed dataset $Y_{1:J} = (Y_j)_{j=1,\dots,J}$ is divided into J groups, with data for each group typically containing multiple observations, e.g. $Y_j = (Y_{j1}, \dots, Y_{jm})$. Each group features some local (i.e. group-specific) parameters $\theta_j \in \mathbb{R}^\ell$, while $\psi \in \mathbb{R}^D$ are global (hyper)-parameters. Above $f(\cdot \mid \theta)$, $p(\cdot \mid \psi)$ and $p_0(\cdot)$ denote some likelihood function, local prior and global prior, respectively. See Section 4.4 for the assumptions we require on each of those. Given model (4.1), posterior inferences are based on the conditional distribution of ψ and $\boldsymbol{\theta} = (\theta_1, \dots, \theta_J)$ given $Y_{1:J}$, which we denote as $\mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi | Y_{1:J})$. Hierarchical models such as (4.1) are the workhorse of Bayesian Statistics and are commonly employed in many applied contexts (see e.g. Gelman and Hill (2007); Gelman et al. (2013) and references therein). In this paper, we are mostly

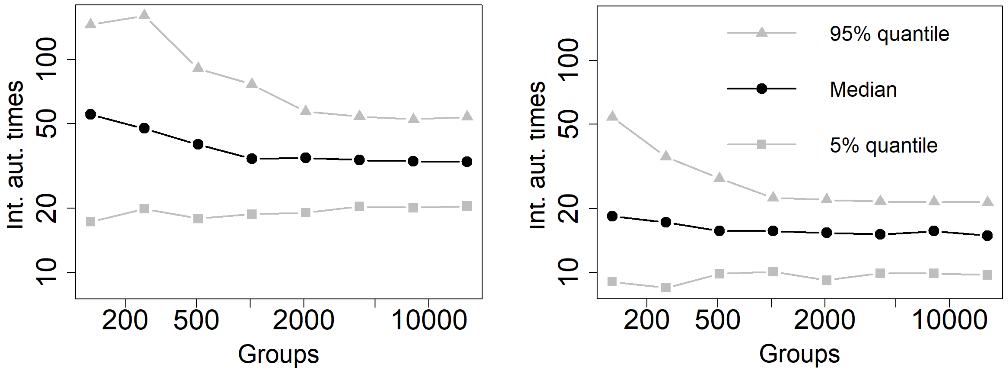


Figure 4.1: Integrated autocorrelation times (on log-scale) of Gibbs samplers targeting the posterior distribution of model (4.1) with specification (4.2). Quantiles refer to repetitions over datasets randomly generated according to the model with true parameters $\mu^* = \tau^* = 1$. Left: $m = 3$. Right: $m = 5$. See Section 4.5 for more details.

interested in the high-dimensional regime where $J \to \infty$, so that both the number of datapoints and parameters, i.e. $n = Jm$ and $p = J\ell + D$ respectively, diverge.

One iteration of a Gibbs sampler targeting $\mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi | Y_{1:J})$ sequentially samples each parameter from its full-conditional distribution, i.e. it performs the updates $\theta_j \sim \mathcal{L}(\mathrm{d}\theta_j | Y_{1:J}, \psi)$ for $j = 1, \ldots, J$ and $\psi \sim \mathcal{L}(\mathrm{d}\psi | Y_{1:J}, \boldsymbol{\theta})$. Algorithms based on conditional updates are well-suited to model (4.1), since they naturally exploit the underlying sparse dependence structure. In particular, the conditional independence of $\theta_1, \ldots, \theta_J$ given $Y_{1:J}$ and ψ implies that the sequence of updates from the low-dimensional distributions $\mathcal{L}(\mathrm{d}\theta_j | Y_{1:J}, \psi)$ for $j = 1, \ldots, J$ is equivalent to an exact joint update from the high-dimensional distribution $\mathcal{L}(\mathrm{d}\boldsymbol{\theta} | Y_{1:J}, \psi)$. Also, since local parameters interact only with local data conditional on ψ, i.e. $\mathcal{L}(\mathrm{d}\theta_j | Y_{1:J}, \psi) = \mathcal{L}(\mathrm{d}\theta_j | Y_j, \psi)$, one iteration of the Gibbs sampler can typically be implemented with a computational cost that scales linearly with J. For the sake of comparisons, a similar cost is required by a single likelihood evaluation or a single posterior gradient evaluation for model (4.1). See also Remark 13 in Section 4.4 for related discussion.

The key question to properly assess the effectiveness of Gibbs samplers targeting model (4.1) is how fast the resulting Markov chain converges to its stationary distribution $\mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi | Y_{1:J})$. Interestingly, such chain often enjoys dimension-free convergence speed, meaning that the number of iterations required to converge does not grow (or grows only logarithmically) with J. Figure 4.1 illustrates numerically this behaviour on a hierarchical logistic model, where the likelihood and prior in (4.1) are specified as

$$f(y \mid \theta) = \binom{m}{y} \frac{e^{y\theta}}{(1 + e^{\theta})^m}, \quad p(\theta \mid \psi) = N(\theta \mid \mu, \tau^{-1}), \quad \psi = (\mu, \tau), \tag{4.2}$$

with $y \in \{0, \ldots, m\}$ and m being a positive integer. The prior for $\psi = (\mu, \tau)$ is set to $\mu \mid \tau \sim N\left(0, 10^3/\tau\right)$ and $\tau \sim \text{Gamma}(0.1, 0.1)$. Full details on the simulation set-up of Figure 4.1 are described in Section 4.5. The results suggest that the number of iterations required by the Gibbs sampler to draw each sample from $\mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi | Y_{1:J})$ remains bounded as J grows and asymptotes to a finite value as $J \to \infty$. Combined with cost per iteration, this implies a computational complexity that grows linearly with J. Note that this complexity is smaller than the one of popular gradient-based MCMC methods when applied to these models (see Section 4.1 for more details), supporting the idea that Gibbs samplers can achieve state-of-the-art performances for hierarchical models with sparse dependence structures.

In Section 4.4 we provide rigorous support to the above empirical evidences. In particular, we study the asymptotic behavior of mixing times of Gibbs samplers targeting model (4.1). There we prove that mixing times remain bounded as $J \to \infty$ under mild assumptions on the likelihood f and the global prior p_0. We instead require stronger assumptions on the local priors $p(\cdot \mid \psi)$, which we assume to be in the exponential family. Our results (see e.g. Theorem 20) are average-case ones and hold with high probability with respect to the law of the data-generating process. To do so we assume the observed data $Y_{1:J}$ to be randomly generated. This allows to use tools of Bayesian asymptotics, such as Bernstein-von Mises type statements (see e.g. Chapter 10 of Van der Vaart (2000)), to characterize the asymptotic posterior behaviour as $J \to \infty$ and then extract information about the limiting behaviour of the associated sequence of MCMC algorithms.

Related literature

The literature on performances of MCMC methods is very broad. The most well-studied classes of algorithm are probably gradient-based ones, such as Langevin (Roberts and Tweedie, 1996) and Hamiltonian (Neal, 2011) Monte Carlo, see e.g. Dalalyan (2017); Durmus and Moulines (2017); Dwivedi et al. (2019) and related literature. Available results suggest that the number of iterations (or target gradient evaluations) required by those algorithm to converge to stationarity increases with dimensionality, e.g. growing as $\mathcal{O}(J^\alpha)$ with the dimensionality J, for some $\alpha > 0$ that depends on the setup and type of algorithm (Roberts and Rosenthal, 1998; Beskos et al., 2013; Wu et al., 2022). In the context of hierarchical models, given that each target gradient evaluation has a linear cost in J, this leads to a computational cost to sample from $\mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi|Y_{1:J})$ that scales super-linearly with J, e.g. as $\mathcal{O}(J^{1+\alpha})$ with $\alpha > 0$. Comparing these results to the one we develop here for Gibbs samplers suggests that, while being state-of-the-art black-box schemes to sample from generic high-dimensional distributions with appropriate regularity conditions (e.g. log-concavity), default gradient-based MCMC schemes can be suboptimal for high-dimensional hierarchical models. See also Papaspiliopoulos et al. (2023) for related numerical evidences.

Compared to gradient-based MCMC, results for Gibbs-type schemes are less abundant and more model-dependent. Notable recent examples include Yang and Rosenthal (2022); Jin and Hobert (2022); Qin and Hobert (2022), which provide convergence bounds for hierarchical models, similar to (4.1), with Gaussian and Poisson likelihoods. Another recent result is given by Qin and Hobert (2019), which provides dimension-free convergence bounds for Gibbs samplers for high-dimensional probit regression models under appropriate regimes. Providing sharp non-asymptotic analyses like the ones above requires proof techniques, such as drift-and-minorization techniques (Rosenthal, 1995) and random mappings Qin and Hobert (2019), that are usually likelihood-specific and potentially hard to construct. For example, they may require to devise and study a suitable Lyapunov function that depends on the specific choices of both likelihood and priors in (4.1) (see e.g. formulae (6) and (33) in Jin and Hobert (2022) and Yang and Rosenthal (2022), respectively). On the other hand, these approaches provide non-asymptotic bounds that apply to fixed sample size and dimensionality, thus being complimentary to the high-dimensional asymptotic analysis we develop here.

Interestingly, there are relatively few papers combining the tools of Bayesian asymptotics and MCMC theory in rigorous ways. The work in Belloni and Chernozhukov (2009) uses Bernstein-von Mises Theorem to provide polynomial bounds on the convergence of random walk

Metropolis-Hastings schemes. After that, very recent papers use similar techniques to provide complexity analysis of MCMC schemes, see e.g. Nickl and Wang (2022); Negrea et al. (2022); Tang and Yang (2022) dealing with gradient-based methods, the first in the context of inverse problems. A brief discussion about the use of asymptotic posterior characterisations to study the convergence properties of Gibbs samplers is given in Roberts and Sahu (2001). A more in-depth use of Bayesian asymptotics to study data augmentation procedures is given in Kamatani (2014), which also considers hierarchical models. See Remark 14 in Section 4.4 for more details on the results in Kamatani (2014). Finally, an interesting exception is given by Bayesian variable selection models, where multiple works have exploited the asymptotic behaviour of the posterior distribution to characterize the computational performances of Bayesian methods Yang et al. (2016); Atchadé (2021); Zhou et al. (2022).

Sketch of the main arguments and structure of the paper

The argument we employ to study Gibbs samplers targeting $\mathcal{L}(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi \mid Y_{1:J})$ can be decomposed in three main parts. First, if $p(\cdot \mid \psi)$ belongs to the exponential family, there exists a set of sufficient statistics $\boldsymbol{T} = \boldsymbol{T}(\boldsymbol{\theta})$, whose dimensionality does not depend on J, such that $\mathcal{L}\left(\mathrm{d}\psi \mid \boldsymbol{\theta}, Y_{1:J}\right) = \mathcal{L}\left(\mathrm{d}\psi \mid \boldsymbol{T}(\boldsymbol{\theta}), Y_{1:J}\right)$. Lemma 44 in Section 4.4 shows that, as a result, the Gibbs sampler on $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi \mid Y_{1:J}\right)$ has the same mixing times as the one on $\mathcal{L}\left(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi \mid Y_{1:J}\right)$. This allows to focus on the latter distribution which, unlike the former, is intractable but fixed dimensional. Note that this dimensionality reduction does not require the likelihood f to admit sufficient statistics (see Remark 12) and is a peculiar property of Gibbs samplers, since it exploits the presence of exact updates. The second step consists in studying the asymptotic behaviour of $\mathcal{L}\left(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi \mid Y_{1:J}\right)$ as J increases. In particular, Proposition 23 shows that a suitable rescaling of $(\boldsymbol{T}, \psi)$ converges to a multivariate Gaussian distribution in total variation distance. The proof combines a classical Bernstein-von Mises Theorem for ψ (Lemma 45) with a less standard Central Limit Theorem for $\boldsymbol{T}$ conditional on ψ (Lemma 46). More details can be found in Section 4.4. The final and key point is then to connect the convergence of the target distributions, in this case $\{\mathcal{L}\left(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi \mid Y_{1:J}\right)\}_{J\geq 1}$, to the convergence of the associated Gibbs sampler operators. Theorem 18 proves that the limiting behaviour of a sequence of Gibbs samplers is equivalent to the behaviour of the Gibbs sampler on the limiting distribution: this is shown in total variation distance and under warm start assumption. The fundamental link is given by Proposition 21, which provides an upper bound on the distance between Gibbs sampler operators in terms of the one between the target distributions. Since those results are of independent interest and are not specific to hierarchical models, we start by developing those in a general setup in Section 4.2. Then, Section 4.3 recalls the Bernstein-von Mises Theorem and illustrates the results of Section 4.2 to the fixed-dimensional setting. Section 4.4 develops the main results of the paper dealing with general hierarchical models (see e.g. Theorem 20) and Section 4.5 verifies the general conditions for some specific likelihood families, e.g. Gaussian, binomial and categorical, together with providing numerical simulations and extension to different graphical model structures. Since a warm start initialization for the sampler is assumed throughout, the availability of feasible starts is discussed in Section 4.6. Finally, Section 4.7 discusses extensions and future work.

4.2 Gibbs sampler and asymptotics

In this section, after recalling basic definitions about Gibbs kernels and mixing times, we connect the convergence of a sequence of target distributions to the convergence of the associated Gibbs kernels. This leads to Theorem 18, which characterizes the limiting behaviour of the Gibbs samplers mixing times. Throughout this section, the target distributions are assumed to have fixed dimensionality.

Setup and notation

Let $(\pi_n)_{n\geq 1} = \left(\pi_n(\cdot \mid Y^{(n)})\right)_{n\geq 1}$ be a sequence of probability distributions on a common product space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_K$, where each π_n is allowed to depend on some observed data $Y^{(n)} \in \mathcal{Y}^{(n)}$. In our applications, $\pi_n(\cdot \mid Y^{(n)})$ represents the posterior distribution of some unknown parameter $\mathbf{x} \in \mathcal{X}$ conditioned on the data $Y^{(n)}$. For the sake of brevity, we will often omit the explicit dependence on $Y^{(n)}$.

Let P_n be the Markov transition kernel of the deterministic-scan Gibbs sampler targeting π_n, defined as the product of K kernels

$$P_n = P_{n,1} \cdots P_{n,K}\,. \tag{4.3}$$

For each $i \in \{1, \dots, K\}$, $P_{n,i}$ is the transition kernel on $\mathcal{X}$ that updates the i-th coordinate drawing it from its conditional distribution $\pi_n(\mathrm{d}x_i|\mathbf{x}^{(-i)})$, where $\mathbf{x}^{(-i)} = (x_j)_{j\neq i}$, while leaving the other components unchanged. Equivalently

$$P_{n,i}\left(\mathbf{x}, S_{\mathbf{x},i,A}\right) = \int_A \pi_n\left(\mathrm{d}y_i \mid \mathbf{x}^{(-i)}\right), \quad A \subset \mathcal{X}_i, \quad i = 1, \dots, n,$$

with $S_{\mathbf{x},i,A} = \{\mathbf{y} \in \mathcal{X} \,:\, y_j = x_j \,\forall\, j \neq i \text{ and } y_i \in A\}$. It is easy to show that $P_{n,i}$ is reversible with respect to π_n for every i, so that π_n is the invariant distribution of P_n (Roberts and Rosenthal, 2004; Hobert, 2011; Chlebicka et al., 2023).

Given $\epsilon \in (0,1)$, define the ϵ-total variation mixing time of P_n with starting distribution $\mu_n \in \mathcal{P}(\mathcal{X})$, where $\mathcal{P}(\mathcal{X})$ denotes the set of probability distribution on $\mathcal{X}$, as

$$t^{(n)}_{mix}(\epsilon, \mu_n) = \inf\left\{t \geq 0 \,:\, \left\|\mu_n P^t_n - \pi_n\right\|_{TV} < \epsilon\right\}, \tag{4.4}$$

where P^t denotes the t-th power of P, $\mu_n P^t_n(A) = \int_{\mathcal{X}} P^t_n(\mathbf{x}, A)\mu_n(\mathrm{d}\mathbf{x})$ for any $A \subseteq \mathcal{X}$ and $\|\cdot\|_{TV}$ denotes the total variation norm. By definition, mixing times quantify the number of Markov chain's iterations required to obtain a sample from the target distribution π_n up to error ϵ. We will focus on worst-case mixing times with respect to M-warm starts. The set of M-warm starts relative to a distribution π is defined as

$$\mathcal{N}(\pi, M) = \{\mu \in \mathcal{P}(\mathcal{X}) \,:\, \mu(A) \leq M\pi(A) \text{ for all } A \subseteq \mathcal{X}\}\,, \qquad M \geq 1,\, \pi \in \mathcal{P}(\mathcal{X})\,, \tag{4.5}$$

and the associated worst-case mixing times for P_n targeting π_n are

$$t^{(n)}_{mix}(\epsilon, M) = \sup_{\mu_n \in \mathcal{N}(\pi_n, M)} t^{(n)}_{mix}(\epsilon, \mu_n)\,. \tag{4.6}$$

Remark 3. *While being common in the literature, see e.g. Dalalyan (2017); Dwivedi et al. (2019); Tang and Yang (2022) for gradient-based methods, the warm start assumption can be quite stringent and potentially unrealistic. In particular, assuming that the algorithm can be initialised by sampling the starting configuration from a warm start with relatively small M (e.g. one that does not grow exponentially fast with dimensionality) may be unrealistic. In Section*

4.6 we show that in the specific case of hierarchical models as in (4.1) *a feasible start, i.e. a starting distribution which can be implemented in practice and allows to control the value of M, is available under some assumptions.*

Assumptions on the sequence of target distributions

We consider settings where a rescaled version of the sequence $(\pi_n)_{n\geq 1}$ converges to a well defined limiting distribution as $n \to \infty$. This is often the case in a Bayesian context where some version of the Bernstein von-Mises theorem holds (see e.g. Theorem 19 below). The convergence of $(\pi_n)_{n\geq 1}$ occurs with high probability assuming the data $Y^{(n)}$ is randomly generated from some distribution. In particular, we assume for the rest of this section that $Y^{(n)}$ is random with distribution $Q^{(n)} \in \mathcal{P}\left(\mathcal{Y}^{(n)}\right)$. The following assumption specifies the convergence we require for $(\pi_n)_{n\geq 1}$:

(A1) There exists $\tilde{\pi} \in \mathcal{P}(\mathcal{X})$ and a sequence of transformations $\phi_n : \mathcal{X} \to \mathcal{X}$ that act *coordinate-wise*, i.e. where

$$\phi_n(\mathbf{x}) = \left(\phi_{n,1}(x_1), \dots, \phi_{n,K}(x_K)\right) , \qquad \mathbf{x} \in \mathcal{X} \tag{4.7}$$

with $\phi_{n,j} : \mathcal{X}_j \to \mathcal{X}_j$ injective and measurable, such that

$$\|\tilde{\pi}_n - \tilde{\pi}\|_{TV} \to 0 \qquad \text{as } n \to \infty , \tag{4.8}$$

in $Q^{(n)}$-probability, i.e. such that $\lim_{n\to\infty} Q^{(n)}(\|\tilde{\pi}_n - \tilde{\pi}\|_{TV} > \epsilon) = 0$ for every $\epsilon \in (0,1)$, where $\tilde{\pi}_n = \pi_n \circ \phi_n^{-1}$ is the law of $\tilde{\mathbf{x}} = \phi_n(\mathbf{x})$ under $\mathbf{x} \sim \pi_n$.

Remark 4. *The necessity of rescaling $\boldsymbol{x}$ by some transformation ϕ_n in* (4.7) *comes from the typical behaviour of posterior distributions in Bayesian models. Indeed, without rescaling, π_n often converges to a random variable which is degenerate to a Dirac delta at a fixed value (e.g. the underlying data-generating parameter). Thus, in order to have a non-trivial limit and total variation convergence, which is essential for our purposes, a suitable rescaling is needed. In our context the specific form of this transformation is dictated by the theory of Bayesian asymptotics, see e.g. Theorem 19 below. Moreover, we assume ϕ_n to act coordinate-wise because this class of transformations leaves Gibbs samplers invariant (see e.g. Lemma 42 below), while general one-to-one transformations can alter the Gibbs sampler dynamics and change its convergence speed (Papaspiliopoulos et al., 2007b).*

Remark 5. *The results we develop below could be extended to more general versions of assumption (A1), including ones where the co-domain of ϕ_n is not equal to the domain, i.e. $\phi_n : \mathcal{X} \to \mathcal{Z}$ for some $\mathcal{Z}$, and where the limiting distribution $\tilde{\pi}$ is random, i.e. allowed to depend on the sequence $(Y^{(n)})_n$. Since* (A1) *is enough for our purposes and motivating applications, we do not consider such extensions here to keep notation simple.*

Let $\tilde{P}$ and $\tilde{P}_n$ be the kernels of the Gibbs samplers targeting $\tilde{\pi}$ and $\tilde{\pi}_n$, respectively. The following lemma shows that studying total variation convergence from M-warm starts for the sequence of kernels $(P_n)_{n\geq 1}$ is equivalent to doing it for the sequence $(\tilde{P}_n)_{n\geq 1}$. The proof, which can be found in Appendix C, relies on the coordinate-wise and bijective requirements of (A1).

Lemma 42. *Under Assumption (A1) we have*

$$\sup_{\mu_n \in \mathcal{N}(\pi_n, M)} \left\|\mu_n P_n^t - \pi_n\right\|_{TV} = \sup_{\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)} \left\|\tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n\right\|_{TV} .$$

Convergence of Gibbs samplers operators

Since by $(A1)$ the stationary distribution of $\tilde{P}_n$, the Gibbs samplers targeting $\tilde{\pi}_n$, converges to the one of $\tilde{P}$, one may be tempted to translate such convergence at the level of the kernels, e.g. $\|\tilde{P}_n(\mathbf{x},\cdot) - \tilde{P}(\mathbf{x},\cdot)\|_{TV} \to 0$ for ($\tilde{\pi}$-almost) every $\mathbf{x} \in \mathcal{X}$. However this is not only false for generic Markov operators, but even in the special class of Gibbs sampler operators: one can have $\|\tilde{\pi}_n - \tilde{\pi}\|_{TV} \to 0$ as $n \to \infty$, while $\|\tilde{P}_n(\mathbf{x},\cdot) - \tilde{P}(\mathbf{x},\cdot)\|_{TV} \nrightarrow 0$ for any $\mathbf{x} \in \mathcal{X}$, see e.g. Example $A.1$ in Appendix A. The reason is that convergence of the joint distribution $\tilde{\pi}_n$ in total variation distance does not imply convergence of the associated conditional distributions, that are the building blocks of the Gibbs sampler operator. However, it turns out that a control on the total variation distance between two target distributions is in general sufficient to control the distance between the corresponding Gibbs sampler operators applied to warm starts. The following Proposition makes the connection precise. Interestingly, no assumptions on the target distribution and Gibbs samplers are required.

Proposition 21. *Let P_1 and P_2 be the transition kernels of Gibbs samplers targeting $\pi_1 \in \mathcal{P}(\mathcal{X})$ and $\pi_2 \in \mathcal{P}(\mathcal{X})$, respectively. Then we have*

$$\|\mu P_1 - \mu P_2\|_{TV} \leq 2MK\, \|\pi_1 - \pi_2\|_{TV}\,, \tag{4.9}$$

for every $\mu \in \mathcal{N}(\pi_1, M) \cup \mathcal{N}(\pi_2, M)$ and $M \geq 1$.

Proposition 21 translates convergence of the stationary distributions, given by $(A1)$, into convergence of the Gibbs samplers operators when a warm start is considered. It is worth noting that a bound of this form cannot hold for generic Markov transition kernels. Indeed, consider transition kernels P_1 and P_2 with the same stationary distribution π: by basic properties of the total variation distance it holds $\|\mu P_1 - \mu P_2\|_{TV} \leq 2\,\|\mu - \pi\|_{TV}$. The latter bound cannot be improved in general, meaning that it is possible to find ergodic kernels P_1 and P_2 that get arbitrarily close to the above upper bound, see Example A.2 in Appendix A.

Proposition 21 is used in the proof of Theorem 18, which shows that the limiting behaviour of P_n, in terms of distance to stationarity from M-warm starts, is completely characterized by the behaviour of the limiting operator $\tilde{P}$. The proof of Theorem 18 also relies on the fact that the total variation distance between π_1 and π_2 provides a control on the distance between the two sets $\mathcal{N}(\pi_1, M)$ and $\mathcal{N}(\pi_2, M)$, as shown in the following Lemma.

Lemma 43. *Let $\pi_1, \pi_2 \in \mathcal{P}(\mathcal{X})$. Then, for every $\mu_1 \in \mathcal{N}(\pi_1, M)$, there exists $\mu_2 \in \mathcal{N}(\pi_2, M)$ such that $\|\mu_1 - \mu_2\|_{TV} \leq M\,\|\pi_1 - \pi_2\|_{TV}$.*

Lemma 43 implies that, under assumption $(A1)$, for every $\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)$ there exists a sequence $\{\tilde{\mu}_n\}_n$ such that $\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)$ and $\|\tilde{\mu}_n - \tilde{\mu}\|_{TV} \to 0$ as $n \to \infty$ in $Q^{(n)}$-probability. We can now state Theorem 18.

Theorem 18. *Let assumption $(A1)$ holds. Then for every $t \in \mathbb{N}$ and $M \geq 1$ it holds*

$$\lim_{n\to\infty} \sup_{\mu_n \in \mathcal{N}(\pi_n, M)} \left\|\mu_n P_n^t - \pi_n\right\|_{TV} = \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\|\tilde{\mu}\tilde{P}^t - \tilde{\pi}\right\|_{TV}\,,$$

in $Q^{(n)}$-probability.

Remark 6. *An alternative approach to derive convergence statements on the sequence of Gibbs kernels would be to consider stronger forms of convergence for the sequence $(\tilde{\pi}_n)_{n\geq 1}$ than the one* *in total variation distance in* (4.8)*. However, we prefer to derive results under weaker convergence requirements for* $(\tilde{\pi}_n)_{n\geq 1}$ *to allow for a more direct use of standard asymptotic results in the Bayesian literature (e.g. common formulations of the Bernstein-von Mises theorem), which are usually derived in terms of weaker metrics such as total variation one.*

Implications for mixing times

Denote the mixing times of $\tilde{P}$ as

$$\tilde{t}_{mix}(\epsilon, M) = \sup_{\tilde{\mu}\in\mathcal{N}(\tilde{\pi},M)} \inf\left\{t \geq 1 \,:\, \left\|\tilde{\mu}\tilde{P}^t - \tilde{\pi}\right\|_{TV} < \epsilon\right\}.$$

The following corollary of Theorem 18 shows how to use $\tilde{t}_{mix}(\epsilon, M)$ to deduce statements on the behaviour of the sequence of mixing times of interest, $(t^{(n)}_{mix}(\epsilon, M))_{n\geq 1}$.

Corollary 8. *Let assumption* $(A1)$ *holds. If* $(M, \epsilon) \in [1,\infty)\times(0,1)$ *is such that* $\tilde{t}_{mix}(\epsilon, M) < \infty$*, then*

$$Q^{(n)}\left(t^{(n)}_{mix}(\epsilon, M) \leq \tilde{t}_{mix}(\epsilon, M)\right) \to 1 \tag{4.10}$$

as $n\to\infty$*. Otherwise, if* $(M, \epsilon) \in [1,\infty)\times(0,1)$ *is such that* $\tilde{t}_{mix}(\epsilon, M) = \infty$*, then it holds*

$$Q^{(n)}\left(t^{(n)}_{mix}(\underline{\epsilon}, M) < T\right) \to 0$$

as $n\to\infty$*, for every* $\underline{\epsilon} < \epsilon$ *and* $T > 0$*.*

Remark 7 (Mixing times bounded in probability)**.** *When* $\tilde{t}_{mix}(\epsilon, M) < \infty$*, the statement in* (4.10) *implies that* $t^{(n)}_{mix}(\epsilon, M) = \mathcal{O}_P(1)$ *as* $n\to\infty$*, i.e. that the sequence of random variables* $(t^{(n)}_{mix}(\epsilon, M))_{n\geq 1}$ *is bounded in probability. The latter means that for every* $\delta > 0$ *there exist an integer* N_δ *and a real constant* $B_\delta < \infty$ *such that* $Q^{(n)}(t^{(n)}_{mix}(\epsilon, M) \leq B_\delta) \geq 1-\delta$ *for every* $n \geq N_\delta$*, which holds by* (4.10) *taking* $B_\delta = \tilde{t}_{mix}(\epsilon, M)$*.*

By Corollary 8, establishing whether $\tilde{P}$ is ergodic (in the sense of yielding finite mixing times) or not is enough to discriminate between sequences of kernels $(P_n)_{n\geq 1}$ whose mixing times diverge as $n\to\infty$ as opposed to ones that do not (see e.g. Figure 4.4 in Section 4.5 for an illustration). Since ergodicity of Gibbs samplers can be established under very mild assumptions (Roberts and Sahu, 1994), in practice one can expect $\tilde{P}$ to be ergodic and thus $(t^{(n)}_{mix}(\epsilon, M))_{n\geq 1}$ to be bounded in probability whenever $(A1)$ holds for a well-behaved, non-singular limiting distribution $\tilde{\pi}$. Sections 4.4 and 4.5 combine Corollary 8 with dimensionality reduction techniques to provide results on Gibbs samplers targeting high-dimensional hierarchical models.

Remark 8 (Alternative metrics)**.** *It is natural to wonder whether the result of Corollary 8 may hold for weaker metrics, like the one induced by the Wasserstein distance. However, it is possible to find examples where the convergence of the stationary distributions (in Wasserstein distance) does not imply convergence of the associated mixing times (neither the ones defined based on the TV distance nor the ones defined based on the Wasserstein one). The intuition is that the limiting distribution in weaker metrics (e.g. Wasserstein, weak convergence, etc) may ignore features of the joint distribution, such as full conditionals behaviours, that have a relevant* *impact on Gibbs sampler dynamics. For example, a sequence of increasingly correlated random variables (whose Gibbs samplers converge slower and slower) may converge to a single point mass, for which independence and immediate convergence automatically holds. See Example A.3 in Appendix A.*

Explicit limiting bounds

Corollary 8 can also be used to derive quantitative bounds on the limiting behaviour of the mixing times $(t_{mix}^{(n)}(\epsilon, M))_{n\geq 1}$. In particular, if one is able to establish explicit bounds on $\tilde{t}_{mix}(\epsilon, M)$, then (4.10) implies a corresponding bound in high probability on $t_{mix}^{(n)}(\epsilon, M)$ for large n. While deriving quantitative bounds on Gibbs samplers mixing times is in general hard, the limiting distribution $\tilde{\pi}$ is often more tractable than the original sequence $(\pi_n)_{n\geq 1}$, a common case being the one where $\tilde{\pi}$ is multivariate Gaussian while $(\pi_n)_{n\geq 1}$ is not. In those scenarios explicit bounds on $\tilde{t}_{mix}(\epsilon, M)$ can be derived using available results on the convergence properties of Gibbs samplers targeting multivariate Gaussian distributions, see e.g. Amit (1991); Khare and Zhou (2009); Roberts and Sahu (1997). For example, Theorem 2 in Amit (1991) provides an explicit bound for deterministic scan Gibbs samplers on Gaussian targets in L^2-distance (and therefore total variation Andrieu et al. (2022)).

In Sections 4.4 and 4.5 we will apply this strategy mostly to cases where $K = 2$, meaning that $\tilde{P}$ is a two-block Gibbs sampler. In this situation, one can use spectral gaps to bound Gibbs samplers mixing times, as shown in the Corollary 9. Given a π-invariant kernel P with $\pi \in \mathcal{P}(\mathcal{X})$ we define its spectral gap as

$$\mathrm{Gap}(P) = \inf_{f\,:\,\pi(f^2)<\infty,\,\mathrm{Var}_\pi(f)>0} \left\{ \frac{\int_{\mathcal{X}^2} \left[f(\mathbf{y}) - f(\mathbf{x})\right]^2 \pi(\mathrm{d}\mathbf{x}) P(\mathbf{x}, \mathrm{d}\mathbf{y})}{2\mathrm{Var}_\pi(f)} \right\},$$

where $f : \mathcal{X} \to \mathbb{R}$ are measurable functions, $\pi(f) = \int_{\mathcal{X}} f(\mathbf{x})\pi(\mathrm{d}\mathbf{x})$ and $\mathrm{Var}_\pi(f) = \int_{\mathcal{X}} \left[f(\mathbf{x}) - \pi(f)\right]^2 \pi(\mathrm{d}\mathbf{x})$. We refer to Rosenthal and Rosenthal (2015) and the proof of Corollary 9 for discussion on why spectral gaps, which are commonly used for π-reversible chains, can be used to analyse two-block Gibbs samplers, which are technically not reversible. We also note that Corollary 9 is only one possible approach to bound $\tilde{t}_{mix}(\epsilon, M)$ and that any quantitative bound on the latter can be combined with Corollary 8 to deduce limiting statements on $(t_{mix}^{(n)}(\epsilon, M))_{n\geq 1}$.

Corollary 9. *Let $K = 2$, assumption $(A1)$ be satisfied and $Gap(\tilde{P}) > 0$. Then, for every $(M, \epsilon) \in [1, \infty) \times (0, 1)$ it holds*

$$Q^{(n)} \left(t_{mix}^{(n)}(\epsilon, M) \leq 1 + \frac{\log(M/2) - \log(\epsilon)}{-\log(1 - Gap(\tilde{P}))} \right) \to 1 \qquad as\ n \to \infty\,.$$

Given the result of Corollary 9, it is natural to ask whether the convergence proved in Theorem 18 could be rephrased in terms of spectral gaps, i.e. $\mathrm{Gap}(P_n) \to \mathrm{Gap}(\tilde{P})$. However, once again, convergence in total variation is too weak for this purpose: indeed it is not difficult to find examples where $(A1)$ holds and the associated Gibbs sampler spectral gaps do not converge, even under the stronger condition requiring $\|\tilde{P}_n(\mathbf{x}, \cdot) - \tilde{P}(\mathbf{x}, \cdot)\|_{TV} \to 0$ for any $\mathbf{x} \in \mathcal{X}$, see Example A.4 in Appendix A. Controlling directly the spectral gaps would require extremely stringent conditions on the convergence of $\tilde{\pi}_n$ to $\tilde{\pi}$ that are rarely satisfied (e.g. uniform convergence of the associated densities on the log-scale, i.e. $\sup_{\mathbf{x}\in\mathcal{X}} |\log \tilde{\pi}_n(\mathbf{x}) - \log \tilde{\pi}(\mathbf{x})| \to 0$). An alternative approach to the direct warm-start mixing time analysis that we perform here, would be to consider asymptotic behaviours of *approximate* spectral measures, such as approximate spectral gaps, see e.g. Atchadé (2021); Tang and Yang (2022).

4.3 Illustrative example: fixed-dimensional parametric models

We first consider the fixed-dimensional case. While this is not our main interest or motivating application, it allows to show the type of results we will derive and also introduce notation about classical Bayesian asymptotic results that we will use. In this setting $\pi_n(\mathrm{d}\psi) = p(\mathrm{d}\psi \mid Y^{(n)})$ is the posterior distribution of the Bayesian model defined as

$$Y_i \mid \psi \overset{iid}{\sim} f(Y \mid \psi), \quad \psi \sim p_0(\psi), \tag{4.11}$$

where $\psi = (\psi_1, \dots, \psi_K)$, with $\mathcal{X} \in \mathbb{R}^K$, and $Y^{(n)} = (Y_1, \dots, Y_n)$, with $Y_i \in \mathcal{Y}$, $i = 1, \dots, n$, so that $\mathcal{Y}^{(n)} = \mathcal{Y}^n$. Moreover, if $Y_i \overset{\text{iid}}{\sim} Q$ for some $Q \in \mathcal{P}(\mathcal{Y})$, we denote with $Q^{(n)}$ and $Q^{(\infty)}$ the associated product measures. We study the mixing times of the Gibbs sampler that updates one coordinate of ψ at the time as n grows. In order to apply the results of Theorem 18 we need a suitable transformation of ψ, that is given by the celebrated Bernstein-von Mises Theorem, which we now recall. The version we provide here, which makes stronger than needed assumptions, can be obtained combining Theorem 10.1 in Van der Vaart (2000), with other remarks in Chapter 10 therein, incuding Lemmas 10.4 and 10.6.

Theorem 19 (Bernstein-von Mises)**.** *Consider model* (4.11) *and let the map $\psi \to f(\cdot \mid \psi)$ be one-to-one. Let the map $\psi \to \sqrt{f(y \mid \psi)}$ be continously differentiable for every $y \in \mathcal{Y}$, with non-singular and continuous Fisher Information $\mathcal{I}(\psi)$. Let the prior measure be absolutely continuous in a neighborhood of $\psi^* \in \mathcal{X}$ with a continuous positive density at ψ^*. Finally, let Ψ be a compact neighborhood of ψ^* for which there exists a sequence of tests u_n such that*

$$\begin{gathered}\int_{\mathcal{Y}^{(n)}} u_n(y_1, \dots, y_n) \prod_{i=1}^n f(dy_i \mid \psi^*) \to 0, \\ \sup_{\psi \notin \Psi} \int_{\mathcal{Y}^{(n)}} [1 - u_n(y_1, \dots, y_n)] \prod_{i=1}^n f(dy_i \mid \psi) \to 0, \quad \text{as } n \to \infty\,.\end{gathered} \tag{4.12}$$

Then, if $Y_i \overset{iid}{\sim} Q_{\psi^*}$ for $i = 1, 2, \dots$ with Q_{ψ^*} admitting density $f(y \mid \psi^*)$, it holds

$$\left\| \mathcal{L}\left(d\tilde{\psi} \mid Y^{(n)}\right) - N\left(\mathcal{I}^{-1}(\psi^*)\Delta_{n,\psi^*}, \mathcal{I}^{-1}(\psi^*)\right) \right\|_{TV} \to 0, \qquad \text{as } n \to \infty$$

in $Q_{\psi^*}^{(\infty)}$-probability, where $\tilde{\psi} = \sqrt{n}(\psi - \psi^*)$ and $\Delta_{n,\psi^*} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \nabla \log f(Y_i \mid \psi)\Big|_{\psi=\psi^*}$.

Remark 9. *Differentiability of $\sqrt{f(y \mid \psi)}$ and continuity of $\mathcal{I}(\psi)$ imply that the model is* differentiable in quadratic mean*, which allows to prove local asymptotic normality of the log-likelihood function. See Theorem* 7.2 *and Lemma* 7.6 *in Van der Vaart (2000).*

Remark 10. *A* test *is a measurable function $u : \mathcal{Y}^{(n)} \to [0, 1]$. The integrals in* (4.12) *represent probabilities of errors of first and second kind, respectively, when the null hypothesis $H_0 : \psi = \psi^*$ is rejected with probability $u(y_1, \dots, y_n)$.*

Loosely speaking, Theorem 19 implies that, if the model is well-specified and ψ is suitably rescaled, the posterior distribution converges to a multivariate normal. The result holds under some identifiability requirements: first of all, the true parameter ψ^* must belong to the support of the prior; moreover, we must be able to separate ψ^* from the complements of its neighborhood, given infinitely many data. Such assumption is mild in most interesting cases and it is implied by the existence of uniformly consistent estimators for ψ (that is guaranteed if the support of p_0 is compact). See Chapter 10 in Van der Vaart (2000) for more details. Finally, the Fisher Information matrix must be non singular.

Remark 11. *Notice that Theorem 19 requires the model to be (perfectly) well-specified, which rarely happens in practice. However there exist extended versions for the case of misspecified likelihoods (Kleijn and van der Vaart, 2012), where the limiting distribution is still Gaussian with a different covariance matrix. Indeed, we expect the results of this and the following sections to hold in a similar way under misspecification: of course the different limiting distribution will have an impact on the final result, especially in the application of Corollary 9.*

We can now use Theorem 18 and Corollary 8 to bound the mixing times of the Gibbs sampler associated to model (4.11) as n diverges.

Proposition 22. *Let model* (4.11) *satisfy the hypotheses of Theorem 19 and let P_n be the Gibbs sampler kernel targeting $\pi_n(d\psi) = p(d\psi \mid Y^{(n)})$ by updating one coordinate of $\psi = (\psi_1, \dots, \psi_K)$ at a time. Then, for every $(M, \epsilon) \in [1, \infty) \times (0, 1)$ there exists $T(\psi^*, \epsilon, M) < \infty$ such that*

$$\lim_{n\to\infty} Q_{\psi^*}^{(n)} \left(t_{mix}^{(n)}(\epsilon, M) \leq T(\psi^*, \epsilon, M) \right) = 1 \, .$$

Proposition 22 shows that, under the conditions of Theorem 19 and starting from an M-warm distribution, the number of iterations required to get ϵ-close to the posterior distribution does not grow as $n \to \infty$. An application to the normal model with unknown mean and precision is given by Corollary $C.7$ in Section $C.10$ of Appendix C.

The main take-away of this Section is that, under relatively mild conditions, the Gibbs sampler behaves well with models of fixed dimensionality and growing number of observations. In the remaining of the paper we consider the more challenging setting of hierarchical models, where the number of parameters grows with the number of observations: in particular we will explore situations in which the number of required iterations remains fixed even with a growing dimensionality of the problem.

4.4 Hierarchical models with exponential family priors

We consider a general class of hierarchical models, with data divided in J groups, each having a set of group-specific parameters θ_j. The latter share a common prior with hyper-parameters ψ. Recalling (4.1), the model under consideration is

$$Y_j \mid \theta_j \sim f(\cdot \mid \theta_j)\,, \quad \theta_j \mid \psi \overset{\text{iid}}{\sim} p(\cdot \mid \psi)\,, \quad \psi \sim p_0(\cdot). \tag{4.13}$$

We assume that the prior for $\theta_j \in \mathbb{R}^\ell$ belongs to the exponential family, that is

$$p(\theta \mid \psi) = h(\theta) \exp\left\{ \sum_{s=1}^{S} \eta_s(\psi) T_s(\theta) - A(\psi) \right\}, \tag{4.14}$$

where $\psi \in \mathbb{R}^D$, $h : \mathbb{R}^\ell \to \mathbb{R}_+$ is a non-negative function and $\eta_s(\psi)$, $T_s(\theta)$ and $A(\psi)$ are known real-valued functions with domains $\mathbb{R}^D$, $\mathbb{R}^\ell$ and $\mathbb{R}^D$ respectively. We will always assume the family to be minimal, that is both $(\eta_1(\psi), \dots, \eta_S(\psi))$ and $(T_1(\theta), \dots, T_S(\theta))$ are linearly independent. On the other hand, we let $f(y \mid \theta)$ be an arbitrary likelihood function with data $y \in \mathbb{R}^m$ and parameters $\theta \in \mathbb{R}^\ell$, dominated by a suitable σ-finite measure (usually Lebesgue or counting one).

Denoting $\boldsymbol{\theta} = (\theta_1, \dots, \theta_J)$, $Y_{1:J} = (Y_1, \dots, Y_J)$ and $\pi_J(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi) = \mathcal{L}\left(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi \mid Y_{1:J}\right)$, we are interested in studying the two-block Gibbs sampler targeting $\pi_J(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi)$, i.e. the kernel defined as

$$P_J\left(\left(\boldsymbol{\theta}^{(t-1)}, \psi^{(t-1)}\right), \left(\mathrm{d}\boldsymbol{\theta}^{(t)}, \mathrm{d}\psi^{(t)}\right)\right) = \pi_J\left(\mathrm{d}\boldsymbol{\theta}^{(t)} \mid \psi^{(t-1)}\right) \pi_J\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{\theta}^{(t)}\right). \tag{4.15}$$

Throughout Section 4.4 we denote by $\left(\boldsymbol{\theta}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$ the Markov chain with operator P_J, and by $t_{mix}^{(J)}$ the associated mixing times, i.e.

$$t_{mix}^{(J)}(\epsilon, \mu) = \inf\left\{t \geq 0 \,:\, \left\|\mu P_J^t - \pi_J\right\|_{TV} < \epsilon\right\}, \quad t_{mix}^{(J)}(\epsilon, M) = \sup_{\mu \in \mathcal{N}(\pi_J, M)} t_{mix}^{(J)}(\epsilon, \mu).$$

Dimensionality reduction

In order to apply Corollary 8 to characterize $t_{mix}^{(J)}$, we would need to study the asymptotic distribution of π_J as $J \to \infty$. The latter is a distribution over $\ell J + D$ parameters, therefore its dimensionality grows with the size of the data. However, the next lemma shows that the convergence properties of P_J can be described through a Gibbs sampler on an intractable, but fixed-dimensional target, namely $\hat{\pi}_J(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi) = \mathcal{L}\left(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi \mid Y_{1:J}\right)$ where $\boldsymbol{T} = \left(\sum_{j=1}^J T_1(\theta_j), \dots, \sum_{j=1}^J T_S(\theta_j)\right)$, with T_s as in (4.14). Let $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)_{t\geq 1} = \left(\boldsymbol{T}(\boldsymbol{\theta}^{(t)}), \psi^{(t)}\right)_{t\geq 1}$ be the stochastic process obtained as a time-wise mapping of $\left(\boldsymbol{\theta}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$ under $(\boldsymbol{\theta}, \psi) \mapsto (\boldsymbol{T}(\boldsymbol{\theta}), \psi)$. The latter process contains all the information characterising the convergence of $\left(\boldsymbol{\theta}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$, in the sense made precise in the following lemma. Below we denote by $\hat{P}_J$ the kernel of the two-block Gibbs sampler targeting $\hat{\pi}_J$.

Lemma 44. *For each $J \geq 1$, the process $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$ is a Markov chain, its transition kernel coincides with $\hat{P}_J$, and its mixing times $\hat{t}_{mix}^{(J)}$ satisfy*

$$\sup_{\mu \in \mathcal{N}(\pi_J, M)} t_{mix}^{(J)}(\epsilon, \mu) = \sup_{\nu \in \mathcal{N}(\hat{\pi}_J, M)} \hat{t}_{mix}^{(J)}(\epsilon, \nu) \qquad (M, \epsilon) \in [1, \infty) \times (0, 1).$$

Remark 12 (Prior and likelihood assumptions)**.** *In order to reduce the dimensionality of the Markov chain under consideration, Lemma 44 requires the existence of sufficient statistics only for the prior density of the group-specific parameters. It does not require any condition on the likelihood function in model* (4.13)*. In particular, we have $\mathcal{L}\left(d\psi \mid \boldsymbol{\theta}, Y_{1:J}\right) = \mathcal{L}\left(d\psi \mid \boldsymbol{T}(\boldsymbol{\theta}), Y_{1:J}\right)$, while $\mathcal{L}\left(dY_{1:J} \mid \boldsymbol{\theta}, \psi\right) \neq \mathcal{L}\left(dY_{1:J} \mid \boldsymbol{T}(\boldsymbol{\theta}), \psi\right)$ in general.*

Lemma 44 allows to focus the analysis on the convergence speed of $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$, which is a chain whose dimensionality does not grow with the size of the data. Note that its target distribution $\hat{\pi}_J$ is usually not available in closed form, and the corresponding two-block Gibbs sampler $\hat{P}_J$ cannot be implemented directly (unless by implementing the original algorithm P_J and keeping track of $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$). In this sense the latter chain is useful for convergence analysis purposes but less so as an algorithmic shortcut.

The result of Lemma 44 is a peculiar property of the Gibbs sampler, which naturally ignores ancillary information about ψ in $\boldsymbol{\theta}$. Indeed, the proof of Lemma 44 crucially relies on the fact that the algorithm is performing exact conditional updates and analogous reductions do not occur for most other MCMC schemes (e.g. Metropolis-Hastings based schemes, including gradient-based ones).

This dimensionality reduction trick can be applied beyond hierarchical models and has already been employed in similar settings, mainly with the idea of obtaining suitable drift functions (Rosenthal, 1995): for example, in Qin and Hobert (2019) it is used to derive the convergence complexity of a data augmentation algorithm for the Bayesian probit regression model, while in Rajaratnam and Sparks (2015) a similar tecnique allows to study the geometric convergence rate of a Gibbs sampler for high dimensional Bayesian linear regression.

Regularity assumptions and main result

In order to apply the techniques of Theorem 18, we need to provide an asymptotic characterization of $\tilde{\pi}_J$. To do so we require the technical assumptions listed in this section. The assumptions will be verified in specific examples in Section 4.5 and 4.5.

The approach we use to analyse $\tilde{\pi}_J$, which is discussed after Theorem 20, is based on the decomposition $\tilde{\pi}_J(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi) = \tilde{\pi}_J(\mathrm{d}\psi)\tilde{\pi}_J(\mathrm{d}\boldsymbol{T} \mid \psi)$. The first set of assumptions contains standard regularity and identifiability conditions to study the marginal distribution $\tilde{\pi}_J(\mathrm{d}\psi)$. In particular, assumptions $(B1) - (B3)$ allow the application of Theorem 19 to the posterior distribution of ψ. Their applicability has been discussed in Section 4.3. We denote the marginal likelihood of the model, obtained by integrating out the group specific parameter θ, as

$$g(y \mid \psi) = \int_{\mathbb{R}^\ell} f(y \mid \theta) p(\theta \mid \psi) \, \mathrm{d}\theta \,, \tag{4.16}$$

and its Fisher Information matrix as

$$[\mathcal{I}(\psi)]_{d,d'} = E\left[\left\{\partial_{\psi_d} \log g(Y \mid \psi)\right\} \left\{\partial_{\psi_{d'}} \log g(Y \mid \psi)\right\}\right], \quad d, d' = 1, \dots, D.$$

We will assume the following:

$(B1)$ There exists $\psi^* \in \mathbb{R}^D$ such that $Y_j \overset{\text{iid}}{\sim} Q_{\psi^*}$ for $j = 1, 2, \dots,$ where Q_{ψ^*} admits density $g(y \mid \psi^*)$. Moreover the map $\psi \to g(\cdot \mid \psi)$ is one-to-one and the map $\psi \to \sqrt{g(x \mid \psi)}$ is continuously differentiable for every x. Finally, the prior density p_0 is continuous and strictly positive in a neighborhood of ψ^*.

$(B2)$ There exist a compact neighborhood Ψ of ψ^* and a sequence of tests $u_j : \mathbb{R}^{mJ} \to [0, 1]$ such that $\int_{\mathbb{R}^{mJ}} u_j (y_1, \dots, y_J) \prod_{j=1}^J g(y_j \mid \psi^*) \, \mathrm{d}y_{1:J} \to 0$ and
$\sup_{\psi \notin \Psi} \int_{\mathbb{R}^{mJ}} [1 - u_j (y_1, \dots, y_J)] \prod_{j=1}^J g(y_j \mid \psi) \, \mathrm{d}y_{1:J} \to 0$, as $J \to \infty$.

$(B3)$ The Fisher Information matrix $\mathcal{I}(\psi)$ is non-singular and continuous w.r.t. ψ.

The second set of regularity assumptions (B4)-(B6) are described and discussed in Appendix B. They deal with smoothness and regularity of the conditional distribution $\tilde{\pi}_J(\boldsymbol{T}|\psi)$ and they allow to derive a suitable conditional Central Limit Theorem in total variation for $\tilde{\pi}_J(\boldsymbol{T}|\psi)$ as $J \to \infty$.

We can now state the main result of this section. Below we denote the product measures associated to Q_{ψ^*} by $Q_{\psi^*}^{(J)}$ and $Q_{\psi^*}^{(\infty)}$.

Theorem 20. *Consider model* (4.13) *and the Gibbs sampler defined as in* (4.15), *with mixing times* $t_{mix}^{(J)}(\epsilon, M)$. *Then, under assumptions* $(B1)$-$(B6)$, *for every* $(M, \epsilon) \in [1, \infty) \times (0, 1)$ *there exists* $T(\psi^*, \epsilon, M) < \infty$ *such that*

$$Q_{\psi^*}^{(J)}\left(t_{mix}^{(J)}(\epsilon, M) \leq T(\psi^*, \epsilon, M)\right) \to 1,$$

as $J \to \infty$. *It follows that* $t_{mix}^{(J)}(\epsilon, M) = \mathcal{O}_P(1)$ *as* $J \to \infty$.

Remark 13. *Theorem 20 provides a formal proof of the linear in* J *cost for Gibbs samplers on hierarchical models. Indeed, it proves that a bounded (in* J*) number of iterations suffices to get a good mixing: assuming that the cost of a single iteration scales linearly with* J*, which is typically the case, this implies an overall computational cost of order* $\mathcal{O}_P(J)$*. Note that a single evaluation of the likelihood of* $(\boldsymbol{\theta}, \psi)$*, or the associated gradients, which is required at every iteration of usual gradient-based methods, yields a cost of the same order.*

Remark 14. *The conclusions of Theorem 20 are similar in spirit to those of (Kamatani, 2014, Thm.1). Also there the convergence of Gibbs Samplers targeting two-level hierarchical models is studied using tools from Bayesian asymptotics. The results therein, which deal with convergence of ergodic averages when the algorithm is started in stationarity, are quite different from ours, which deal with mixing times. Nonetheless they also support the idea that Gibbs samplers targeting two-level hierarchical models can exhibit* $\mathcal{O}_P(1)$ *convergence as* $J \to \infty$.

Posterior convergence lemmas for Theorem 20

The proof of Theorem 20 can be found in Appendix C. It relies on Lemma 44, which allows to focus on the two-blocks Gibbs sampler targeting $\tilde{\pi}_J(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi)$, and on Lemmas 45 and 46 below. These two lemmas imply that $\tilde{\pi}_J(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi)$ satisfies assumption $(A1)$ as $J \to \infty$ and that the associated limiting kernel is ergodic, thus allowing to apply Corollary 8.

In order to prove $(A1)$ for $\tilde{\pi}_J(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi) = \mathcal{L}(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi \mid Y_{1:J})$, we need to identify a suitable transformation of $(\boldsymbol{T}, \psi)$, denoted by $\left(\tilde{\boldsymbol{T}}, \tilde{\psi}\right)$. We define a one-to-one transformation of ψ as

$$\tilde{\psi} = \sqrt{J}(\psi - \psi^*) - \Delta_J, \quad \Delta_J = \frac{1}{\sqrt{J}} \sum_{j=1}^{J} \mathcal{I}^{-1}(\psi^*) \nabla \log g(Y_j \mid \psi^*). \tag{4.17}$$

The asymptotic distribution of $\tilde{\psi}$ follows directly through Theorem 19, as summarized in the next lemma.

Lemma 45. *Define* $\tilde{\psi}$ *as in* (4.17). *Under assumptions* $(B1) - (B3)$ *it holds*

$$\left\| \mathcal{L}(d\tilde{\psi} \mid Y_{1:J}) - N\left(\mathbf{0}, \mathcal{I}^{-1}(\psi^*)\right) \right\|_{TV} \to 0,$$

as $J \to \infty$, *in* $Q_{\psi^*}^{(\infty)}$*-probability.*

Let $M^{(1)}(\psi \mid y) = \left(M_1^{(1)}(\psi \mid y), \ldots, M_S^{(1)}(\psi \mid y)\right) \in \mathbb{R}^S$ with $M_s^{(1)}(\psi \mid y) = E\left[T_s(\theta_j) \mid Y_j = y, \psi\right]$ and

$$[C(\psi)]_{s,d} = E_{Y_j}\left[\partial_{\psi_d} M_s^{(1)}(\psi \mid Y_j)\right], \quad [V(\psi)]_{s,s'} = E_{Y_j}\left[\mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi\right)\right], \tag{4.18}$$

with $s, s' = 1, \dots S$ and $d = 1, \dots, D$. We use the notation $E_{Y_j}[\cdot]$ for expectations with respect to the law of Y_j as defined in $(B1)$. Then we define a one-to-one transformation of $\boldsymbol{T}$ as

$$\tilde{\boldsymbol{T}} = \frac{1}{\sqrt{J}} \sum_{j=1}^{J} \left[T(\theta_j) - M^{(1)}\left(\psi^* \mid Y_j\right) \right] - C(\psi^*)\Delta_J, \tag{4.19}$$

with $C(\psi^*)$ defined in (38). The next lemma proves the required asymptotic normality of $\tilde{\boldsymbol{T}}$, conditional to $\tilde{\psi}$.

Lemma 46. *Let $\tilde{\boldsymbol{T}}$ be as in* (4.19)*. Under assumptions* $(B1)$*-*$(B6)$ *for every $\tilde{\psi}$ it holds*

$$\left\| \mathcal{L}(d\tilde{\boldsymbol{T}} \mid Y_{1:J}, \tilde{\psi}) - N\left(C(\psi^*)\tilde{\psi}, V(\psi^*)\right) \right\|_{TV} \to 0,$$

as $J \to \infty$, for $Q_{\psi^*}^{(\infty)}$-almost every $(Y_1, Y_2, \dots)$.

Lemma $C.18$ in Section $C.14$ of Appendix C combines Lemmas 45 and 46 to prove that $\mathcal{L}(\mathrm{d}\tilde{\boldsymbol{T}}, \tilde{\psi} \mid Y_{1:J})$ converges in total variation to a multivariate Gaussian vector with non singular covariance matrix, which allows to apply Corollary 8 as desired.

Remark 15. *The definition of $\tilde{\boldsymbol{T}}$ and Lemma 46 are an important part of the proof of Theorem 20. Lemma 46 relies on the fact that, conditional to $\tilde{\psi}$ and $Y_{1:J}$, $\boldsymbol{T}$ is a sum of independent (but not identically distributed) terms. The proof of convergence in total variation requires more than the usual tools from Lindeberg-Feller Central Limit Theorem, as discussed in Appendix B after assumptions* $(B5)$ *and* $(B6)$*.*

Analysis of the limiting chain

As a byproduct of the proof of Theorem 20, it is possible to characterize the limiting distribution of the rescaled vector $\left(\tilde{\boldsymbol{T}}, \tilde{\psi}\right)$, as the next proposition shows.

Proposition 23. *Consider the same assumptions of Theorem 20. Then*

$$\left\| \mathcal{L}(d\tilde{\boldsymbol{T}}, d\tilde{\psi} \mid Y_{1:J}) - N\left(\mathbf{0}, \Sigma\right) \right\|_{TV} \to 0,$$

as $J \to \infty$, in $Q_{\psi^*}^{(\infty)}$-probability, where

$$\Sigma = \begin{bmatrix} V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*) & C(\psi^*)\mathcal{I}^{-1}(\psi^*) \\ \mathcal{I}^{-1}(\psi^*)C^\top(\psi^*) & \mathcal{I}^{-1}(\psi^*) \end{bmatrix} \tag{4.20}$$

with $C(\psi^*)$ and $V(\psi^*)$ defined in (38)*.*

The expression for the limiting covariance in (4.20) can be used to investigate the convergence properties of the limiting Gibbs sampler, since the spectral gap is explicitly computable from that. We can then apply Corollary 9 and obtain the following result.

Corollary 10. *Under the assumptions of Theorem 20, for every $(M, \epsilon) \in [1, \infty) \times (0, 1)$, we have $Q^{(J)}_{\psi^*}\left(t^{(J)}_{mix}(\epsilon, M) \leq T(\psi^*, \epsilon, M)\right) \to 1$ as $J \to \infty$, with*

$$\begin{aligned} T(\psi^*, \epsilon, M) &= 1 + \frac{\log(M/2) - \log(\epsilon)}{-\log(1 - \gamma(\psi^*))}, \\ \gamma(\psi^*) &= \min\left\{\frac{1}{1+\lambda_i} : \lambda_i \textit{ eigenvalue of } V^{-1}(\psi^*)C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}. \end{aligned}$$

Thus, once the limiting distribution is obtained, an upper bound on the mixing times can be derived by computing the eigenvalues of a $S \times S$ matrix. As an application, the next corollary provides the value of γ when $S = D = 1$.

Corollary 11. *Consider the same setting of Corollary 10, with $S = D = 1$. Then we have*

$$\gamma(\psi^*) = \frac{Var_{Y_j}\left(E\left[T(\theta_j) \mid \psi^*, Y_j\right]\right)}{Var\left(T(\theta_j) \mid \psi^*\right)}. \tag{4.21}$$

By the law of total variance, we have that $\gamma(\psi^*) \to 0$ if and only if

$$\frac{\mathrm{Var}_{Y_j}\left(E\left[T(\theta_j) \mid \psi^*, Y_j\right]\right)}{E\left[\mathrm{Var}_{Y_j}\left(T(\theta_j) \mid \psi^*, Y_j\right)\right]} \to 0,$$

i.e., loosely speaking, when the data Y_j yield little information about $T(\theta_j)$ and therefore about ψ. This phenomenon arises since model (4.13) is an example of centered parametrization, see e.g. Gelfand et al. (1995); Papaspiliopoulos et al. (2003, 2007a). The formula in (4.21) resembles the definition of the so-called Bayesian fraction of missing information (Liu, 1994), with the notable difference of not involving an infimum over a set of test functions.

4.5 Examples

In this section various examples, which differ by the choice of likelihoods and priors, are discussed.

Hierarchical normal model

Consider the following hierarchical specification:

$$\begin{aligned} Y_{j,i} \mid \theta_j &\sim N\left(\theta_j, \tau_0^{-1}\right), \quad i = 1, \dots, m, \ j = 1, \dots, J \\ \theta_j \mid \mu, \tau_1 &\overset{\text{iid}}{\sim} N(\mu, \tau_1^{-1}), \qquad\qquad\qquad j = 1, \dots, J \\ (\mu, \tau_1) &\sim p_0(\cdot). \end{aligned} \tag{4.22}$$

where (μ, τ_1) are unknown hyperparameters. In this section we assume τ_0 to be fixed and known, see Section 4.5 for the case with τ_0 unknown. The prior p_0 can be any distribution satisfying the assumptions stated in Proposition 24 below. It can be seen that (4.22) is a particular case of model (4.13), with $f(Y_j \mid \theta_j) = \prod_{i=1}^m N(Y_{j,i} \mid \theta_j, \tau_0^{-1})$, $p(\cdot \mid \mu, \tau_1) = N(\mu, \tau_1^{-1})$. The marginal likelihood of Y_j conditional to (μ, τ_1, τ_0) is given by

$$g(y \mid \mu, \tau_1, \tau_0) = N\left(y \mid \mu, \tau_0^{-1} I + \tau_1^{-1} \mathbb{H}\right) \qquad y \in \mathbb{R}^m, \tag{4.23}$$

where I is the $m \times m$ identity matrix and $\mathbb{H}$ is the $m \times m$ matrix of ones.

We consider three Gibbs sampler specifications, which vary depending on which parameters are unknown and treated as random and which blocking rules are used. First, when τ_1 is fixed, we define P_1 as the transition kernel of the Gibbs sampler that targets $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\mu \mid Y_{1:J}\right)$ by alternating updates from $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta} \mid \mu, Y_{1:J}\right)$ and $\mathcal{L}\left(\mathrm{d}\mu \mid \boldsymbol{\theta}, Y_{1:J}\right)$. If instead μ and τ_1 are unknown, we define P_2 and P_3 as the transition kernels of the two Gibbs samplers targeting $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\mu, \mathrm{d}\tau_1 \mid Y_{1:J}\right)$ by alternating updates from $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\mu \mid \tau_1, Y_{1:J}\right)$ and $\mathcal{L}\left(\mathrm{d}\tau_1 \mid \boldsymbol{\theta}, \mu, Y_{1:J}\right)$ for P_2; and $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta} \mid \tau_1, Y_{1:J}\right)$, $\mathcal{L}\left(\mathrm{d}\mu \mid \boldsymbol{\theta}, \tau_1, Y_{1:J}\right)$, $\mathcal{L}\left(\mathrm{d}\tau_1 \mid \boldsymbol{\theta}, \mu, Y_{1:J}\right)$ for P_3. In the following we will show that the asymptotic behaviour of P_2 and P_3 is essentially the same.

It is possible to prove that P_1 falls directly in the setting of Theorem 20, with $T(\theta_j) = \theta_j$ for P_1. Even if P_2 and P_3 are not exactly particular cases of the general theorem, since different update schemes are considered, it turns out that they can be studied with the same tools introduced in the previous section, with $T(\theta_j) = \left(\theta_j, (\theta_j - \mu^*)^2\right)$.

The next proposition shows that the settings introduced above lead to well-behaved asymptotic regimes. Here $t_{mix,l}^{(J)}(\epsilon, M)$ denotes the mixing times of the Gibbs sampler defined by P_l with $l \in \{1, 2, 3\}$.

Proposition 24. *Let $Y_j \overset{iid}{\sim} Q_{\psi^*}$, with Q_{ψ^*} admitting density $g(y \mid \psi^*)$ as in* (4.23)*, where $\psi^* = (\mu^*, \tau_1^*, \tau_0^*)$, and consider model* (4.22) *with $\tau_0 = \tau_0^*$. Consider the Gibbs sampler with operator P_l, with $l \in \{1, 2, 3\}$, and let the prior density p_0 be continuous and strictly positive in a neighborhood of μ^* when $l = 1$ and (μ^*, τ_1^*) when $l \in \{2, 3\}$. Finally, when $l = 1$ let $\tau_1 = \tau_1^*$. Then for every $(M, \epsilon) \in [1, \infty) \times (0, 1)$ there exists $T_l\left(\psi^*, \epsilon, M\right) < \infty$ such that*

$$Q_{\psi^*}^{(J)}\left(t_{mix,l}^{(J)}(\epsilon, M) \leq T_l\left(\psi^*, \epsilon, M\right)\right) \to 1 \qquad \text{as } J \to \infty,\ l = 1, 2, 3\,. \tag{4.24}$$

Under model (4.22), the matrices in Corollary 10 can be explicitly computed, leading to the following result.

Corollary 12. *Under the same assumptions and notation of Proposition 24, for every $(M, \epsilon) \in [1, \infty) \times (0, 1)$,* (4.24) *holds with*

$$T_l\left(\psi^*, \epsilon, M\right) = 1 + \frac{\log(M/2) - \log(\epsilon)}{-\log\left(1 - \gamma_l(\psi^*)\right)}, \qquad l = 1, 2, 3\,,$$

where

$$\gamma_1(\psi^*) = \left(1 + \frac{\tau_1^*}{m\tau_0^*}\right)^{-1} \quad \text{and} \quad \gamma_2(\psi^*) = \gamma_3(\psi^*) = \gamma_1(\psi^*)^2\,. \tag{4.25}$$

The expressions for the asymptotic gaps in (4.25) are insightful in many ways. First, μ^* does not appear in any of the spectral gaps, meaning that the limiting value of the mean parameter seems not to play a role in the asymptotic behaviour of the Gibbs sampler. Moreover, the gaps are a function of the ratio $(m\tau_0^*)^{-1}\tau_1^*$, that is the ratio of the prior and likelihood precisions, respectively. In particular the gaps converge to 0, i.e. the upper bound on the mixing times diverges, if and only if $(m\tau_0^*)^{-1}\tau_1^* \to \infty$, which happens when the prior is increasingly more informative than the data. As discussed after Corollary 11, such phenomenon arises since all the three formulations are an example of centered parametrization (Gelfand et al., 1995; Papaspiliopoulos et al., 2003). On the contrary, the gaps converge to 1, i.e. asymptotically a single iteration suffices, if and only if $(m\tau_0^*)^{-1}\tau_1^* \to 0$.

When τ_1 is fixed and $p_0(\mu)$ is Gaussian, then $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\mu \mid Y_{1:J}\right)$ is a multivariate Gaussian and P_1 is amenable to finite-sample analysis. In fact, the expression for $\gamma_1(\psi^*)$ appeared previously in the literature, see e.g. Papaspiliopoulos et al. (2003). The result in Corollary 12 is, however, different since it is asymptotic and it applies also to general priors.

On the contrary, a finite-sample analysis of P_2 are P_3 is hard even when $p_0(\mu)$ is Gaussian (see e.g. Jin and Hobert (2022); Qin and Hobert (2022); Yang and Rosenthal (2022)) and $\gamma_2(\psi^*)$ and $\gamma_3(\psi^*)$ did not appear previously in the literature, to the best of our knowledge. It is interesting that, regardless of the value of $(m, \mu^*, \tau_1^*, \tau_0^*)$, including the random precision parameter, when moving from P_1 to either P_2 or P_3, always slows down the sampler (asymptotically), since $\gamma_1(\psi^*) > \gamma_i(\psi^*)$ for $i = 2, 3$, and that the two blocking rules of P_2 and P_3 are asymptotically equivalent in terms of mixing times, since $\gamma_2(\psi^*) = \gamma_3(\psi^*)$.

Models with binary and categorical data

Let now $f(y \mid \theta)$ be a probability mass function, whose point masses are denoted by $y_0, \dots, y_m$, with $m < \infty$, such that for every $\theta \in \mathbb{R}^K$ we have

$$\sum_{r=0}^{m} f(y_r \mid \theta) = 1, \quad f(y_r \mid \theta) > 0, \quad r = 0, \dots, m. \tag{4.26}$$

The assumption in (4.26) is mild and holds for most likelihoods usually employed with categorical data, e.g. multinomial logit and probit. We focus on hierarchical models with normal priors, i.e.

$$Y_j \mid \theta_j \sim f(Y_j \mid \theta_j)\,, \quad \theta_1, \dots, \theta_J \mid \mu, \tau \overset{\text{iid}}{\sim} N(\mu, \tau^{-1})\,, \quad (\mu, \tau) \sim p_0(\cdot)\,. \tag{4.27}$$

For example the case $f(y \mid \theta) = \binom{m}{y}\frac{e^{y\theta}}{(1+e^{\theta})^m}$, with $y = 0, \dots, m$, corresponds to the logistic hierarchical model with Gaussian random effects. The prior p_0 can be any distribution satisfying the assumptions stated in Proposition 25 below. We define P as the transition kernel of the Gibbs sampler that targets $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\mu, \mathrm{d}\tau \mid Y_{1:J}\right)$ by alternating updates from $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta} \mid \mu, \tau, Y_{1:J}\right)$ and $\mathcal{L}\left(\mathrm{d}\mu, \mathrm{d}\tau \mid \boldsymbol{\theta}, Y_{1:J}\right)$. This is a particular case of the setting of Theorem 20, with $\psi = (\mu, \tau)$ and $T(\theta_j) = (\theta_j, \theta_j^2)$. Notice that usually $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta} \mid \mu, \tau, Y_{1:J}\right)$ is not known in closed form (with the notable exception of the probit case, see Durante (2019)), but nonetheless exact sampling is often feasible through adaptive rejection sampling (see e.g. Gilks and Wild (1992)) since each θ_j is one dimensional. The marginal likelihood is given by

$$g(y \mid \psi) = \int_{\mathbb{R}} f(y \mid \theta) N\left(\theta \mid \mu, \tau^{-1}\right) \mathrm{d}\theta. \tag{4.28}$$

The next lemma shows that assumptions $(B4)$-$(B6)$ follow directly from (4.27).

Lemma 47. *Consider model* (4.27) *and let $Y_j \overset{iid}{\sim} Q_{\psi^*}$, with Q_{ψ^*} admitting density $g(y \mid \psi^*)$ as in* (4.28)*, with $\psi^* = (\mu^*, \tau^*)$. Then assumptions* $(B4)$-$(B6)$ *are satisfied.*

Thus, in order to apply Theorem 20, it suffices to prove assumptions $(B2)$ and $(B3)$, i.e. that the parameters ψ are identifiable with non singular Fisher Information matrix. Therefore, as formalized in the next proposition, standard identifiability conditions (which are also necessary to consistently estimate ψ) are sufficient to prove boundedness of the mixing times.

Proposition 25. *Consider model* (4.27) *and let* $Y_j \overset{iid}{\sim} Q_{\psi^*}$*, with* Q_{ψ^*} *admitting density* $g(y \mid \psi^*)$ *as in* (4.28)*, where* $\psi^* = (\mu^*, \tau^*)$*. Consider the Gibbs sampler with operator* P *and let* p_0 *be continuous and strictly positive in a neighborhood of* ψ^**. Let the map* $\psi \to g(\cdot \mid \psi)$ *be one-to-one, with non singular and continuous* $\mathcal{I}(\psi)$*. Finally, assume tests as in* $(B2)$ *exist. Then for every* $(M, \epsilon) \in [1, \infty) \times (0, 1)$ *there exists* $T(\psi^*, \epsilon, M) < \infty$ *such that*

$$Q^{(J)}_{\psi^*}\left(t^{(J)}_{mix}(\epsilon, M) \leq T(\psi^*, \epsilon, M)\right) \to 1 \qquad \text{as } J \to \infty\,.$$

Remark 16. *In most cases* $m \geq 2$ *is required to avoid the pair* (μ, τ) *being not identifiable and the associated Fisher Information matrix being singular. For example Lemma* C.35 *in Section* C.23 *of Appendix C shows that with the logit link* $\mathcal{I}(\psi)$ *is singular if and only if* $m = 1$.

As already discussed in the Section 4.1, the results of Proposition 25 are illustrated on simulated data in Figure 4.1. Since mixing times are very hard to approximate numerically in high-dimensions, we employ the Integrated Autocorrelation Times (IATs) as an empirical measure of convergence time. The IAT associated to a π-invariant Markov chain $X = \{X^{(t)}\}_{t\geq 1}$ and a test function $f \in L^2(\pi)$ is defined as

$$\mathrm{IAT}(f) = 1 + 2\sum_{t=2}^{\infty} \mathrm{Corr}\left(f(X^{(1)}), f(X^{(t)})\right)\,. \tag{4.29}$$

Loosely speaking, IAT(f) is the number of MCMC samples that is equivalent to a single independent sample in terms of estimation of $\int f(x)\pi(dx)$, thus the higher IAT the slower the convergence. When dealing with hierarchical models as in (4.27), we compute the maximum IAT over all the parameters (both global and group specific). We estimate the IAT with the ratio of the number of iterations and the effective sample size, as described in Gong and Flegal (2015), with the effective sample size computed with the R package *mcmcse* (Flegal et al., 2021). For a review of different methods to estimate the IATs, see Thompson (2010). In Figure 4.1 we plot the quantiles of the IATs as a function of the number of groups for the Gibbs sampler, implemented using adaptive rejection sampling (Gilks and Wild, 1992) for the exact updates of local parameters with full conditionals $\mathcal{L}\left(\mathrm{d}\theta_j \mid \mu, \tau, Y_{1:J}\right)$. As expected by Proposition 25, the IATs do not diverge as J increases for both values of m under consideration. Note that variability decreases as J increases and the posterior gets closer to its asymptotic limit.

Corollary 13. *Consider the same setting of Proposition 25. For every* $(M, \epsilon) \in [1, \infty) \times (0, 1)$ *define*

$$T(\psi^*, \epsilon, M) = 1 + \frac{\log(M/2) - \log(\epsilon)}{-\log\left(1 - \gamma(\psi^*)\right)},$$

for $\gamma(\psi^*) \in (0, 1)$ *as in Corollary 10. Then*

$$Q^{(J)}_{\psi^*}\left(t^{(J)}_{mix}(\epsilon, M) \leq T(\psi^*, \epsilon, M)\right) \to 1 \qquad \text{as } J \to \infty\,.$$

The study of the limiting spectral properties, i.e. of $\gamma(\psi^*)$, can be useful to predict under which scenarios the Gibbs sampler will perform well or not for large J. We illustrate this by considering model (4.27) with logit link and known τ set to 1. In this setting, where μ is

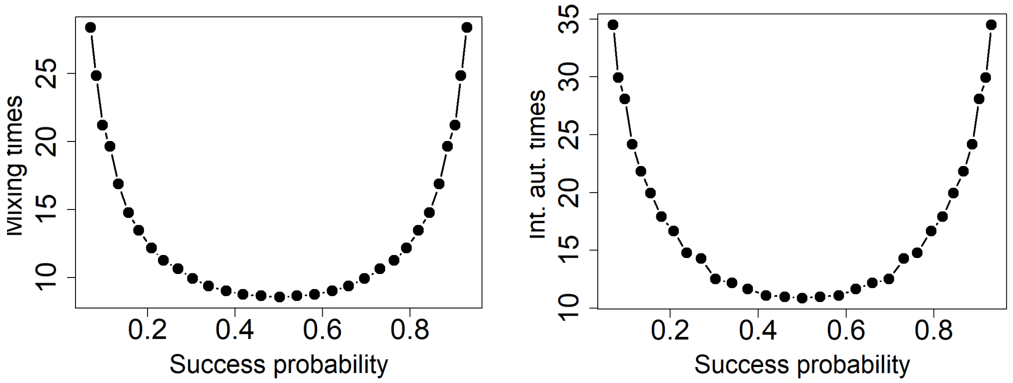


Figure 4.2: Left: upper bounds on mixing times for model (4.27) with τ known, where $\tau^* = 1$, $\mu^* \in (-3, 3)$, $m = 1$, $M = 2$ and $\epsilon = 0.2$. A priori $\mu \sim N\left(0, 10^3\right)$. Right: median IATs with $J = 2000$.

the only global parameter, the value of $\gamma(\psi^*)$ can be computed as in (4.21) through simple one-dimensional numerical integration. In Figure 4.2 we compare the resulting mixing time upper bound, $T\left(\psi^*, \epsilon, M\right)$, with the numerical estimates of IATs defined in (4.29), obtained by running a long MCMC chain with a moderately large value of J. We compare such quantities for different values of the true success probability induced by μ^*, i.e. $\int_{\mathbb{R}} f(1 \mid \theta) N\left(\theta \mid \mu^*, 1\right) \, \mathrm{d}\theta$. Both theoretical and empirical measures of convergence highlight that the performances of the Gibbs sampler deteriorate when the problem is not balanced: such conclusion is coherent with the findings in Johndrow et al. (2019), that considers an asymptotic regime with increasing imbalancedness.

Different graphical models structure

In the previous subsections we have studied applications of Theorem 20 for some specification of the hierarchical model in (4.13). These correspond to the graphical models in the leftmost panel of Figure 4.3. While this structure is very common in Bayesian modeling and it constitutes our main motivating application, the techniques we developed - and in particular the dimensionality reduction and posterior asymptotic approach - can be applied to different classes of models, including other widely used ones. Here we provide two examples, the first is a relatively direct extension of the model in (4.13) with the addition of parameters in the likelihood, the second is a more different setting of Gaussian Process regression where the latent parameters are not independent. See respectively the center and rightmost panels in Figure 4.3 for the resulting graphical models. More generally, we expect our methodology to be potentially useful to analyse samplers for models that feature a fixed set of hyperparameters ψ, conditional to which a growing set of parameters or latent variables is tractable enough for posterior sampling.

Likelihood parameters

Consider again the hierarchical normal model

$$Y_{j,i} \mid \theta_j, \tau_0 \sim N\left(\theta_j, \tau_0^{-1}\right), \quad \theta_j \mid \mu, \tau_1 \overset{\text{iid}}{\sim} N(\mu, \tau_1^{-1}), \quad (\mu, \tau_1, \tau_0) \sim p_0(\cdot), \tag{4.30}$$

with $i = 1, \dots, m$ and $j = 1, \dots, J$. The unknown parameters are now given by the triplet $\psi = (\mu, \tau_1, \tau_0)$. We denote with P the transition kernel of the Gibbs sampler targeting $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\mu, \mathrm{d}\tau_1, \mathrm{d}\tau_0 \mid Y_{1:J}\right)$ by alternating updates from $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\mu \mid \tau_1, \tau_0, Y_{1:J}\right)$ and $\mathcal{L}\left(\mathrm{d}\tau_1, \mathrm{d}\tau_0 \mid \boldsymbol{\theta}, \mu, Y_{1:J}\right)$. This cannot be

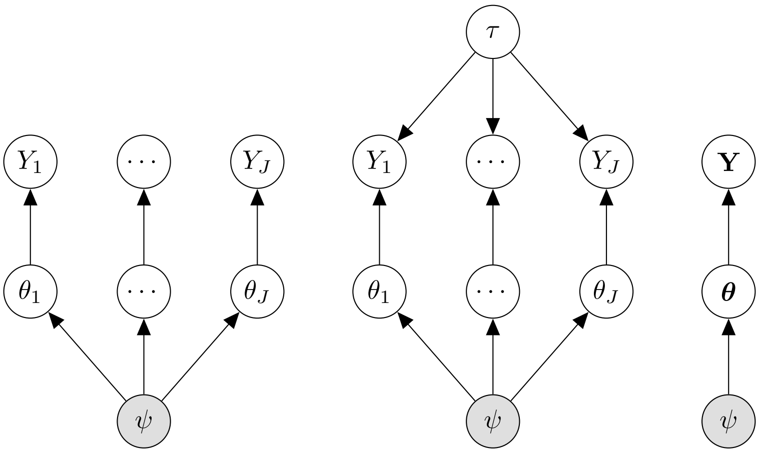


Figure 4.3: Graphical models of different hierarchical structures. Left: one level nested model as in Theorem 20. Center: hyperparameters specifying the likelihood. Right: dependent latent parameters.

seen as a specific case of Theorem 20 with $\psi = (\mu, \tau_1, \tau_0)$, since τ_0 is a parameter of the likelihood f and therefore there is no conditional independence between Y_j and ψ, given θ_j. However, an approach similar to the one of the previous section can be employed. In particular, a result analogous to Lemma 44 can be derived, with $T(\theta_j) = \left(\left(\theta_j - \bar{Y}_j\right)^2, \left(\theta_j - \mu\right)^2\right)$ playing the role of the sufficient statistics and $\bar{Y}_j = \frac{1}{m}\sum_{i=1}^m Y_{j,i}$. It is interesting to notice that T in this case depends also on the data $Y_{1:J}$, exactly because the group specific parameters $\boldsymbol{\theta}$ do not contain all the information regarding ψ. The next proposition shows that also this specification leads to a well-behaved asymptotic regime.

Proposition 26. *Consider model* (4.30) *with* $m \geq 2$ *and let* $Y_j \overset{iid}{\sim} Q_{\psi^*}$*, with* Q_{ψ^*} *admitting density* $g(y \mid \psi^*)$ *as in* (4.23)*, where* $\psi^* = (\mu^*, \tau_1^*, \tau_0^*)$*. Consider the Gibbs sampler with operator* P *and let the prior density* p_0 *be a continuous and strictly positive in a neighborhood of* ψ^**. Then for every* $(M, \epsilon) \in [1, \infty) \times (0, 1)$ *there exists* $T(\psi^*, \epsilon, M) < \infty$ *such that*

$$Q_{\psi^*}^{(J)}\left(t_{mix}^{(J)}(\epsilon, M) \leq T(\psi^*, \epsilon, M)\right) \to 1 \qquad \textit{as } J \to \infty\,. \tag{4.31}$$

An explicit value for $T(\psi^*, \epsilon, M)$ can be found through Corollary 9, as shown in the next corollary.

Corollary 14. *Consider the same setting of Proposition 26. Then, for every* $(M, \epsilon) \in [1, \infty) \times (0, 1)$*,* (4.31) *holds with*

$$T(\psi^*, \epsilon, M) = 1 + \frac{\log(M/2) - \log(\epsilon)}{-\log\left(1 - \gamma(\psi^*)\right)}\,,$$

where

$$\gamma(\psi^*) = \left(1 + \frac{1}{m-1}\left(1 - \frac{\tau_1^*}{m\tau_0^*}\right)^2 + \left(\frac{\tau_1^*}{m\tau_0^*}\right)^2\right)^{-1}.$$

Remark 17. *The assumption* $m \geq 2$ *cannot be relaxed: indeed, if a single observation per group is available, the pair* (τ_1, τ_0) *is not identifiable and the Fisher Information matrix is singular.*

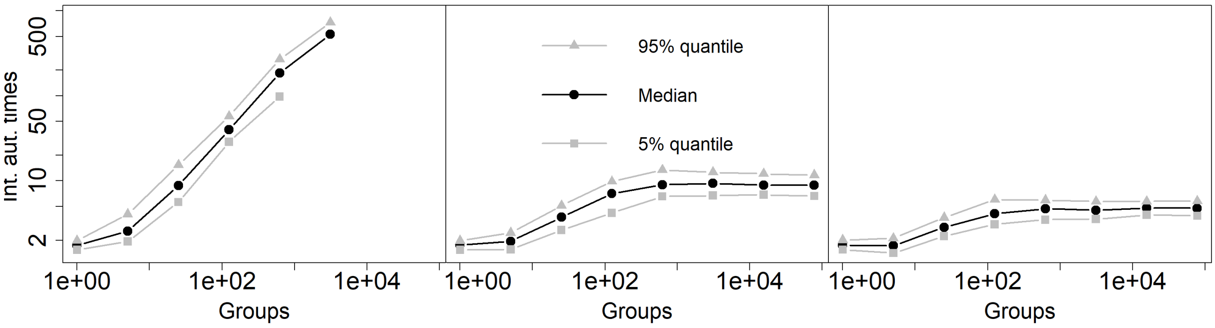


Figure 4.4: Quantiles of the integrated autocorrelations times (on log-scale) for model (4.30) with $\mu^* = 4$, $\tau_0^* = 1$ and $\tau_1^* = 3$. A priori $(\tau_0, \tau_1) \overset{\text{i.i.d.}}{\sim}$ Gamma$(1, 1)$ and $p_0(\mu) \propto 1$. Top left: $m = 1$ (last points not plotted due to numerical instability). Center: $m = 3$. Top right: $m = 5$.

For an empirical illustration of the issues arising in this context, see the top left panel in Figure 4.4 or Section 6.2 *of Rajaratnam and Sparks (2015).*

Unlike the case of Corollary 12, in this setting the limiting gap does not depend on m only through the ratio of prior and likelihood precisions, but also directly on its value. Loosely speaking, a higher value of m allows to better recover the relation between τ_0 and τ_1.

The results of Proposition 26 and Corollary 14 are illustrated on simulated data in Figure 4.4, which depicts the Integrated Autocorrelations Times (IATs) as defined in (4.29). When the model is not identifiable, i.e. $m = 1$ (top left panel), the IATs diverge with the number of groups, while with $m = 3$ and $m = 5$ they stabilize as J increases. Differently from the binomial setting of Figure 4.4, the IATs grow for small values of J before the asymptotic regime kicks in.

Gaussian processes

We now consider the popular setting where the groups are identified by a continuous covariate (e.g. location) and group specific parameters are modeled through a Gaussian process. It turns out that the main arguments of the paper, namely dimensionality reduction and impact of posterior asymptotic characterization, can be applied also in this context. This section, compared to the previous ones, aims to provide a proof of concept rather than a detailed analysis, e.g. we directly assume limiting statements on the posterior distributions of interest. Nonetheless we find it useful to show how widely our methodology could be applied and illustrate interesting directions of ongoing work.

Assume to observe n data points $Y(s_i)$ with $i = 1, \dots, n$, at a set of locations $(s_1, \dots, s_n)$, together with input variables or covariates $x(s_i) \in \mathbb{R}$. We consider Gaussian Process regression models of the form

$$\begin{aligned} Y(s_i) \mid \boldsymbol{\beta} &\sim f(\cdot \mid \beta(s_i), x(s_i)), \quad i = 1, \dots, n \\ \boldsymbol{\beta}^{(n)} \mid \psi &\sim N(\theta \mathbf{1}, \tau_\beta^{-1} R^{(n)}) \\ \psi &\sim p_0(\cdot). \end{aligned} \tag{4.32}$$

where $\boldsymbol{\beta} = (\beta(s_1), \dots, \beta(s_n))^\top$ is a Gaussian Process (GP) observed at $(s_1, \dots, s_n)$ and f is a density function with respect to a suitable dominating measure. Here $\mathbf{1}_n = (1, \dots, 1)^\top$ is an n-dimensional vector and $R^{(n)} = (R_{ij})_{i,j=1,\dots,n}$ is a $n \times n$ correlation matrix, with $R_{ij} =$

Corr $(\beta(s_i), \beta(s_j))$, defined through a suitable kernel function, that we assume to be fixed and known. Typically, strength of correlation among coefficients at different locations depends on their distance, with R_{ij} defined e.g. through a kernel of the Matérn family (see e.g. Section 4.2.1 in Williams and Rasmussen (2006)). In this Section we focus on a single real covariate for notational convenience, but everything could be restated on a general p-dimensional space with little effort: direct analogues of the next lemma and corollaries similarly follow. We first consider cases where the likelihood function has no specific hyper-parameters, such as in the common binary case where $Y(s_j) \mid \boldsymbol{\beta} \sim \text{Bernoulli}(\sigma(\beta(s_j)x(s_j)))$, with σ logistic link function and $Y(s_j) \in \{0,1\}$.

Let P_n be the kernel of the Gibbs sampler which targets $\pi_n(\mathrm{d}\boldsymbol{\beta}, \mathrm{d}\theta, \mathrm{d}\tau_\beta) = \mathcal{L}\left(\mathrm{d}\boldsymbol{\beta}, \mathrm{d}\theta, \mathrm{d}\tau_\beta \mid Y^{(n)}\right)$, by sequentially performing updates from the full conditionals of $\boldsymbol{\beta}$, θ and τ_β. Despite the different graphical model structure, the analysis of mixing times of P_n as $n \to \infty$ can be approached with the techniques we developed above, regardless of the specific likelihood used in (4.32). The first step is to perform a dimensionality reduction analogous to the one in Section 4.4. Define $\psi = (\theta, \tau_\beta)$ and $\boldsymbol{T}(\boldsymbol{\beta}) = \left(T_\theta, T_{\tau_\beta}\right)$, where $T_\theta = \mathbf{1}^\top R^{-1}\boldsymbol{\beta}$, $T_{\tau_\beta} = \boldsymbol{\beta}^\top R^{-1}\boldsymbol{\beta}$, which play the same role of global parameters and sufficient statistics in Lemma 44. Indeed it holds $\mathcal{L}\left(\mathrm{d}\psi \mid \boldsymbol{\beta}, Y^{(n)}) = \mathcal{L}(\mathrm{d}\psi \mid \boldsymbol{T}(\boldsymbol{\beta}), Y^{(n)}\right)$ and we can provide an analogue of Lemma 44 for model (4.32).

Lemma 48. *Let π_n and P_n be defined as above for model* (4.32). *Let $\hat{P}_n$ be the transition kernel of Gibbs sampler targeting $\hat{\pi}_n(d\boldsymbol{T}, d\theta, d\tau_\beta) = \mathcal{L}\left(d\boldsymbol{T}, d\theta, d\tau_\beta \mid Y^{(n)}\right)$ which sequentially performs updates from the full conditionals of $\boldsymbol{T}$, θ and τ_β. Let $(\boldsymbol{T}^{(t)}, d\theta^{(t)}, d\tau_\beta^{(t)})_{t\geq 1}$ be the stochastic process obtained as a time-wise transformation of $(\boldsymbol{\beta}^{(t)}, d\theta^{(t)}, d\tau_\beta^{(t)})_{t\geq 1}$. Then $(\boldsymbol{T}^{(t)}, d\theta^{(t)}, d\tau_\beta^{(t)})_{t\geq 1}$ is a Markov chain, its transition kernel coincides with $\hat{P}_n$, and its mixing times $\hat{t}_{mix}^{(n)}$ satisfy*

$$\sup_{\mu \in \mathcal{N}(\pi_n, M)} t_{mix}^{(n)}(\epsilon, \mu) = \sup_{\nu \in \mathcal{N}(\hat{\pi}_n, M)} \hat{t}_{mix}^{(n)}(\epsilon, \nu) \qquad M \geq 1\,.$$

Also, provided a rescaled version of $(\boldsymbol{T}, \theta, \tau_\beta)$ converges to a suitable limit conditional on the data, the mixing times are bounded with respect to the number of observations.

Corollary 15. *Under model* (4.32), *let $\hat{\pi}_n$ satisfy assumption* (*A*1) *for a given data generating process $Y^{(n)} \sim Q^{(n)}$, with limiting distribution $\tilde{\pi}$. If $(M, \epsilon) \in [1, \infty) \times (0, 1)$ is such that $\tilde{t}_{mix}(\epsilon, M) < \infty$, then it holds*

$$Q^{(n)}\left(t_{mix}^{(n)}(\epsilon, M) \leq \tilde{t}_{mix}(\epsilon, M)\right) \to 1 \qquad \text{as } n \to \infty\,. \tag{4.33}$$

In some cases the likelihood contains some unknown parameters that are also included in the Bayesian model. A common example is the likelihood precision τ_ϵ in normal linear models with spatially varying regression coefficients (see e.g. Gelfand et al. (2003) or Section 2 in Williams and Rasmussen (2006)), where

$$Y(s_i) \mid \boldsymbol{\beta} \sim N(\beta(s_i)x(s_i), \tau_\epsilon^{-1}), \qquad i = 1, \ldots, n. \tag{4.34}$$

Let P_n be the Gibbs sampler kernel targeting $\pi_n(\mathrm{d}\boldsymbol{\beta}, \mathrm{d}\theta, \mathrm{d}\tau_\beta, \mathrm{d}\tau_\epsilon) = \mathcal{L}\left(\mathrm{d}\boldsymbol{\beta}, \mathrm{d}\theta, \mathrm{d}\tau_\beta, \mathrm{d}\tau_\epsilon \mid Y^{(n)}\right)$, by sequentially performing updates from the full conditionals of $\boldsymbol{\beta}$, θ, τ_β and τ_ϵ. Analogously to Section 4.5, the results of Lemma 48 and Corollary 15 extend to this context with $\psi = (\theta, \tau_\beta, \tau_\epsilon)$

and $\boldsymbol{T}$ defined as $\boldsymbol{T} = \left(T_\theta, T_{\tau_\beta}, T_{\tau_\epsilon}\right)$, where $T_{\tau_\epsilon} = \left(Y^{(n)} - D\boldsymbol{\beta}\right)^\top \left(Y^{(n)} - D\boldsymbol{\beta}\right)$ and D is the $n \times n$ diagonal matrix with values $(x(s_1), \dots, x(s_n))$. This is summarized in the next corollary.

Corollary 16. *Under model* (4.32) *with likelihood as in* (4.34)*, assume the conditions of Corollary 15 are satisfied with* $\psi = (\theta, \tau_\beta, \tau_\epsilon)$ *and* $\boldsymbol{T} = \left(T_\theta, T_{\tau_\beta}, T_{\tau_\epsilon}\right)$*. Then* (4.33) *holds.*

Similarly to the hierarchical normal case, studied in Section 4.5, if the precisions $(\tau_\beta, \tau_\epsilon)$ are fixed in specification (4.34), then the spectral gap of P_n can be explicitly studied to deduce limiting bounds on mixing times (see e.g. Bass and Sahu (2016)); while if the precisions are unknown, as it is mostly the case in applications, the performances of P_n have only been empirically studied through simulations. The methodology we introduce here can be used to formally analyze the behaviour of these samplers as $n \to \infty$.

To conclude this section, it is important to note that in this context the kernel P_n may or may not be directly implementable, depending on the specific model formulation. In the commonly used linear case, the full conditional distribution $\pi_n(\mathrm{d}\boldsymbol{\beta} \mid \psi)$ is normal, so that sampling becomes accessible and P_n is directly the algorithm used to sample from π_n. See e.g. Appendix 2 of Bass and Sahu (2016) for details on the implementation, including expressions for the full conditionals. In other cases, e.g. for log-concave likelihoods such as the binary regression ones, adaptive rejection sampling techniques (e.g. Gilks and Wild (1992)) can be used in low dimensions. In the more general case the exact update from $\pi_n(\mathrm{d}\boldsymbol{\beta} \mid \psi)$ is commonly replaced with a Metropolis update from $\pi_n(\mathrm{d}\boldsymbol{\beta} \mid \psi)$ (using e.g. a gradient-based kernel such as MALA or HMC). In the latter case, the Gibbs kernel P_n we analyse here is an idealized version of the practically used Metropolis-within-Gibbs kernel. Under suitable (mild) assumptions, we expect the convergence properties of this idealized scheme to provide a lower bound to the Metropolis-within-Gibbs schemes used in practice. Also, we expect the convergence of the two kernels to be of the same order when the kernel used for the Metropolis updates on the full conditional mixes fast. Providing quantitative results in this direction is an interesting area for future work, which we are currently pursuing. This would extend the applicability of the proof techniques developed in this work to broad classes of non conditionally-conjugate models, such as Gaussian Processes with non-Gaussian likelihood discussed above. See Section 4.7 for more details.

4.6 Feasible start

All the previous results are stated in terms of mixing times from worst case M-warm start, as defined in (4.5). Since starting from $\mu \in \mathcal{N}(\pi_J, M)$ with small M (e.g. not increasing with J) may be in principle infeasible, it is of interest to provide an explicit example of a starting distribution that can be implemented in practice, a so-called feasible start, where the associated value of M can be controlled. In the setting of Theorem 20, the properties of the Gibbs samplers combined with the probabilistic structure of hierarchical models allow to translate the problem of feasible starts into the one of having a good initialisation for the hyper-parameters ψ, as we now show. Indeed, assume that the maximum marginal likelihood estimator $\hat{\psi}_J = \arg\max \prod_{j=1}^J g(Y_j \mid \psi)$, with g as in (4.16), is well-defined. Let $\mu_J \in \mathcal{P}\left(\mathbb{R}^{lJ+D}\right)$ be given by

$$\mu_J(B) = \int_B \mathrm{Unif}\left(\hat{\psi}_J, c/\sqrt{J}\right)(\mathrm{d}\psi) \prod_{j=1}^J p(\theta_j \mid Y_j, \psi)\,\mathrm{d}\boldsymbol{\theta} \qquad B \subset \mathbb{R}^{lJ+D} \tag{4.35}$$

where $c > 0$ is a fixed constant and $\text{Unif}\,(\psi, r)$ denotes the uniform distribution over the closed ball of center ψ and radius $r > 0$. Therefore, the initial point is obtained by sampling from the uniform distribution around the maximum likelihood estimator for ψ and, conditional on this value, from the posterior distribution of the groups specific parameters. The next theorem shows that this choice leads to a good asymptotic behaviour of the mixing times.

Theorem 21. *Consider the same setting of Theorem 20 and let* $\mu_J \in \mathcal{P}\left(\mathbb{R}^{lJ+D}\right)$ *as in* (4.35)*. Then, for every* $\epsilon \in (0,1)$ *there exists* $T\left(\psi^*, \epsilon, c\right) < \infty$ *such that*

$$\liminf_{J\to\infty} Q_{\psi^*}^{(J)}\left(t_{mix}^{(J)}(\epsilon, \mu_J) \leq T\left(\psi^*, \epsilon, c\right)\right) \to 1 \qquad \textit{as } J \to \infty\,.$$

The difference with Theorem 20 is in the specification of the starting distribution, that is now made explicit. Note that whether or not μ_J is a feasible start in practice depends on whether the maximum likelihood estimate $\hat{\psi}_J$ can be computed, using e.g. an Expectation-Maximization algorithm, up to a $\mathcal{O}(1/\sqrt{J})$ error.

Remark 18. *By its definition in* (4.3)*, the Gibbs sampler does not depend on the starting point of the first block. Therefore Theorem 21 extends to any* $\mu_J \in \mathcal{P}\left(\mathbb{R}^{lJ+D}\right)$ *such that*

$$\mu_J\left(\mathbb{R}^{lJ} \times A\right) = \textit{Unif}\left(\hat{\psi}_J, c/\sqrt{J}\right)(A) \qquad\qquad A \subset \mathbb{R}^D\,.$$

4.7 Future works

A first natural extension in this context would be the case where no fixed dimensional sufficient statistic is available, i.e. $p(\cdot \mid \psi)$ in (4.1) does not belong to the exponential family. Since the above dimensionality reduction does not apply there, a possibility is to study the marginal chain induced on ψ; indeed the latter has the same properties of the Gibbs sampler on $(\boldsymbol{\theta}, \psi)$, see e.g. Roberts and Rosenthal (2001). Also, in this work we have focused on the case with well-specified likelihoods but, as discussed after Theorem 19, we expect the misspecified setting to behave in qualitatively similar ways.

Secondly, when dealing with Gibbs samplers, it is often the case that some of the conditional updates cannot be performed exactly. A natural solution is to employ more general coordinate-wise schemes, where exact sampling is replaced by Markov updates with stationary measure given by the conditional distribution. For example in hierarchical models for categorical data (see Section 4.5), while in principle exact conditional sampling is feasible, the parameters θ_j are often sampled in a Metropolis-within-Gibbs fashion, for reasons of computational efficiency and easiness of implementation. While algorithmically convenient, the modification makes theoretical analysis significantly more involved: in particular Proposition 21 ceases to hold and the dimensionality reduction given by Lemma 44 is not available without exact sampling. In ongoing work we are considering a different strategy, by providing lower bounds on the approximate conductance (Lovász and Simonovits, 1993): our preliminary results suggest that, provided the conditional Markov updates have good spectral properties, general coordinate-wise schemes can enjoy the same dimension-free convergence of the Gibbs sampler. Another interesting direction would be to derive results analogous to the ones in Section 4.2 for other MCMC kernels (e.g. gradient-based ones) under appropriate regularity assumptions on the sequence of target distribution, potentially exploiting tools from the recent work in Caprio and Johansen (2023).

Finally, we expect (at least parts of) our methodology to be applicable much beyond hierarchical models as in (4.1). For example, when fitting (finite or infinite) Bayesian mixture models, it is customary to use a Gibbs sampler over a properly augmented space by introducing latent allocation variables (see e.g. Diebolt and Robert (1994)): this leads to a problem of increasing dimensionality, since the number of latent variables grows linearly with n. An asymptotic analysis, as performed in this paper, seems accessible: indeed, posterior concentration results are available (Nguyen, 2013) and a dimensionality reduction similar to Lemma 44 can be exploited. However there are still significant challenges to perform a rigorous analysis in this setting: for example posterior contraction is often proved using Wasserstein distance, that is in general too weak for our purposes. We leave the discussion of such issues to a future work.

A1 Simple counter-examples for Section 4.2

Convergence of the stationary distribution does not imply pointwise convergence of Gibbs operators

Let $\mathcal{X} = [0,1]^2$ and define $A_n = \left[\frac{r_n}{l_n}, \frac{r_n+1}{l_n}\right]$, where

$$r_n = n - 2^{k_n}, \quad l_n = 2^{k_n}, \quad k_n = \lfloor \log_2 n \rfloor,$$

with $\lfloor a \rfloor$ denoting the integer part of a and $n \geq 2$. Therefore $\{A_n\}_n$ is a collection of intervals with decreasing length, such that $x \in A_n$ infinitely often, for every $x \in [0,1]$. We define a sequence $\{\pi_n\}_n \subset \mathcal{P}(\mathcal{X})$ as

$$\pi_n(\mathrm{d}x_1 \mid x_2) = \begin{cases} \mathbb{1}_{[0,1]}(x_1)\,\mathrm{d}x_1, & x_2 \notin A_n \\ \delta_0(\mathrm{d}x_1), & x_2 \in A_n \end{cases}, \quad \pi_n(\mathrm{d}x_2) = \mathbb{1}_{[0,1]}(x_2)\,\mathrm{d}x_2,$$

where $\mathbb{1}_A(x)\,\mathrm{d}x$ denotes the uniform measure on A. Define now

$$\pi(\mathrm{d}x_1, \mathrm{d}x_2) = \mathbb{1}_{[0,1]}(x_1)\mathbb{1}_{[0,1]}(x_2)\mathrm{d}x_1\mathrm{d}x_2$$

and denote $C = \{0\} \times A_n$. For every $B \subset \mathcal{X}$ we have

$$\begin{aligned} |\pi_n(B) - \pi(B)| &\leq |\pi_n(B \cap C) - \pi(B \cap C)| + |\pi_n(B \cap C^c) - \pi(B \cap C^c)| \\ &= \pi_n(B \cap C) \leq \pi_n(C). \end{aligned}$$

Therefore we conclude

$$\|\pi_n - \pi\|_{TV} \leq \pi_n(C) \to 0,$$

as $n \to \infty$. However, if P_n and P are the operators of the associated Gibbs samplers, for every $\mathbf{x} \in \mathcal{X}$ it holds

$$\|P_n(\mathbf{x}, \cdot) - P(\mathbf{x}, \cdot)\|_{TV} \geq |P_n(\mathbf{x}, C) - P(\mathbf{x}, C)|,$$

so that, since $x_2 \in A_n$ infinitely often, we get

$$\|P_n(\mathbf{x}, \cdot) - P(\mathbf{x}, \cdot)\|_{TV} = 1$$

infinitely often. Incidentally, it is not difficult to show that $\mathrm{Gap}(P_n) = 0$ for every n, while $\mathrm{Gap}(P) = 1$. Example 1.4 shows that this mismatch may hold under significantly less pathological scenarios.

Equality of the stationary distributions does not imply closeness of the transition operators

Let $\pi_1 = \pi_2 = \pi$, with π the standard Gaussian distribution. Moreover, let

$$P_1(x, \cdot) = \epsilon\pi(\cdot) + (1-\epsilon)\delta_x(\cdot) \quad \text{and} \quad P_2(x, \cdot) = \epsilon\pi(\cdot) + (1-\epsilon)\delta_{-x}(\cdot),$$

with $\epsilon \in [0,1)$. P_1 and P_2 are uniformly ergodic transition operators with invariant distribution π. Let μ be the truncation of π on the positive real numbers: it is easy to show that $\mu \in \mathcal{N}(\pi, 2)$.

However

$$\|\mu P_1 - \mu P_2\|_{TV} \geq (1-\epsilon)\left[\mu((0,\infty)) - \mu((-\infty,0])\right] = 1-\epsilon.$$

Moreover, it holds that $\|\mu - \pi\|_{TV} = 1/2$, so that we conclude

$$2\|\mu - \pi\|_{TV} - \epsilon \leq \|\mu P_1 - \mu P_2\|_{TV} \leq 2\|\mu - \pi\|_{TV}.$$

Convergence of the stationary distribution in Wasserstein distance does not imply convergence of the mixing times for Gibbs sampler operators

Let $\mathcal{X} = \mathbb{R}^2$ and $\bar{\pi}_n(\mathrm{d}\mathbf{x}) = N(x_1 \mid 0, 1/n)N(x_2 \mid 0, 1/n)\mathrm{d}x_1\mathrm{d}x_2$. Define π_n to be the truncation of $\bar{\pi}_n$ on the set

$$A = \{(-\infty,0]\times(-\infty,0]\}\bigcup\{[0,+\infty)\times[0,+\infty)\}.$$

Let $f : \mathcal{X} \to \mathbb{R}$ be a Lipschitz function with constant 1. Then it holds

$$\int_{\mathcal{X}} \left[f(x_1,x_2) - f(0,0)\right]\pi_n(\mathrm{d}\mathbf{x}) \leq \int_{\mathcal{X}} \sqrt{x_1^2 + x_2^2}\,\pi_n(\mathrm{d}\mathbf{x}) \to 0,$$

as $n \to \infty$, so that $\|\pi_n - \pi\|_W \to 0$, where $\pi(\mathrm{d}\mathbf{x}) = \delta_{(0,0)}(\mathbf{x})$ and $\|\cdot\|_W$ denotes the Wasserstein distance.

If P is the kernel of the Gibbs sampler targeting π, then it is immediate to show that

$$\sup_{\mu\in\mathcal{N}(\pi,M)} \|\mu P - \pi\|_W = 0$$

for every $M \geq 1$, so that the mixing times in Wasserstein distance are equal to 1 for every $\epsilon > 0$.

Instead, denote with μ_n the truncation of π_n on $A_1 = (-\infty,0]\times(-\infty,0]$. It is easy to show that $\mu_n \in \mathcal{N}(\pi_n, 2)$, but

$$\mu_n P_n^t(A_1) - \pi_n(A_1) = \frac{1}{2}$$

for every n and t, where P_n is the kernel of the Gibbs sampler targeting π_n. Since the Wasserstein distance is stronger than the weak one, there exists an absolute constant c such that $\left\|\mu_n P_n^t - \pi_n\right\|_W \geq c$ for every n and t. Therefore, with ϵ small enough and $M \geq 2$, the mixing times of P_n in Wasserstein distance are equal to infinity for every n.

Convergence of the stationary distribution does not imply convergence of the spectral gaps for Gibbs operators

Let $\mathcal{X} = \mathbb{R}^2$ and

$$\pi(\mathrm{d}\mathbf{x}) = N(x_1 \mid 0,1)N(x_2 \mid 0,1)\mathrm{d}x_1\mathrm{d}x_2,$$

where $N(x \mid \mu, \sigma^2)$ is the density function of a gaussian distribution with mean μ and variance σ^2. Define π_n to be the truncation of π on the set A_n, where

$$A_n = \{(-\infty,n]\times(-\infty,n]\}\bigcup\{[n,+\infty)\times[n,+\infty)\}.$$

If P_n and P are the operators of the associated Gibbs samplers, it is not difficult to show that

$$\|\pi_n - \pi\|_{TV} \to 0 \quad \text{and} \quad \|P_n(\mathbf{x},\cdot) - P(\mathbf{x},\cdot)\|_{TV} \to 0$$

as $n \to \infty$, for every $\mathbf{x} \in \mathcal{X}$. However, if $B_n = (-\infty, n] \times (-\infty, n]$ we have

$$\pi_n(B_n) > 0 \quad \text{and} \quad \int_{B_n} P_n\left(\mathbf{x}, B_n^c\right) \pi_n(\mathrm{d}\mathbf{x}) = 0,$$

so that $\mathrm{Gap}(P_n) = 0$ for every n, while $\mathrm{Gap}(P) = 1$.

A2 Regularity assumptions (B4)-(B6) for Theorem 20

Let

$$M_s^{(p)}(\psi \mid y) = E\left[T_s^p(\theta_j) \mid Y_j = y, \psi\right], \tag{36}$$

$$M_{s,s'}^{(p)}(\psi \mid y) = E\left[T_s^p(\theta_j) T_{s'}^p(\theta_j) \mid Y_j = y, \psi\right], \tag{37}$$

be the posterior moments of $\boldsymbol{T}$ given ψ, denote $M^{(p)}(\psi \mid y) = \left(M_1^{(p)}(\psi \mid y), \ldots, M_S^{(p)}(\psi \mid y)\right) \in \mathbb{R}^S$ and

$$\left[C(\psi)\right]_{s,d} = E_{Y_j}\left[\partial_{\psi_d} M_s^{(1)}\left(\psi \mid Y_j\right)\right], \quad \left[V(\psi)\right]_{s,s'} = E_{Y_j}\left[\mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi\right)\right], \tag{38}$$

with $s, s' = 1, \ldots S$ and $d = 1, \ldots, D$. Moreover we write B_δ for the ball of center ψ^* and radius δ, and denote expectations with respect to the law of Y_j as defined in $(B1)$ by $E_{Y_j}[\cdot]$.

$(B4)$ The expectation $M_s^{(p)}(\psi \mid y)$ is well defined for every y and $p = 1, \ldots, 6$. Moreover, there exist $\delta_4 > 0$ and C finite constant such that for every $\psi \in B_{\delta_4}$ it holds $E_{Y_j}\left[\left|\partial_{\psi_d} M_s^{(6)}(\psi \mid Y_j)\right|\right] < C$, $E_{Y_j}\left[\left|\partial_{\psi_d}\partial_{\psi_{d'}} M_s^{(1)}(\psi \mid Y_j)\right|\right] < C$, $E_{Y_j}\left[\left|\partial_{\psi_d} M_{s,s'}^{(1)}(\psi \mid Y_j)\right|\right] < C$ and $E_{Y_j}\left[\left|\partial_{\psi_d}\left\{M_s^{(1)}(\psi \mid Y_j) M_{s'}^{(1)}(\psi \mid Y_j)\right\}\right|\right] < C$ for $s, s' = 1, \ldots, S$ and $d, d' = 1, \ldots, D$. Finally, the matrix $V(\psi^*)$ defined in (38) is non singular.

Assumption $(B4)$ can be understood as a smoothness condition. The posterior distribution of $\boldsymbol{T}$ should not change considerably, if we move from ψ^* to a sufficiently close ψ: this is measured in terms of the derivative of the posterior moments, that must be finite in average. Thanks to $(B4)$ we can prove a suitable conditional Central Limit Theorem to show convergence of a rescaled version of $\boldsymbol{T}$, conditional to ψ and $Y_{1:J}$.

We define the posterior characteristic function of $T(\theta_j) = \left(T_1(\theta_j), \ldots, T_S(\theta_j)\right)$ and $\sum_{j=1}^k T(\theta_j)$, given ψ, as $\varphi\left(t \mid Y_j, \psi\right) = E\left[e^{it^\top T(\theta_j)} \mid Y_j, \psi\right]$ for $t \in \mathbb{R}^S$. and $\varphi^{(k)}\left(t \mid Y_{1:k}, \psi\right) = \prod_{j=1}^k \varphi\left(t \mid Y_j, \psi\right)$, respectively. We will assume:

$(B5)$ There exist $k \geq 1$ and $\delta_5 > 0$ such that

$$\sup_{\psi \in B_{\delta_5}} \int_{\mathbb{R}^S} \left|\varphi^{(k)}\left(t \mid Y_{1:k}, \psi\right)\right|^2 \mathrm{d}t < \infty,$$

for almost every $Y_1, \ldots, Y_k \overset{\text{iid}}{\sim} Q_{\psi^*}$.

$(B6)$ There exist $k' \geq 1$ and $\delta_6 > 0$ such that

$$\sup_{\psi \in B_{\delta_6}} \sup_{|t|>\epsilon} \left| \varphi^{(k')} \left(t \mid Y_{1:k'}, \psi \right) \right| < \phi(\epsilon),$$

for almost every $Y_1, \dots, Y_k \overset{\text{iid}}{\sim} Q_{\psi^*}$, with $\phi(\epsilon) < 1$ for every $\epsilon > 0$.

Assumptions $(B5)$ and $(B6)$ allow the convergence of $\boldsymbol{T}$ to hold for the total variation distance, that is stronger than the weak one, proved through $(B4)$. Loosely speaking, integrability of the characteristic function and its strictly positive distance from 1 guarantee that the distribution is far from being discrete: the latter is exactly the case where weak convergence does not translate to stronger metrics. The problem of proving Central Limit theorems in total variation distance has received considerable attention over the decades: it can be tackled with Fourier-based techniques (Petrov, 1956; Smith, 1953), as we do here, but also with Stein's method (see Ross (2011) for a survey), Malliavin calculus (e.g. Bally and Caramellino (2015)) or through bounds based on entropy (e.g. Bobkov et al. (2014)). Conditions $(B5)$ and $(B6)$ are somewhat reminiscent of the ones in Theorem 19.3 in Bhattacharya and Rao (2010).

A3 Proofs

*

Statement and proof of Lemma 49

Lemma 49. *Let $\mathcal{N} \subset \mathcal{P}(\mathcal{X})$ and $\pi \in \mathcal{P}(\mathcal{X})$. Then*

$$\sup_{\mu \in \mathcal{N}} \inf \left\{ t \geq 1 \, : \, \left\| \mu P^t - \pi \right\|_{TV} < \epsilon \right\} = \inf \left\{ t \geq 1 \, : \, \sup_{\mu \in \mathcal{N}} \left\| \mu P^t - \pi \right\|_{TV} < \epsilon \right\},$$

for every Markov transition kernel P.

Proof. Let

$$t^{(1)} = \sup_{\mu \in \mathcal{N}} \inf \left\{ t \geq 1 \, : \, \left\| \mu P^t - \pi \right\|_{TV} < \epsilon \right\}, \quad t^{(2)} = \inf \left\{ t \geq 1 \, : \, \sup_{\mu \in \mathcal{N}} \left\| \mu P^t - \pi \right\|_{TV} < \epsilon \right\}.$$

Assume $t^{(1)} < \infty$. Then $\left\| \mu P^{t^{(1)}} - \pi \right\|_{TV} < \epsilon$ for every $\mu \in \mathcal{N}$. This implies

$$\sup_{\mu \in \mathcal{N}} \left\| \mu P^{t^{(1)}} - \pi \right\|_{TV} < \epsilon,$$

i.e. $t^{(2)} \leq t^{(1)}$. With a similar reasoning, if $t^{(2)} < \infty$ we have $t^{(1)} \leq t^{(2)}$. Therefore $t^{(1)} = t^{(1)}$ if either $t^{(1)} < \infty$ or $t^{(2)} < \infty$.

Assume now $t^{(1)} = \infty$ and fix $t^* > 0$. By definition of $t^{(1)}$ there exists $\mu \in \mathcal{N}$ such that

$$\left\| \mu P^{t^*} - \pi \right\|_{TV} \geq \epsilon,$$

that implies

$$\sup_{\mu \in \mathcal{N}} \left\| \mu P^{t^*} - \pi \right\|_{TV} \geq \epsilon,$$

i.e. $t^{(2)} > t^*$. Since t^* is arbitrary, we have $t^{(2)} = \infty$. With a similar reasoning, if $t^{(2)} = \infty$ it holds $t^{(1)} = \infty$. □

*

Statement and proof of Lemma 50

Lemma 50. *Let $M \geq 1$, $\pi \in \mathcal{P}(\mathcal{X})$, $\mu \in \mathcal{N}(\pi, M)$ and P be a π-invariant Markov transition kernel. Then $\mu P^t \in \mathcal{N}(\pi, M)$, for every $t \in \mathbb{N}$.*

Proof. Let $A \subseteq \mathcal{X}$. Since $\mu \in \mathcal{N}(\pi, M)$ and P is π-invariant, we have $(\mu P)(A) \leq M(\pi P)(A) = M\pi(A)$. Thus $\mu P \in \mathcal{N}(\pi, M)$ and the result follows by induction on t. □

*

Proof of Lemma 42

Proof. Let $\hat{P}_n = P_n \circ \phi_n^{-1}$ be the push-forward operator of P_n under ϕ_n, defined as

$$\hat{P}_n(\mathbf{x}, B) = P_n\left(\phi_n^{-1}(\mathbf{x}), \phi_n^{-1}(B)\right) \tag{39}$$

for every $\mathbf{x} \in \phi_n(\mathcal{X})$ and $B \subseteq \mathcal{X}$. Since ϕ_n is an injective transformation, $\hat{P}_n$ is a well-defined Markov transition kernel (see e.g. Lemma 1 in Papaspiliopoulos et al. (2020)). Moreover, since ϕ_n is coordinate-wise as in (4.7) we have $\hat{P}_n = \hat{P}_{n,1} \dots \hat{P}_{n,K}$, where

$$\begin{aligned}\hat{P}_{n,i}\left(\mathbf{x}, S_{\mathbf{x},i,A}\right) = P_{n,i}\left(\phi_n^{-1}(\mathbf{x}), S_{\phi_n^{-1}(\mathbf{x}),i,\phi_{n,i}^{-1}(A)}\right) &= \int_{\phi_{n,i}^{-1}(A)} \pi_n\left(\mathrm{d}y_i \mid \phi_n^{-1}(\mathbf{x})^{(-i)}\right) \\ &= \int_A \tilde{\pi}_n\left(\mathrm{d}y_i \mid \mathbf{x}^{(-i)}\right), \quad A \subset \mathcal{X}_i,\end{aligned}$$

so that $\hat{P}_n$ is exactly the operator of the Gibbs sampler targeting $\tilde{\pi}_n$, i.e. $\tilde{P}_n = \hat{P}_n$.

Therefore, since ϕ_n is an injective transformation, by Corollary 2 in Roberts and Rosenthal (2001) we have

$$\left\|\mu_n P_n^t - \pi_n\right\|_{TV} = \left\|\tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n\right\|_{TV},$$

with $\tilde{\mu}_n = \mu_n \circ \phi_n^{-1}$. To conclude the proof, we show that $\tilde{\mu}_n \in \mathcal{N}\left(\tilde{\pi}_n, M\right)$ if and only if $\mu_n \in \mathcal{N}\left(\pi_n, M\right)$. Indeed, to prove the implication from right to left, by definition of push-forward measure we have

$$\tilde{\mu}_n(A) = \mu_n\left(\phi_n^{-1}(A)\right) = \int_{\phi_n^{-1}(A)} \frac{\mathrm{d}\mu_n}{\mathrm{d}\pi_n}(\mathbf{x})\, \pi_n(\mathrm{d}\mathbf{x}) \leq M\pi_n\left(\phi_n^{-1}(A)\right) = M\tilde{\pi}_n(A),$$

for every set $A \subset \mathcal{X}$. Equivalently we obtain the other implication. □

Proof of Proposition 21

For any $\pi \in \mathcal{P}(\mathcal{X})$ and Q Markov transition kernel with state space $\mathcal{X}$, we define $(\pi \otimes Q) \in \mathcal{P}\left(\mathcal{X} \times \mathcal{X}\right)$ as

$$\left(\pi \otimes Q\right)(B) = \int_B Q(\mathbf{x}, \mathrm{d}\mathbf{y})\pi(\mathrm{d}\mathbf{x})$$

for every $B \subseteq \mathcal{X} \times \mathcal{X}$.

Lemma 51. *Let $\pi_1, \pi_2 \in \mathcal{P}(\mathcal{X})$ and Q be a Markov transition kernel with state space $\mathcal{X}$. Then*

$$\|\pi_1 \otimes Q - \pi_2 \otimes Q\|_{TV} = \|\pi_1 - \pi_2\|_{TV}.$$

Proof. By definition of total variation distance we have

$$\begin{aligned}
&\|\pi_1 \otimes Q - \pi_2 \otimes Q\|_{TV} \\
&= \sup_{f\,:\,\mathcal{X}\times\mathcal{X}\to[0,1]} \left| \int_{\mathcal{X}\times\mathcal{X}} f(\mathbf{x},\mathbf{y})Q(\mathbf{x},\mathrm{d}\mathbf{y})\pi_1(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}\times\mathcal{X}} f(\mathbf{x},\mathbf{y})Q(\mathbf{x},\mathrm{d}\mathbf{y})\pi_2(\mathrm{d}\mathbf{x}) \right| \\
&= \sup_{f\,:\,\mathcal{X}\times\mathcal{X}\to[0,1]} \left| \int_{\mathcal{X}} \left(\int_{\mathcal{X}} f(\mathbf{x},\mathbf{y})Q(\mathbf{x},\mathrm{d}\mathbf{y}) \right) \pi_1(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}} \left(\int_{\mathcal{X}} f(\mathbf{x},\mathbf{y})Q(\mathbf{x},\mathrm{d}\mathbf{y}) \right) \pi_2(\mathrm{d}\mathbf{x}) \right| \\
&\leq \sup_{g\,:\,\mathcal{X}\to[0,1]} \left| \int_{\mathcal{X}} g(\mathbf{x})\pi_1(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}} g(\mathbf{x})\pi_2(\mathrm{d}\mathbf{x}) \right| = \|\pi_1 - \pi_2\|_{TV}.
\end{aligned}$$

Also, taking $f(\mathbf{x},\mathbf{y}) = g(\mathbf{x})$ for every $(\mathbf{x},\mathbf{y}) \in \mathcal{X} \times \mathcal{X}$ we have

$$\begin{aligned}
\|\pi_1 - \pi_2\|_{TV} = &\sup_{g\,:\,\mathcal{X}\to[0,1]} \left| \int_{\mathcal{X}} g(\mathbf{x})\pi_1(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}} g(\mathbf{x})\pi_2(\mathrm{d}\mathbf{x}) \right| \\
\leq &\sup_{f\,:\,\mathcal{X}\times\mathcal{X}\to[0,1]} \left| \int_{\mathcal{X}\times\mathcal{X}} f(\mathbf{x},\mathbf{y})Q(\mathbf{x},\mathrm{d}\mathbf{y})\pi_1(\mathrm{d}\mathbf{x}) - \int_{\mathcal{X}\times\mathcal{X}} f(\mathbf{x},\mathbf{y})Q(\mathbf{x},\mathrm{d}\mathbf{y})\pi_2(\mathrm{d}\mathbf{x}) \right| \\
= &\|\pi_1 \otimes Q - \pi_2 \otimes Q\|_{TV}\,.
\end{aligned}$$

□

For $j = 1, 2$, denote the kernel of the Gibbs sampler targeting π_j as $P_j = P_{j,1} \dots P_{j,K}$, where

$$P_{j,i}\left(\mathbf{x}, S_{\mathbf{x},i,A}\right) = \int_A \pi_j \left(\mathrm{d}y_i \mid \mathbf{x}^{(-i)} \right), \quad A \subset \mathcal{X}_i,$$

with $S_{\mathbf{x},i,A} = \{\mathbf{y} \in \mathcal{X} : y_j = x_j \,\forall\, j \neq i \text{ and } y_i \in A\}$ as in the main. By definition, $P_i(\mathbf{x}, \mathrm{d}\mathbf{y})$ depends only on $\mathbf{x}^{(-i)}$. Thus we can define $\left(\pi^{(-i)} \otimes Q\right) \in \mathcal{P}\left(\mathcal{X}^{(-i)} \times \mathcal{X}\right)$ as

$$\left(\pi^{(-i)} \otimes P_i\right)(B) = \int_B P_i\left(\mathbf{x}^{(-i)}, \mathrm{d}\mathbf{y}\right) \pi\left(\mathrm{d}\mathbf{x}^{(-i)}\right),$$

for every $B \subset \mathcal{X}^{(-i)} \times \mathcal{X}$ and similarly for

$$\left(\pi^{(-1)} \otimes P\right) \in \mathcal{P}\left(\mathcal{X}^{(-1)} \times \mathcal{X}\right) \quad \text{and} \quad \left(\pi^{(-i)} \otimes \prod_{j \geq i} P_j\right) \in \mathcal{P}\left(\mathcal{X}^{(-i)} \times \mathcal{X}\right),$$

with $i = 1 \dots, K$. Given this notation we have the following Lemmas.

Lemma 52. *We have*

$$\|\mu P_1 - \mu P_2\|_{TV} \leq M \left\| \pi_2^{(-1)} \otimes P_1 - \pi_2^{(-1)} \otimes P_2 \right\|_{TV}$$

for every $\mu \in \mathcal{N}(\pi_2, M)$ and $M \geq 1$.

Proof. By definition of total variation distance

$$\|\mu P_1 - \mu P_2\|_{TV} = \sup_{f\,:\,\mathcal{X}\to[0,1]} \left| \int_{\mathcal{X}} f(\mathbf{y})\mu P_1(\mathrm{d}\mathbf{y}) - \int_{\mathcal{X}} f(\mathbf{y})\mu P_2(\mathrm{d}\mathbf{y}) \right|.$$

Then, by definition of $\mathcal{N}(\pi_2, M)$, it holds

$$\begin{aligned}
&\|\mu P_1 - \mu P_2\|_{TV} \\
&= M \sup_{f\,:\,\mathcal{X}\to[0,1]} \left| \int_{\mathcal{X}^K} \frac{f(\mathbf{y})}{M} \int_{\mathcal{X}^{(-1)}} \frac{\mathrm{d}\mu^{(-1)}}{\mathrm{d}\pi_2^{(-1)}}(\mathbf{x}^{(-1)}) P_1(\mathbf{x}^{(-1)}, \mathrm{d}\mathbf{y}) \pi_2\left(\mathrm{d}\mathbf{x}^{(-1)}\right) \right. \\
&\qquad\qquad \left. - \int_{\mathcal{X}} \frac{f(\mathbf{y})}{M} \int_{\mathcal{X}^{(-1)}} \frac{\mathrm{d}\mu^{(-1)}}{\mathrm{d}\pi_2^{(-1)}}(\mathbf{x}^{(-1)}) P_2(\mathbf{x}^{(-1)}, \mathrm{d}\mathbf{y}) \pi_2\left(\mathrm{d}\mathbf{x}^{(-1)}\right) \right| \\
&\leq M \sup_{g\,:\,\mathcal{X}^{(-1)}\times\mathcal{X}\to[0,1]} \left| \int_{\mathcal{X}^{(-1)}\times\mathcal{X}} g(\mathbf{x}^{(-1)}, \mathbf{y}) P_1\left(\mathbf{x}^{(-1)}, \mathrm{d}\mathbf{y}\right) \pi_2\left(\mathrm{d}\mathbf{x}^{(-1)}\right) \right. \\
&\qquad\qquad \left. - \int_{\mathcal{X}^{(-1)}\times\mathcal{X}} g(\mathbf{x}^{(-1)}, \mathbf{y}) P_2\left(\mathbf{x}^{-1}, \mathrm{d}\mathbf{y}\right) \pi_2\left(\mathrm{d}\mathbf{x}^{(-1)}\right) \right| \\
&= M \left\| \pi_2^{(-1)} \times P_1 - \pi_2^{(-1)} \otimes P_2 \right\|_{TV}.
\end{aligned}$$

□

Lemma 53. *We have*

$$\begin{aligned}
\left\| \pi_1^{(-i)} \otimes \prod_{j\geq i} P_{1,j} - \pi_2^{(-i)} \otimes \prod_{j\geq i} P_{2,j} \right\|_{TV} &\leq 2\,\|\pi_1 - \pi_2\|_{TV} \\
&+ \left\| \pi_1^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{1,j} - \pi_2^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{2,j} \right\|_{TV}
\end{aligned} \tag{40}$$

for every $i = 1, \ldots, K-1$ and

$$\left\| \pi_1^{(-K)} \otimes P_{1,K} - \pi_2^{(-K)} \otimes P_{2,K} \right\|_{TV} = \|\pi_1 - \pi_2\|_{TV}.$$

Proof. We start by proving (40). Notice that, by definition of $P_{1,i}$ and $P_{2,i}$, we have

$$\begin{aligned}
&\int_{\mathcal{X}^{(-i)}\times\mathcal{X}} g\left(\mathbf{x}^{(-i)}, \mathbf{y}\right) \prod_{j\geq i} P_{1,j}\left(\mathbf{x}^{(-i)}, \mathrm{d}\mathbf{y}\right) \pi_1^{(-i)}\left(\mathrm{d}\mathbf{x}^{(-i)}\right) \\
&\qquad = \int_{\mathcal{X}\times\mathcal{X}^{(-i)}} h\left(\mathbf{x}, \mathbf{y}^{(-i)}\right) \prod_{j\geq i+1} P_{1,j}\left(\mathbf{x}^{(-i-1)}, \mathrm{d}\mathbf{y}\right) \pi_1(\mathrm{d}\mathbf{x})
\end{aligned}$$

and

$$\begin{aligned}
&\int_{\mathcal{X}^{(-i)}\times\mathcal{X}} g\left(\mathbf{x}^{(-i)}, \mathbf{y}\right) \prod_{j\geq i} P_{2,j}\left(\mathbf{x}^{(-i)}, \mathrm{d}\mathbf{y}\right) \pi_2^{(-i)}\left(\mathrm{d}\mathbf{x}^{(-i)}\right) \\
&\qquad = \int_{\mathcal{X}\times\mathcal{X}^{(-i)}} h\left(\mathbf{x}, \mathbf{y}^{(-i)}\right) \prod_{j\geq i+1} P_{2,j}\left(\mathbf{x}^{(-i-1)}, \mathrm{d}\mathbf{y}\right) \pi_2(\mathrm{d}\mathbf{x}),
\end{aligned}$$

where $g : \mathcal{X}^{(-i)} \times \mathcal{X} \to \mathbb{R}$ is any measurable function and h is the composition of g and the function $c : \mathcal{X}^{(-i)} \times \mathcal{X} \to \mathcal{X} \times \mathcal{X}^{(-i)}$ that relocates the $(K-1+i)$-th element of a vector after the $(i-1)$-th element. Since there is a one-to-one relationship between functions g and h, we have

$$\left\| \pi_1^{(-i)} \otimes \prod_{j \geq i} P_{1,j} - \pi_2^{(-i)} \otimes \prod_{j \geq i} P_{2,j} \right\|_{TV} = \left\| \pi_1 \otimes \prod_{j \geq i+1} P_{1,j} - \pi_2 \otimes \prod_{j \geq i+1} P_{2,j} \right\|_{TV}. \tag{41}$$

Then by triangular inequality and Lemma 51 we have

$$\begin{aligned} \left\| \pi_1 \otimes \prod_{j \geq i+1} P_{1,j} - \pi_2 \otimes \prod_{j \geq i+1} P_{2,j} \right\|_{TV} &\leq \left\| \pi_1 \otimes \prod_{j \geq i+1} P_{1,j} - \pi_2 \otimes \prod_{j \geq i+1} P_{1,j} \right\|_{TV} \\ &\quad + \left\| \pi_2 \otimes \prod_{j \geq i+1} P_{1,j} - \pi_2 \otimes \prod_{j \geq i+1} P_{2,j} \right\|_{TV} \\ &\leq \|\pi_1 - \pi_2\|_{TV} + \left\| \pi_2 \otimes \prod_{j \geq i+1} P_{1,j} - \pi_2 \otimes \prod_{j \geq i+1} P_{2,j} \right\|_{TV}. \end{aligned} \tag{42}$$

Notice that $\prod_{j \geq i+1} P_{1,j}$ and $\prod_{j \geq i+1} P_{2,j}$ do not depend on x_{i+1} by construction, that implies

$$\begin{aligned} &\left\| \pi_2 \otimes \prod_{j \geq i+1} P_{1,j} - \pi_2 \otimes \prod_{j \geq i+1} P_{2,j} \right\|_{TV} \\ &= \sup_{h : \mathcal{X} \times \mathcal{X} \to [0,1]} \left| \int_{\mathcal{X} \times \mathcal{X}} h(\mathbf{x}, \mathbf{y}) \prod_{j \geq i+1} P_{1,j}\left(\mathbf{x}^{(-(i+1))}, \mathrm{d}\mathbf{y}\right) \pi_2(\mathrm{d}\mathbf{x}) \right. \\ &\qquad \left. - \int_{\mathcal{X} \times \mathcal{X}} h(\mathbf{x}, \mathbf{y}) \prod_{j \geq i+1} P_{2,j}\left(\mathbf{x}^{(-(i+1))}, \mathrm{d}\mathbf{y}\right) \pi_2(\mathrm{d}\mathbf{x}) \right|, \end{aligned}$$

so that we have

$$\begin{aligned} &\left\| \pi_2 \otimes \prod_{j \geq i+1} P_{1,j} - \pi_2 \otimes \prod_{j \geq i+1} P_{2,j} \right\|_{TV} \\ &= \sup_{h : \mathcal{X} \times \mathcal{X} \to [0,1]} \left| \int_{\mathcal{X}^{(-(i+1))} \times \mathcal{X}} \int_{\mathcal{X}_{i+1}} h(\mathbf{x}, \mathbf{y}) \pi_2\left(\mathrm{d}x_{i+1} \mid x^{(-(i+1))}\right) \prod_{j \geq i+1} P_{1,j}\left(\mathbf{x}^{(-(i+1))}, \mathrm{d}\mathbf{y}\right) \pi_2\left(\mathrm{d}\mathbf{x}^{(-(i+1))}\right) \right. \\ &\qquad \left. - \int_{\mathcal{X}^{(-(i+1))} \times \mathcal{X}} \int_{\mathcal{X}_{i+1}} h(\mathbf{x}, \mathbf{y}) \pi_2\left(\mathrm{d}x_{i+1} \mid x^{(-(i+1))}\right) \prod_{j \geq i+1} P_{2,j}\left(\mathbf{x}^{(-(i+1))}, \mathrm{d}\mathbf{y}\right) \pi_2\left(\mathrm{d}\mathbf{x}^{(-(i+1))}\right) \right| \\ &\leq \left\| \pi_2^{(-(i+1))} \otimes \prod_{j \geq i+1} P_{1,j} - \pi_2^{(-(i+1))} \otimes \prod_{j \geq i+1} P_{2,j} \right\|_{TV}. \end{aligned}$$

Moreover, it is clear that

$$\left\|\pi_2^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{1,j} - \pi_2^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{2,j}\right\|_{TV} \leq \left\|\pi_2 \otimes \prod_{j\geq i+1} P_{1,j} - \pi_2 \otimes \prod_{j\geq i+1} P_{2,j}\right\|_{TV},$$

thus combining the two above inequalities we get

$$\left\|\pi_2 \otimes \prod_{j\geq i+1} P_{1,j} - \pi_2 \otimes \prod_{j\geq i+1} P_{2,j}\right\|_{TV} = \left\|\pi_2^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{1,j} - \pi_2^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{2,j}\right\|_{TV}. \tag{43}$$

Combining (41), (42) and (43) with the fact that

$$\begin{aligned}\left\|\pi_2^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{1,j} - \pi_2^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{2,j}\right\|_{TV} &\leq \|\pi_1 - \pi_2\|_{TV} \\ &+ \left\|\pi_1^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{1,j} - \pi_2^{(-(i+1))} \otimes \prod_{j\geq i+1} P_{2,j}\right\|_{TV}\end{aligned}$$

we finally obtain (40). When $i = K$ the result follows by noticing that

$$\pi_1^{(-K)} \otimes P_{1,K} = \pi_1 \quad \text{and} \quad \pi_2^{(-K)} \otimes P_{2,K} = \pi_2$$

by definition. □

Proof of Proposition 21. Without loss of generality, let $\mu \in \mathcal{N}(\pi_2, M)$. By Lemma 52 and the triangle inequality we have

$$\begin{aligned}\|\mu P_1 - \mu P_2\|_{TV} &\leq M \left\|\pi_2^{(-1)} \otimes P_1 - \pi_2^{(-1)} \otimes P_2\right\|_{TV} \\ &\leq M \|\pi_1 - \pi_2\|_{TV} + M \left\|\pi_1^{(-1)} \otimes P_1 - \pi_2^{(-1)} \otimes P_2\right\|_{TV}\end{aligned}$$

and the result follows by applying K times Lemma 53. □

Proof of Lemma 43

Proof. With an abuse of notation, let $\pi_1(x)$, $\pi_2(x)$ and $\mu_1(x)$ be densities of π_1, π_2 and μ_1 with respect to a common dominating measure, such as $\tau = \pi_1 + \pi_2$. Let $\bar{\mu}$ be the measure on $\mathcal{X}$ with density $\bar{\mu}(x) = \min\{\mu_1(x), M\pi_2(x)\}$ for $x \in \mathcal{X}$. By construction $\bar{\mu}$ is a sub-probability since

$$\bar{\mu}(\mathcal{X}) = \int_{\mathcal{X}} \bar{\mu}(x)\tau(\mathrm{d}x) \leq \int_{\mathcal{X}} \mu_1(x)\tau(\mathrm{d}x) = 1.$$

Therefore, we can define a probability distribution $\mu_2 \in \mathcal{P}(\mathcal{X})$ with density

$$\mu_2(x) = \bar{\mu}(x) + \alpha \max\{M\pi_2(x) - \mu_1(x), 0\}, \qquad x \in \mathcal{X}$$

where

$$\alpha = \frac{1 - \int \bar{\mu}(x)\tau(\mathrm{d}x)}{\int_{\mathcal{X}} \max\{M\pi_2(x) - \mu_1(x), 0\}\tau(\mathrm{d}x)} \in (0, 1).$$

Notice that $\mu_2(x) \leq M\pi_2(x)$ for every $x \in \mathcal{X}$ since

$$\mu_2(x) = \begin{cases} M\pi_2(x), & \text{if } \mu_1(x) > M\pi_2(x)\,, \\ (1-\alpha)\mu_1(x) + \alpha M\pi_2(x), & \text{if } \mu_1(x) \leq M\pi_2(x)\,. \end{cases}$$

Thus $\mu_2 \in \mathcal{N}(\pi_2, M)$. By definition of total variation distance and of $\tilde{\mu}$, we have

$$\begin{aligned} \|\mu_1 - \mu_2\|_{TV} =& \int_{\mathcal{X}} \max\left\{\mu_1(x) - \mu_2(x), 0\right\} \tau(\mathrm{d}x) = \int_{\mathcal{X}} \max\left\{\mu_1(x) - M\pi_2(x), 0\right\} \tau(\mathrm{d}x) \\ \leq& M \int_{\mathcal{X}} \max\left\{\pi_1(x) - \pi_2(x), 0\right\} \tau(\mathrm{d}x) = M\left\|\pi_1 - \pi_2\right\|_{TV}. \end{aligned}$$

□

Proof of Theorem 18

Proof. By Lemma 42 the statement is equivalent to

$$\lim_{n\to\infty} \sup_{\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)} \left\|\tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n\right\|_{TV} = \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\|\tilde{\mu}\tilde{P}^t - \tilde{\pi}\right\|_{TV} \tag{44}$$

in $Q^{(n)}$-probability, where $\tilde{P}_n$ is the kernel of the Gibbs sampler targeting $\tilde{\pi}$.

Consider $\|\tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n\|_{TV}$ with $\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)$. By Lemma 43, there exists $\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)$ such that

$$\|\tilde{\mu}_n - \tilde{\mu}\|_{TV} \leq M\left\|\tilde{\pi}_n - \tilde{\pi}\right\|_{TV}. \tag{45}$$

By the triangular inequality we can decompose $\|\tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n\|_{TV}$ as follows

$$\left\|\tilde{\mu}_n \tilde{P}_n^t - \tilde{\pi}_n\right\|_{TV} \leq \left\|\tilde{\mu}_n \tilde{P}_n^t - \tilde{\mu}\tilde{P}_n^t\right\|_{TV} + \left\|\tilde{\mu}\tilde{P}_n^t - \tilde{\mu}\tilde{P}^t\right\|_{TV} + \left\|\tilde{\mu}\tilde{P}^t - \tilde{\pi}\right\|_{TV} + \left\|\tilde{\pi}_n - \tilde{\pi}\right\|_{TV}. \tag{46}$$

Combining (45) with the monotonicity of the total variation distance with respect to the application of transition kernels, we obtain

$$\left\|\tilde{\mu}_n \tilde{P}_n^t - \tilde{\mu}\tilde{P}_n^t\right\|_{TV} \leq \|\tilde{\mu}_n - \tilde{\mu}\|_{TV} \leq M\left\|\tilde{\pi}_n - \tilde{\pi}\right\|_{TV}\,. \tag{47}$$

For the second term in (46), we want to prove that if $\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)$ we have

$$\left\|\tilde{\mu}\tilde{P}_n^t - \tilde{\mu}\tilde{P}^t\right\|_{TV} \leq 2MKt\left\|\tilde{\pi}_n - \tilde{\pi}\right\|_{TV} \tag{48}$$

for every $t \geq 1$. Indeed, the case $t = 1$ holds by Proposition 21. Assume now (48) holds for $t-1$, with $t \geq 2$. Then by the triangular inequality we have

$$\begin{aligned} \left\|\tilde{\mu}\tilde{P}_n^t - \tilde{\mu}\tilde{P}^t\right\|_{TV} &\leq \left\|\tilde{\mu}\tilde{P_n}^t - \mu\tilde{P}^{t-1}\tilde{P}_n\right\|_{TV} + \left\|\mu\tilde{P}^t - \mu\tilde{P}^{t-1}\tilde{P}_n\right\|_{TV} \\ &\leq \left\|\tilde{\mu}\tilde{P}_n^{t-1} - \tilde{\mu}\tilde{P}^{t-1}\right\|_{TV} + \left\|\mu\tilde{P}^{t-1}\tilde{P} - \mu\tilde{P}^{t-1}\tilde{P}_n\right\|_{TV}. \end{aligned}$$

By induction hypothesis we have

$$\left\|\tilde{\mu}\tilde{P}_n^{t-1} - \tilde{\mu}\tilde{P}^{t-1}\right\|_{TV} \leq 2MK(t-1)\left\|\tilde{\pi}_n - \tilde{\pi}\right\|_{TV}. \tag{49}$$

Moreover, by Lemma 50 we have that $\tilde{\mu}\tilde{P}^{t-1} \in \mathcal{N}(\tilde{\pi}, M)$, so that from the case $t = 1$ we obtain

$$\left\|\mu\tilde{P}^{t-1}\tilde{P} - \mu\tilde{P}^{t-1}\tilde{P}_n\right\|_{TV} \leq 2MK \left\|\tilde{\pi}_n - \tilde{\pi}\right\|_{TV}. \tag{50}$$

Then (48) follows by (49) and (50). Combining (46), (47) and (48), for every $\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)$ there exists $\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)$ such that

$$\left\|\tilde{\mu}_n\tilde{P}_n^t - \tilde{\pi}_n\right\|_{TV} \leq (2MKt + M + 1)\left\|\tilde{\pi}_n - \tilde{\pi}\right\|_{TV} + \left\|\tilde{\mu}\tilde{P}^t - \tilde{\pi}_n\right\|_{TV}.$$

Thus

$$\sup_{\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)} \left\|\tilde{\mu}_n\tilde{P}_n^t - \tilde{\pi}_n\right\|_{TV} \leq (2MKt + M + 1)\left\|\tilde{\pi}_n - \tilde{\pi}\right\|_{TV} + \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\|\tilde{\mu}\tilde{P}^t - \tilde{\pi}\right\|_{TV}.$$

It follows that, for any $\epsilon > 0$, we have

$$\begin{aligned} &Q^{(n)}\left(\sup_{\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)} \left\|\tilde{\mu}_n\tilde{P}_n^t - \tilde{\pi}_n\right\|_{TV} - \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\|\tilde{\mu}\tilde{P}^t - \tilde{\pi}\right\|_{TV} \geq \epsilon\right) \\ &\leq Q^{(n)}\left(\left\|\tilde{\pi}_n - \tilde{\pi}\right\|_{TV} \geq (2MKt + M + 1)^{-1}\epsilon\right) \to 0, \end{aligned} \tag{51}$$

as $n \to \infty$ by $(A1)$ and $(2MKt + M + 1)^{-1}\epsilon > 0$.

We now prove the reverse inequality of (51) to establish (44). Given $\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)$, by Lemma 43, there exists $\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)$ such that $\left\|\tilde{\mu} - \tilde{\mu}_n\right\|_{TV} \leq M\left\|\tilde{\pi}_n - \tilde{\pi}\right\|_{TV}$. Then we proceed analogously to above, first decomposing $\left\|\tilde{\mu}\tilde{P}^t - \tilde{\pi}\right\|_{TV}$ as

$$\left\|\tilde{\mu}\tilde{P}^t - \tilde{\pi}\right\|_{TV} \leq \left\|\tilde{\mu}\tilde{P}^t - \tilde{\mu}\tilde{P}_n^t\right\|_{TV} + \left\|\tilde{\mu}\tilde{P}_n^t - \tilde{\mu}_n\tilde{P}_n^t\right\|_{TV} + \left\|\tilde{\mu}_n\tilde{P}_n^t - \tilde{\pi}_n\right\|_{TV} + \left\|\tilde{\pi}_n - \tilde{\pi}\right\|_{TV} \tag{52}$$

and then applying Proposition 21 using an argument analogous to above to get

$$\left\|\tilde{\mu}\tilde{P}^t - \tilde{\pi}\right\|_{TV} \leq \left\|\tilde{\mu}_n\tilde{P}_n^t - \tilde{\pi}_n\right\|_{TV} + (2MKt + M + 1)\left\|\tilde{\pi}_n - \tilde{\pi}\right\|_{TV}.$$

It follows

$$\sup_{\mu_n \in \mathcal{N}(\tilde{\pi}_n, M)} \left\|\tilde{\mu}_n\tilde{P}_n^t - \tilde{\pi}_n\right\|_{TV} \geq \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\|\tilde{\mu}\tilde{P}^t - \tilde{\pi}\right\|_{TV} - (2MKt + M + 1)\left\|\tilde{\pi}_n - \tilde{\pi}\right\|_{TV}.$$

Fixing $\epsilon > 0$ arbitrary constant we have

$$\begin{aligned} &Q^{(n)}\left(\sup_{\tilde{\mu}_n \in \mathcal{N}(\tilde{\pi}_n, M)} \left\|\tilde{\mu}_n\tilde{P}_n^t - \tilde{\pi}_n\right\|_{TV} - \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\|\tilde{\mu}\tilde{P}^t - \tilde{\pi}\right\|_{TV} \leq -\epsilon\right) \\ &\leq Q^{(n)}\left(\left\|\tilde{\pi}_n - \tilde{\pi}\right\|_{TV} \geq \frac{\epsilon}{2MKt + M + 1}\right) \to 0, \end{aligned} \tag{53}$$

as $n \to \infty$ by $(A1)$ and $(2MKt + M + 1)^{-1}\epsilon > 0$. The result follows by combining (51) and (53). □

Proof of Corollary 8

Proof. Thanks to Lemma 49 we can write

$$t_{mix}^{(n)}(\epsilon, M) = \inf \left\{ t \geq 1 \,:\, \sup_{\mu_n \in \mathcal{N}(\pi_n, M)} \left\| \mu_n P_n^t - \pi_n \right\|_{TV} < \epsilon \right\}$$

and

$$\tilde{t}_{mix}(\epsilon, M) = \inf \left\{ t \geq 1 \,:\, \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\| \tilde{\mu} \tilde{P}^t - \tilde{\pi} \right\|_{TV} < \epsilon \right\}.$$

Assume $(A1)$ and denote $t^* = \tilde{t}_{mix}(\epsilon, M) < \infty$ for brevity. By definition of t^* we have $\delta = \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\| \tilde{\mu} \tilde{P}^{t^*} - \tilde{\pi} \right\|_{TV} < \epsilon$. Thus

$$\begin{aligned} Q^{(n)} \left(t_{mix}^{(n)}(\epsilon, M) \leq t^* \right) &= Q^{(n)} \left(\sup_{\mu_n \in \mathcal{N}(\pi_n, M)} \left\| \mu_n P_n^{t^*} - \pi_n \right\|_{TV} < \epsilon \right) \\ &= Q^{(n)} \left(\sup_{\mu_n \in \mathcal{N}(\pi_n, M)} \left\| \mu_n P_n^{t^*} - \pi_n \right\|_{TV} - \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\| \tilde{\mu} \tilde{P}^{t^*} - \tilde{\pi} \right\|_{TV} < \epsilon - \delta \right) \\ &\to 1, \end{aligned}$$

as $n \to \infty$ by Theorem 18.

As regards the second part of the statement, let $(A1)$ hold and fix $T > 0$. Denote $\delta = \sup_{\tilde{\mu} \in \mathcal{N}(\tilde{\pi}, M)} \left\| \tilde{\mu} \tilde{P}^T - \tilde{\pi} \right\|_{TV}$ and notice that by assumption $\delta \geq \epsilon > \underline{\epsilon}$. Thus

$$\begin{aligned} \liminf_{n \to \infty} Q^{(n)} \left(t_{mix}^{(n)}(\underline{\epsilon}, M) < T \right) &= \liminf_{n \to \infty} Q^{(n)} \left(\sup_{\mu_n \in \mathcal{N}(\pi_n, M)} \left\| \mu_n P_n^T - \pi_n \right\|_{TV} < \underline{\epsilon} \right) \\ &= \liminf_{n \to \infty} Q^{(n)} \left(\delta - \sup_{\mu_n \in \mathcal{N}(\pi_n, M)} \left\| \mu_n P_n^T - \pi_n \right\|_{TV} \geq \delta - \underline{\epsilon} \right) \\ &\to 0, \end{aligned}$$

as $n \to \infty$ by Theorem 18. □

Proof of Corollary 9

We need a preliminary well known lemma, whose proof we include for self-containedness.

Lemma 54. *Let P be a Gibbs sampler kernel with $K = 2$ and target $\pi \in \mathcal{P}(\mathcal{X}_1 \times \mathcal{X}_2)$. Then*

$$\left\| \mu P^t - \pi \right\|_{TV} \leq \frac{M}{2} \left(1 - Gap(P) \right)^t,$$

for every $\mu \in \mathcal{N}(\pi, M)$ and $t \geq 1$.

Proof. Let $\mu \in \mathcal{N}(\pi, M)$ and $t \geq 1$. By Corollary 1 in Roberts and Rosenthal (2001) we have

$$\left\| \mu P^t - \pi \right\|_{TV} = \left\| \mu^{(-1)} \hat{P}^t - \pi^{(-1)} \right\|_{TV}, \tag{54}$$

where $\hat{P}$ is the Markov transition kernel on $\mathcal{X}_2$ defined as

$$\hat{P}(x_2, \mathrm{d}y_2) = \int_{\mathcal{X}_1} \pi(\mathrm{d}y_2 \mid y_1)\pi(\mathrm{d}y_1 \mid x_2) \qquad\qquad x_2 \in \mathcal{X}_2\,.$$

Note that $\hat{P}$ is $\pi^{(-1)}$-reversible. Also, for every $f \in L^2(\pi^{(-1)})$, i.e. $f : \mathcal{X}_2 \to \mathbb{R}$ such that $\|f\|_2^2 = \pi^{(-1)}(f^2)$ is finite, we have

$$\begin{aligned}
&\int_{\mathcal{X}_2^2} f(x_2)f(y_2)\hat{P}(x_2, \mathrm{d}y_2)\pi(\mathrm{d}x_2) \\
&\quad = \int_{\mathcal{X}_2^2} f(x_2)f(y_2)\int_{\mathcal{X}_1} \pi(\mathrm{d}y_2 \mid y_1)\pi(\mathrm{d}y_1 \mid x_2)\pi(\mathrm{d}x_2) \\
&\quad = \int_{\mathcal{X}_1} \left[\int_{\mathcal{X}_2} f(y_2)\pi(\mathrm{d}y_2 \mid y_1)\right]\left[\int_{\mathcal{X}_2} f(x_2)\pi(\mathrm{d}x_2 \mid y_1)\right]\pi(\mathrm{d}y_1) \\
&\quad = \int_{\mathcal{X}_1} \left[\int_{\mathcal{X}_2} f(y_2)\pi(\mathrm{d}y_2 \mid y_1)\right]^2 \pi(\mathrm{d}y_1) \geq 0,
\end{aligned}$$

so that $\hat{P}$ is also positive semi-definite. Since $\hat{P}$ is reversible and positive semi-definite, we have (see e.g. equation (5) in Andrieu et al. (2022)) that

$$\left|\left|\hat{P}^t(f)\right|\right|_2 \leq ||f||_2 \left(1 - \mathrm{Gap}(\hat{P})\right)^t, \tag{55}$$

for every f such that $\pi(f) = 0$. Choosing $f = \frac{\mathrm{d}\mu^{(-1)}}{\mathrm{d}\pi^{(-1)}} - 1$ and using the reversibility of $\hat{P}$ (see e.g. Section 2.1 in Khare and Zhou (2009)) we also have

$$\left\|\mu^{(-1)}\hat{P}^t - \pi^{(-1)}\right\|_{TV} \leq \frac{1}{2}\left|\left|\mu^{(-1)}\hat{P}^t(f)\right|\right|_2, \tag{56}$$

where $\mu^{(-1)}\hat{P}^t(f) = \int f(x_2)\mu^{(-1)}\hat{P}^t(\mathrm{d}x_2)$. With the same choice of f, we have

$$||f||_2^2 = \int \left(\frac{\mathrm{d}\mu^{(-1)}}{\mathrm{d}\pi^{(-1)}}(x_2) - 1\right)^2 \pi^{-1}(\mathrm{d}x_2) \leq M^2$$

since $\mu^{(-1)} \in \mathcal{N}(\pi^{(-1)}, M)$. Thus, combining (55) with (56) we obtain

$$\left\|\mu P^t - \pi\right\|_{TV} \leq \frac{M}{2}\left(1 - \mathrm{Gap}(\hat{P})\right)^t.$$

Finally, for every $f : \mathcal{X}_2 \to \mathbb{R}$ with $||f||_2 < \infty$ it holds

$$\frac{\int_{\mathcal{X}_2^2} \left[f(y_2) - f(x_2)\right]^2 \pi(\mathrm{d}x_2)\hat{P}(x_2, \mathrm{d}y_2)}{2\mathrm{Var}_\pi^{(-1)}(f)} = \frac{\int_{\mathcal{X}^2} \left[g(\mathbf{y}) - g(\mathbf{x})\right]^2 \pi(\mathrm{d}\mathbf{x})P(\mathbf{x}, \mathrm{d}\mathbf{y})}{2\mathrm{Var}_\pi(f)},$$

where $g(\mathbf{x}) = f(x_2)$. Therefore $\mathrm{Gap}(\hat{P}) \geq \mathrm{Gap}(P)$ and we get

$$\left\|\mu P^t - \pi\right\|_{TV} \leq \frac{M}{2}\left(1 - \mathrm{Gap}(P)\right)^t,$$

as desired. □

Proof of Corollary 9. By Lemma 54 we obtain

$$\tilde{t}_{mix}(\epsilon, M) \leq 1 + \frac{\log(M/2) - \log(\epsilon)}{-\log\left(1 - \mathrm{Gap}(\tilde{P})\right)},$$

and the result follows by the first part of Corollary 8. □

Proof of Proposition 22

Proof. By Theorem 19, assumption $(A1)$ is satisfied with

$$\phi_n(\psi) = \sqrt{n}(\psi - \psi^*) - \mathcal{I}^{-1}(\psi^*)\Delta_{n,\psi^*},$$

and $\tilde{\pi} = N\left(\mathbf{0}, \mathcal{I}^{-1}(\psi^*)\right)$. Since $\tilde{\pi}$ is the distribution of a multivariate normal with non singular covariance matrix, then it is easy to show $\tilde{t}_{mix}(\epsilon, M) < \infty$ for every $(M, \epsilon) \in [1, \infty) \times (0, 1)$, see e.g. Theorem 2 in Amit (1991). □

Statement and proof of Corollary 17

We illustrate the result of Proposition 22 on a simple example of model (4.11) with normal likelihood and unknown mean and precision, that is

$$f(y \mid \mu, \tau) = N\left(y \mid \mu, \tau^{-1}\right), \tag{57}$$

where $K = 2$ and $\psi = (\mu, \tau)$. Notice that, even if a conjugate prior exists, it is common to place independent priors on μ and τ, for which the Gibbs sampler defined in (4.3) becomes a reasonable option.

Corollary 17. *Consider model* (4.11) *with likelihood as in* (57). *Let* $Y_i \overset{iid}{\sim} Q_{\psi^*}$, *with* Q_{ψ^*} *admitting density* $f(y \mid \psi^*)$ *and* $\psi^* = (\mu^*, \tau^*) \in \mathbb{R} \times \mathbb{R}_+$. *Moreover let* p_0 *be absolutely continuous in a neighborhood of* ψ^* *with a continuous positive density at* ψ^*. *Consider the Gibbs sampler defined in* (4.3). *Then, for every* $M \geq 1$ *and* $\epsilon > 0$ *we have*

$$Q^{(n)}_{\psi^*}\left(t^{(n)}_{mix}(\epsilon, M) \leq 1\right) \to 1,$$

as $n \to \infty$.

For the proof we need a preliminary Lemma, whose proof we include for self-containedness and because it will be useful to refer to later on.

Lemma 55. *Consider the same setting of Corollary 17. Then conditions* (4.12) *are satisfied.*

Proof of Lemma 55. Define

$$\Psi = \Psi_1 \times \Psi_2 = [\mu^* - 1, \mu^* + 1] \times \left[\frac{\tau^*}{2}, 2\tau^*\right]$$

compact neighborhood of ψ^* and

$$u_n(Y_1, \ldots, Y_n) = 1 - \mathbb{1}_{g_1(Y_{1:n}) \leq c_1} \, \mathbb{1}_{g_2(Y_{1:n}) \leq c_2},$$

where $c_1 = 1/2$, $c_2 = (2\tau^*)^{-1}$ and

$$g_1(Y_{1:n}) = \left|\bar{Y} - \mu^*\right|, \quad \text{and} \quad g_2(Y_{1:n}) = \left|\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 - \frac{1}{\tau^*}\right|,$$

with $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Since $Y_i \overset{\text{iid}}{\sim} N(\mu, \tau^{-1})$, then $g_1(Y_{1:n})$ and $g_2(Y_{1:n})$ are equal in distribution, respectively, to

$$h_1(Z_{1:n}, \mu, \tau) = \left|\frac{1}{\sqrt{\tau}}\bar{Z} + \mu - \mu^*\right|, \quad h_2(Z_{1:n}, \mu, \tau) = \left|\frac{1}{\tau}\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2 - \frac{1}{\tau^*}\right|,$$

where $Z_i \overset{\text{iid}}{\sim} N(0, 1)$. By the Law of Large numbers we have

$$\bar{Z} \to 0, \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2 \to 1$$

almost surely as $n \to \infty$. This implies

$$\begin{aligned}\int u_n(y_1, \ldots, y_n) \prod_{i=1}^{n} f(\mathrm{d}y_i \mid \psi^*) \leq & P\left(h_1(Z_{1:n}, \mu^*, \tau^*) > c_1\right) \\ & + P\left(h_2(Z_{1:n}, \mu^*, \tau^*) > c_2\right) \to 0,\end{aligned}$$

as $n \to \infty$. Also, we have

$$\begin{aligned}\sup_{\psi \notin \Psi} \int \left[1 - u_n(y_1, \ldots, y_n)\right] \prod_{i=1}^{n} f(\mathrm{d}y_i \mid \psi) \leq & \sup_{\tau \notin \Psi_2} P\left(h_2(Z_{1:n}, \mu, \tau) \leq c_2\right) \\ & + \sup_{\mu \notin \Psi_1, \tau \in \Psi_2} P\left(h_1(Z_{1:n}, \mu, \tau) \leq c_1\right).\end{aligned}$$

Now notice that by the reverse triangle inequality we have

$$\begin{aligned}\sup_{\tau \notin \Psi_2} P\left(h_2(Z_{1:n}, \mu, \tau) \leq c_2\right) &= \sup_{\tau \notin \Psi_2} P\left(\left|\frac{1}{\tau}\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2 - \frac{1}{\tau^*}\right| \leq c_2\right) \\ &\leq \sup_{\tau \notin \Psi_2} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(Z_i - \bar{Z}\right)^2 - 1\right| \geq \left|1 - \frac{\tau}{\tau^*}\right| - c_2\tau\right) \to 0,\end{aligned}$$

by definition of Ψ_2, as $n \to \infty$. Finally, again by reverse triangle inequality, we have

$$\sup_{\mu \notin \Psi_1, \tau \in \Psi_2} P\left(h_1(Z_{1:n}, \mu, \tau) \leq c_1\right) \leq \sup_{\mu \notin \Psi_1, \tau \in \Psi_2} P\left(|\bar{Z}| \geq \sqrt{\tau}\left(|\mu - \mu^*| - c_1\right)\right) \to 0,$$

as $n \to \infty$. □

Proof of Corollary 17. In this case $\psi = (\mu, \tau)$ and

$$f(y \mid \psi) = \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(y-\mu)^2}.$$

By Lemma 55 conditions (4.12) are satisfied. Also, the map $\psi \to f(y \mid \psi)$ is one-to-one, the map $\psi \to \sqrt{f(y \mid \psi)}$ is continuously differentiable, and the Fisher information matrix is

$$\mathcal{I}(\psi) = \begin{bmatrix} \frac{\tau}{2} & 0 \\ 0 & \frac{1}{2\tau} \end{bmatrix},$$

which is non singular and continuous as a function of ψ. Thus the conditions of Theorem 19 and Proposition 22 are satisfied. Finally, since we are considering a two-blocks Gibbs sampler, by Corollary 9 we have

$$T\left(\psi^*, \epsilon, M\right) = 1 + \frac{\log(M/2) - \log(\epsilon)}{-\log\left(1 - \operatorname{Gap}(\tilde{P})\right)},$$

where $\tilde{P}$ is the Gibbs sampler targeting a bivariate normal distribution with covariance matrix given by $\mathcal{I}^{-1}(\psi^*)$. Since the latter is diagonal, the Gibbs sampler coincides with independent sampling, so that $\operatorname{Gap}(\tilde{P}) = 1$. □

Proof of Lemma 44

Proof. Denote by $\left(\boldsymbol{\theta}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$ the Markov chain with kernel P_J defined in (4.15). The Markovianity of the induced sequence $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$ follows by the one of $\left(\psi^{(t)}\right)_{t\geq 1}$, which is well known (Diaconis et al., 2008; Roberts and Rosenthal, 2001). We now show that $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$ admits $\hat{P}_J$ as kernel. The conditional distribution of $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)$ given $\left(\boldsymbol{T}^{(t-1)}, \psi^{(t-1)}\right)$ is given by

$$\begin{aligned}\mathcal{L}\left(\mathrm{d}\boldsymbol{T}^{(t)}, \mathrm{d}\psi^{(t)} \mid \boldsymbol{T}^{(t-1)}, \psi^{(t-1)}\right) &= \mathcal{L}\left(\mathrm{d}\boldsymbol{T}^{(t)} \mid \boldsymbol{T}^{(t-1)}, \psi^{(t-1)}\right) \mathcal{L}\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{T}^{(t)}, \psi^{(t-1)}, \boldsymbol{T}^{(t-1)}\right) \\ &= \hat{\pi}_J\left(\mathrm{d}\boldsymbol{T}^{(t)} \mid \psi^{(t-1)}\right) \mathcal{L}\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{T}^{(t)}, \psi^{(t-1)}\right),\end{aligned}$$

where the last equality follows by (4.15) and the definition of $\hat{\pi}_J$. By the exponential family assumption in (4.14), $\boldsymbol{T}$ is a set of sufficient statistics for ψ, so that

$$\pi_J\left(\mathrm{d}\psi \mid \boldsymbol{\theta}\right) = \mathcal{L}\left(\mathrm{d}\psi \mid \boldsymbol{\theta}, Y_{1:J}\right) = \mathcal{L}\left(\mathrm{d}\psi \mid \boldsymbol{T}(\boldsymbol{\theta}), Y_{1:J}\right) = \hat{\pi}_J\left(\mathrm{d}\psi \mid \boldsymbol{T}(\boldsymbol{\theta})\right). \tag{58}$$

Combining (4.15) and (58) we have

$$\begin{aligned}\mathcal{L}\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{T}^{(t)}, \psi^{(t-1)}\right) &= \int \pi_J\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{\theta}\right) \pi_J\left(\mathrm{d}\boldsymbol{\theta} \mid \boldsymbol{T}^{(t)}, \psi^{(t-1)}\right) \\ &= \int \hat{\pi}_J\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{T}(\boldsymbol{\theta})\right) \pi_J\left(\mathrm{d}\boldsymbol{\theta} \mid \boldsymbol{T}^{(t)}, \psi^{(t-1)}\right) = \hat{\pi}_J\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{T}^{(t)}\right)\end{aligned} \tag{59}$$

since $\boldsymbol{T}(\boldsymbol{\theta}) = \boldsymbol{T}^{(t)}$ almost surely under $\pi_J\left(\mathrm{d}\boldsymbol{\theta} \mid \boldsymbol{T}^{(t)}, \psi^{(t-1)}\right)$. Thus we can conclude

$$\begin{aligned}\mathcal{L}\left(\mathrm{d}\boldsymbol{T}^{(t)}, \mathrm{d}\psi^{(t)} \mid \boldsymbol{T}^{(t-1)}, \psi^{(t-1)}\right) &= \hat{\pi}_J\left(\mathrm{d}\boldsymbol{T}^{(t)} \mid \psi^{(t-1)}\right)\hat{\pi}_J\left(\mathrm{d}\psi^{(t)} \mid \boldsymbol{T}^{(t)}\right)\\ &= \hat{P}_J\left(\left(\boldsymbol{T}^{(t-1)}, \psi^{(t-1)}\right), \left(\mathrm{d}\boldsymbol{T}^{(t)}, \mathrm{d}\psi^{(t)}\right)\right),\end{aligned}$$

as desired. From the above one can easily deduce that $\left(\boldsymbol{\theta}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$ and $\left(\boldsymbol{T}^{(t)}, \psi^{(t)}\right)_{t\geq 1}$ are *co-deinitializing* as in Roberts and Rosenthal (2001) and thus, by Corollary 2 therein, for every $\mu \in \mathcal{P}\left(\mathbb{R}^{\ell J} \times \mathbb{R}^D\right)$ we have

$$\left\|\mu P_J^t - \pi_J\right\|_{TV} = \left\|\nu \hat{P}_J^t - \hat{\pi}_J\right\|_{TV}, \tag{60}$$

where $\nu \in \mathcal{P}\left(\mathbb{R}^S \times \mathbb{R}^D\right)$ is the push forward of μ under $(\boldsymbol{\theta}, \psi) \mapsto (\boldsymbol{T}(\boldsymbol{\theta}), \psi)$. Moreover, by (4.5) we have that $\nu \in \mathcal{N}(\hat{\pi}_J, M)$ whenever $\mu \in \mathcal{N}(\pi_J, M)$. It follows that $\sup_{\mu\in\mathcal{N}(\pi_J,M)} t_{mix}^{(J)}(\epsilon, \mu) \leq \sup_{\nu\in\mathcal{N}(\hat{\pi}_J,M)} \hat{t}_{mix}^{(J)}(\epsilon, \nu)$. For the reverse inequality, fix $\nu \in \mathcal{N}(\hat{\pi}_J, M)$ and take $\mu(\mathrm{d}\boldsymbol{\theta}, \mathrm{d}\psi) = \int \pi_J(\mathrm{d}\boldsymbol{\theta} \mid \boldsymbol{T}, \psi)\, \nu(\mathrm{d}\boldsymbol{T}, \mathrm{d}\psi)$. By (4.5) we have $\mu \in \mathcal{N}(\pi_J, M)$ and thus (60). It follows $\sup_{\nu\in\mathcal{N}(\hat{\pi}_J,M)} \hat{t}_{mix}^{(J)}(\epsilon, \nu) \leq \sup_{\mu\in\mathcal{N}(\pi_J,M)} t_{mix}^{(J)}(\epsilon, \mu)$ as desired. □

Proof of Lemma 45

Proof. The result follows immediately from Theorem 19, whose assumptions are given exactly by assumption $(B1) - (B3)$, with likelihood $g(y \mid \psi)$. □

Proof of Lemma 46

The proof is divided in two main steps: the result is firstly proved under the weak metric (Lemma 58) and it is then extended to the total variation distance.

First of all we need two technical lemmas, that we prove for completeness.

Lemma 56. *Let S and p be two positive integers. Then there exists a constant $C = C(S, p)$ such that*

$$|\boldsymbol{x}|^p \leq 1 + C\sum_{s=1}^{S} x_s^{2p}$$

for every $\boldsymbol{x} \in \mathbb{R}^S$.

Proof. Since $(1 - |\mathbf{x}|^p)^2 \geq 0$, we have $|\mathbf{x}|^p \leq 1 + |\mathbf{x}|^{2p}$. Moreover, by the Multinomial Theorem, we get

$$|\mathbf{x}|^{2p} = \left(\sum_{s=1}^{S} x_s^2\right)^p = \sum_{\boldsymbol{k}\in\mathbb{P}} \binom{p}{k_1 \ldots k_S} \prod_{s=1}^{S} x_s^{2k_s},$$

where $\mathbb{P} = \left\{\boldsymbol{k} = (k_1, \ldots, k_S) : k_s \text{ positive integer}, \sum_{s=1}^{S} k_s = p\right\}$. Since

$$\prod_{s=1}^{S} x_s^{2k_s} \leq \left(\max_s |x_s|\right)^{2p} \leq \sum_{s=1}^{S} x_s^{2p},$$

the result follows by choosing $C = \sum_{\boldsymbol{k}\in\mathbb{P}} \binom{p}{k_1 \dots k_S}$. □

Lemma 57. *Under assumption* ($B3$), *the random variables* $\Delta_J = (\Delta_{J,1}, \dots, \Delta_{J,D})$ *defined in* (4.17) *are such that for every* $\beta > 0$ *we have*

$$\frac{1}{J^\beta}\Delta_{J,d} \quad \to \quad 0,$$

$Q_{\psi^*}^{(\infty)}$*-almost surely as* $J \to \infty$ *for every* $d = 1, \dots, D$.

Proof. Recall that

$$\Delta_{J,d} = \frac{1}{\sqrt{J}} \sum_{j=1}^{J} \left[\mathcal{I}^{-1}(\psi^*)\nabla \log g(Y_j \mid \psi^*)\right]_d =: \frac{1}{\sqrt{J}} \sum_{j=1}^{J} X_{j,d}$$

and $\mathcal{I}^{-1}(\psi^*)\partial_{\psi_d} \log g(Y_j \mid \psi^*)$ has zero mean and finite variance, by ($B3$). Therefore, by Chebychev inequality

$$P\left(\left|\frac{1}{J^\beta}\Delta_{J,d}\right| > \epsilon\right) \leq \frac{\mathrm{Var}\left(X_{1,d}\right)}{\epsilon^2 J^{1+2\beta}},$$

for every $\epsilon > 0$. This implies

$$\sum_{J=1}^{\infty} P\left(\left|\frac{1}{J^\beta}\Delta_{J,d}\right| > \epsilon\right) \leq \sum_{J=1}^{\infty} \frac{\mathrm{Var}\left(X_{1,d}\right)}{\epsilon^2 J^{1+2\beta}} < \infty,$$

and the result follows by Borel-Cantelli Lemma. □

Weak convergence

In order to ease the following exposition, denote

$$\psi^{(J)} := \psi^* + \frac{\tilde{\psi} + \Delta_J}{\sqrt{J}}, \quad J \geq 1\,. \tag{61}$$

The next lemma proves convergence of $\tilde{\boldsymbol{T}}$ using the weak metric, denoted by $\|\cdot\|_W$.

Lemma 58. *Define* $\tilde{\psi}$ *and* $\tilde{\boldsymbol{T}}$ *as in* (4.17) *and* (4.19), *respectively. Under assumptions* ($B1$) – ($B4$), *for every* $\tilde{\psi} \in \mathbb{R}^D$ *it holds*

$$\left\|\mathcal{L}(d\tilde{\boldsymbol{T}} \mid Y_{1:J}, \tilde{\psi}) - N\left(C(\psi^*)\tilde{\psi}, V(\psi^*)\right)\right\|_W \to 0, \tag{62}$$

$Q_{\psi^*}^{(\infty)}$*-almost surely as* $J \to \infty$.

Proof. For ease of notation, denote

$$\mu = C(\psi^*)\tilde{\psi} \quad \text{and} \quad \Xi := V(\psi^*).$$

By definition of $M_s^{(p)}$, we have

$$E\left[T_s^p(\theta_j) \mid Y_j, \psi^{(J)}\right] = M_s^{(p)}\left(\psi^{(J)} \mid Y_j\right).$$

Conditional on $\tilde{\psi}$, the group specific statistics $T_s(\theta_j)$ are independent across $j = 1, \dots, J$. Thus, by Lyapunov version of Central Limit Theorem, in order to obtain (62) it suffices to show

$$\frac{1}{\sqrt{J}} \sum_{j=1}^{J} \left[M^{(1)}\left(\psi^{(J)} \mid Y_j\right) - M^{(1)}\left(\psi^* \mid Y_j\right) \right] - C(\psi^*)\Delta_J \quad \to \quad \mu \tag{63}$$

$$\frac{1}{J} \sum_{j=1}^{J} \operatorname{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^{(J)}\right) \quad \to \quad \Xi_{s,s'} \tag{64}$$

$$\frac{1}{J^{3/2}} \sum_{j=1}^{J} E_{Y_j}\left[\left| T(\theta_j) - M^{(1)}\left(\psi^* \mid Y_j\right) \right|^3 \mid Y_j, \psi^{(J)} \right] \quad \to \quad 0, \tag{65}$$

$Q^{(\infty)}_{\psi^*}$-almost surely as $J \to \infty$, with $s, s' = 1, \dots, S$. We prove the three above results sequentially below, which concludes the proof of (62). □

Proof of (63). For any $s = 1, \dots, S$, by (61) and the multivariate Taylor formula it holds

$$M_s^{(1)}\left(\psi^{(J)} \mid Y_j\right) - M_s^{(1)}\left(\psi^* \mid Y_j\right) = \sum_{d=1}^{D} \frac{\tilde{\psi}_d + \Delta_{J,d}}{\sqrt{J}} \partial_{\psi_d} M_s^{(1)}\left(\psi^* \mid Y_j\right) + R_2(Y_j),$$

where

$$R_2(Y_j) = \sum_{d,d'=1}^{D} \frac{(\tilde{\psi}_d + \Delta_{J,d})(\tilde{\psi}_{d'} + \Delta_{J,d'})}{J} \int_0^1 (1-t) \partial_{\psi_d} \partial_{\psi_{d'}} M_s^{(1)}\left(\psi^* + t\frac{\tilde{\psi} + \Delta_J}{\sqrt{J}} \mid Y_j\right) \mathrm{d}t.$$

Therefore

$$\begin{aligned} &\frac{1}{\sqrt{J}} \sum_{j=1}^{J} \left[M_s^{(1)}\left(\psi^{(J)} \mid Y_j\right) - M_s^{(1)}\left(\psi^* \mid Y_j\right) \right] = \\ &\qquad = \sum_{d=1}^{D} (\tilde{\psi}_d + \Delta_{J,d}) \frac{1}{J} \sum_{j=1}^{J} \partial_{\psi_d} M_s^{(1)}\left(\psi^* \mid Y_j\right) + \frac{1}{\sqrt{J}} \sum_{j=1}^{J} R_2(Y_j)\,, \end{aligned} \tag{66}$$

where

$$\begin{aligned} &\frac{1}{\sqrt{J}} \sum_{j=1}^{J} R_2(Y_j) = \\ &\sum_{d,d'=1}^{D} \frac{(\tilde{\psi}_d + \Delta_{J,d})(\tilde{\psi}_{d'} + \Delta_{J,d'})}{J^{1/4}} \frac{1}{J^{5/4}} \sum_{j=1}^{J} \int_0^1 (1-t) \partial_{\psi_d} \partial_{\psi_{d'}} M_s^{(1)}\left(\psi^* + t\frac{\tilde{\psi} + \Delta_J}{\sqrt{J}} \mid Y_j\right) \mathrm{d}t. \end{aligned} \tag{67}$$

As regards (67), for every $d, d' = 1, \dots, D$ by Lemma 57 it holds

$$\frac{(\tilde{\psi}_d + \Delta_{J,d})(\tilde{\psi}_{d'} + \Delta_{J,d'})}{J^{1/4}} = \frac{\tilde{\psi}_d \tilde{\psi}_{d'}}{J^{1/4}} + \tilde{\psi}_d \frac{\Delta_{J,d'}}{J^{1/4}} + \tilde{\psi}_{d'} \frac{\Delta_{J,d}}{J^{1/4}} + \frac{\Delta_{J,d}}{J^{1/8}} \frac{\Delta_{J,d'}}{J^{1/8}} \quad \to \quad 0, \tag{68}$$

$Q^{(\infty)}_{\psi^*}$-almost surely as $J \to \infty$. Moreover, with the change of variables $x = t/J^{1/4}$ we have

$$\begin{aligned}
&\left|\frac{1}{J^{5/4}}\sum_{j=1}^{J}\int_0^1 (1-t)\partial_{\psi_d}\partial_{\psi_{d'}} M_s^{(1)}\left(\psi^* + t\frac{\tilde{\psi}+\Delta_J}{\sqrt{J}} \mid Y_j\right)\,\mathrm{d}t\right| \\
&\leq \int_0^{J^{1/4}} \frac{1}{J}\sum_{j=1}^{J}\left|\partial_{\psi_d}\partial_{\psi_{d'}} M_s^{(1)}\left(\psi^* + x\frac{\tilde{\psi}+\Delta_J}{J^{1/4}} \mid Y_j\right)\right|\,\mathrm{d}x \\
&\leq \int_{-J^{1/4}}^{J^{1/4}} \frac{1}{J}\sum_{j=1}^{J}\left|\partial_{\psi_d}\partial_{\psi_{d'}} M_s^{(1)}\left(\psi^* + x \mid Y_j\right)\right|\,\mathrm{d}x,
\end{aligned}$$

where the last inequality follows from $\left|\frac{\tilde{\psi}+\Delta_J}{J^{1/4}}\right| \leq 1$ for J high enough, thanks to Lemma 57. Moreover, $\frac{1}{J^{1/4}} < \delta_4$ for J high enough, so that

$$\begin{aligned}
&\left|\frac{1}{J^{5/4}}\sum_{j=1}^{J}\int_0^1 (1-t)\partial_{\psi_d}\partial_{\psi_{d'}} M_s^{(1)}\left(\psi^* + t\frac{\tilde{\psi}+\Delta_J}{\sqrt{J}} \mid Y_j\right)\,\mathrm{d}t\right| \\
&\leq \int_{\delta_4}^{\delta_4} \frac{1}{J}\sum_{j=1}^{J}\left|\partial_{\psi_d}\partial_{\psi_{d'}} M_s^{(1)}\left(\psi^* + x \mid Y_j\right)\right|\,\mathrm{d}x \\
&= \frac{1}{J}\sum_{j=1}^{J}\int_{\delta_4}^{\delta_4} \left|\partial_{\psi_d}\partial_{\psi_{d'}} M_s^{(1)}\left(\psi^* + x \mid Y_j\right)\right|\,\mathrm{d}x.
\end{aligned}$$

By the Law of Large Numbers and $(B4)$ it holds

$$\begin{aligned}
&\frac{1}{J}\sum_{j=1}^{J}\int_{\delta_4}^{\delta_4} \left|\partial_{\psi_d}\partial_{\psi_{d'}} M_s^{(1)}\left(\psi^* + x \mid Y_j\right)\right|\,\mathrm{d}x \\
&\qquad \to \int_{-\delta_4}^{\delta_4} E\left[\left|\partial_{\psi_d}\partial_{\psi_{d'}} M_s^{(1)}\left(\psi^* + x \mid Y_j\right)\right|\right]\,\mathrm{d}x < 2C\delta_4. \qquad (69)
\end{aligned}$$

By combining (68) and (69), we can conclude

$$\left|\frac{1}{\sqrt{J}}\sum_{j=1}^{J} R_2(Y_j)\right| \quad \to \quad 0,$$

$Q^{(\infty)}_{\psi^*}$-almost surely as $J \to \infty$. As regards (66), by the Law of Large Numbers we have

$$\frac{1}{J}\sum_{j=1}^{J}\partial_{\psi_d} M_s^{(1)}\left(\psi^* \mid Y_j\right) \quad \to \quad E\left[\partial_{\psi_d} M_s^{(1)}\left(\psi^* \mid Y_j\right)\right] = C_{s,d}(\psi^*),$$

that is finite thanks to $(B4)$. Therefore, we can conclude that for any $s = 1, \dots, S$ we have

$$M_s^{(1)}\left(\psi^{(J)} \mid Y_j\right) - M_s^{(1)}\left(\psi^* \mid Y_j\right) - \sum_{d=1}^{D} C_{s,d}(\psi^*)\Delta_{J,d} \quad \to \quad \sum_{d=1}^{D} C_{s,d}(\psi^*)\tilde{\psi}_d,$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$ and thus (63) holds. □

Proof of (64). For every $s, s' = 1, \dots, S$ by multivariate Taylor formula it holds

$$\mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^{(J)}\right) = \mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^*\right) + R_{1,cov}(Y_j),$$

where

$$R_{1,cov}(Y_j) = \sum_{d=1}^{D} \frac{\tilde{\psi}_d + \Delta_{J,d}}{\sqrt{J}} \int_0^1 (1-t)\partial_{\psi_d}\mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^* + t\frac{\tilde{\psi} + \Delta_J}{\sqrt{J}}\right) \mathrm{d}t.$$

Notice that

$$\frac{1}{J}\sum_{j=1}^{J} R_{1,cov}(Y_j) = \sum_{d=1}^{D} \frac{\tilde{\psi}_d + \Delta_{J,d}}{J^{1/4}} \int_0^1 (1-t)\frac{1}{J^{5/4}}\sum_{j=1}^{J}\partial_{\psi_d}\mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^* + t\frac{\tilde{\psi} + \Delta_J}{\sqrt{J}}\right) \mathrm{d}t.$$

With the same arguments of before we have $\frac{\tilde{\psi}_d + \Delta_{J,d}}{J^{1/4}} \to 0$ and

$$\begin{aligned} &\left|\int_0^1 (1-t)\frac{1}{J^{5/4}}\sum_{j=1}^{J}\partial_{\psi_d}\mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^* + t\frac{\tilde{\psi} + \Delta_J}{\sqrt{J}}\right) \mathrm{d}t\right| \\ &\qquad \le \frac{1}{J}\sum_{j=1}^{J}\int_{-\delta_4}^{\delta_4} \left|\partial_{\psi_d}\mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^* + x\right)\right| \mathrm{d}x \\ &\qquad \to \quad \int_{-\delta_4}^{\delta_4} E\left[\left|\partial_{\psi_d}\mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^* + x\right)\right|\right] \mathrm{d}x \end{aligned}$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$. Notice that by $(B4)$ we have

$$\begin{aligned} &E\left[\left|\partial_{\psi_d}\mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^* + x\right)\right|\right] \\ &\quad \le E\left[\left|\partial_{\psi_d} M_{s,s'}^{(1)}\left(\psi^* + x \mid Y_j\right)\right|\right] + E\left[\left|\partial_{\psi_d}\left\{M_s^{(1)}\left(\psi^* + x \mid Y_j\right) M_{s'}^{(1)}\left(\psi^* + x \mid Y_j\right)\right\}\right|\right] \\ &\quad \le 2C, \end{aligned}$$

for every $x \in (-\delta_4, \delta_4)$. Therefore, we can conclude

$$\left|\frac{1}{J}\sum_{j=1}^{J} R_{1,cov}(Y_j)\right| \to 0,$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$. Thus, by the Law of Large Numbers we have

$$\frac{1}{J}\sum_{j=1}^{J}\mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^*\right) \quad \to \quad E\left[\mathrm{Cov}\left(T_s(\theta_j), T_{s'}(\theta_j) \mid Y_j, \psi^*\right)\right],$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$. □

Proof of (65). By Lemma 56 we have

$$\frac{1}{J^{3/2}}\sum_{j=1}^{J}E_{Y_j}\left[\left|T(\theta_j)-M^{(1)}\left(\psi^*\mid Y_j\right)\right|^3\mid Y_j,\psi^{(J)}\right]$$
$$\leq\frac{1}{\sqrt{J}}+C\frac{1}{J^{3/2}}\sum_{s=1}^{S}\sum_{j=1}^{J}M^{(6)}\left(\psi^{(J)}\mid Y_j\right)+C\frac{1}{J^{3/2}}\sum_{s=1}^{S}\sum_{j=1}^{J}\left[M^{(1)}\left(\psi^*\mid Y_j\right)\right]^6.$$

By Jensen inequality $\left[M^{(1)}\left(\psi^*\mid Y_j\right)\right]^6\leq M^{(6)}\left(\psi^*\mid Y_j\right)$ and by the Law of Large Numbers

$$\frac{1}{J}\sum_{s=1}^{S}\sum_{j=1}^{J}M^{(6)}\left(\psi^*\mid Y_j\right)\quad\rightarrow\quad\sum_{s=1}^{S}E\left[T_s^6(\theta_j)\mid\psi^*\right]<\infty$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J\rightarrow\infty$. Thus to prove (65) it suffices to show

$$\frac{1}{J^{3/2}}\sum_{s=1}^{S}\sum_{j=1}^{J}M^{(6)}\left(\psi^{(J)}\mid Y_j\right)\quad\rightarrow\quad 0$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J\rightarrow\infty$. For every $s=1,\dots,S$ by multivariate Taylor formula it holds

$$M_s^{(6)}\left(\psi^{(J)}\mid Y_j\right)=M_s^{(6)}\left(\psi^*\mid Y_j\right)+R_{1,6}(Y_j),$$

where

$$R_{1,6}(Y_j)=\sum_{d=1}^{D}\frac{\tilde{\psi}_d+\Delta_{J,d}}{\sqrt{J}}\int_0^1(1-t)\partial_{\psi_d}M_s^{(6)}\left(\psi^*+t\frac{\tilde{\psi}+\Delta_J}{\sqrt{J}}\mid Y_j\right)\mathrm{d}t.$$

Notice that

$$\frac{1}{J}\sum_{j=1}^{J}R_{1,6}(Y_j)=\sum_{d=1}^{D}\frac{\tilde{\psi}_d+\Delta_{J,d}}{J^{1/4}}\int_0^1(1-t)\frac{1}{J^{5/4}}\sum_{j=1}^{J}\partial_{\psi_d}M_s^{(6)}\left(\psi^*+t\frac{\tilde{\psi}+\Delta_J}{\sqrt{J}}\mid Y_j\right)\mathrm{d}t,$$

and with the same arguments of before we have $\frac{\tilde{\psi}_d+\Delta_{J,d}}{J^{1/4}}\rightarrow 0$ $Q_{\psi^*}^{(\infty)}$-almost surely as $J\rightarrow\infty$ and

$$\left|\int_0^1(1-t)\frac{1}{J^{5/4}}\sum_{j=1}^{J}\partial_{\psi_d}M_s^{(6)}\left(\psi^*+t\frac{\tilde{\psi}+\Delta_J}{\sqrt{J}}\mid Y_j\right)\mathrm{d}t\right|$$
$$\leq\frac{1}{J}\sum_{j=1}^{J}\int_{-\delta_4}^{\delta_4}\left|\partial_{\psi_d}M_s^{(6)}(\psi^*+x\mid Y_j)\right|\mathrm{d}x$$
$$\rightarrow\int_{-\delta_4}^{\delta_4}E\left[\left|\partial_{\psi_d}M_s^{(6)}(\psi^*+x\mid Y_j)\right|\right]\mathrm{d}x<2\delta_4 C,$$

by $(B4)$. Therefore, we can conclude

$$\left|\frac{1}{J}\sum_{j=1}^{J}R_{1,6}(Y_j)\right|\rightarrow 0,$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J \to \infty$. Moreover, by the Law of Large Numbers we have

$$\frac{1}{J}\sum_{j=1}^{J} M_s^{(6)}\left(\psi^* \mid Y_j\right) \quad \to \quad E\left[M_s^{(6)}\left(\psi^* \mid Y_j\right)\right] = E\left[T_s^6(\theta_j) \mid \psi^*\right],$$

by $(B1)$ and the definition of conditional expectation. Therefore

$$\frac{1}{J^{3/2}}\sum_{j=1}^{J} M_s^{(6)}\left(\psi^* + \frac{\tilde{\psi}_d + \Delta_{J,d}}{\sqrt{J}} \mid Y_j\right) \to 0,$$

from which (65) follows. □

Total variation convergence

We extend the weak convergence to total variation using characteristic functions, in particular exploiting the conditions in Lemma 61. Here we first state some other technical lemmas that will be required later on.

Lemma 59. *Let X be a $\mathbb{R}^S$-valued random vector with zero mean and characteristic function $\varphi_X(u)$. Then for every $u \in \mathbb{R}^S$*

$$\varphi_X(u) = 1 - \frac{1}{2}E\left[(u^\top X)^2\right] + \frac{\theta}{6}E\left[|u^\top X|^3\right],$$

for some $\theta = \theta(u) \in \mathbb{C}$ such that $|\theta| \leq 1$.

Proof. Taylor formula for the complex exponential reads

$$e^{ix} = 1 + ix - \frac{x^2}{2} + \frac{x^3}{6}e^{iz},$$

where $z \in \mathbb{C}$ is such that $0 \leq |z| \leq |x|$. By $x = u^\top X$, we have

$$\varphi_X(u) = 1 + iE\left[u^\top X\right] - \frac{1}{2}E\left[\left(u^\top X\right)^2\right] + \frac{\theta}{6}E\left[\left|u^\top X\right|^3\right],$$

with $\theta = e^{iz}$, recalling that $|e^{iz}| \leq 1$ for any z. The result follows from $E\left[u^\top X\right] = 0$. □

Lemma 60. *Let $X \in \mathbb{R}^S$ and $Y \in \mathbb{R}^S$ be independent random vectors with the same distribution. Then*

$$\varphi_{X-Y}(u) = \left|\varphi_X(u)\right|^2.$$

Proof. By independence we can write

$$\varphi_{X-Y}(u) = E\left[e^{iu^\top X}\right]E\left[e^{-iu^\top X}\right],$$

where

$$E\left[e^{iu^\top X}\right] = E\left[\cos u^\top X\right] + iE\left[\sin u^\top X\right] = a + ib,$$

for suitable a and b. Since $\cos x$ is even and $\sin x$ is odd, we can write

$$\left|\varphi_{X-Y}(u)\right| = \left|(a+ib)(a-ib)\right| = a^2 + b^2 = \left|\varphi_X(u)\right|^2$$

Since $X - Y$ has a symmetric density by construction $|\varphi_{X-Y}(u)| = \varphi_{X-Y}(u)$ and the result follows. □

Corollary 18. *Let X be a $\mathbb{R}^S$-valued random vector with characteristic function $\varphi_X(u)$. Then*

$$|\varphi_X(u)|^2 \leq e^{-u^\top Var(X)u + \frac{2|u|^3}{3}\left[1+C\sum_{s=1}^S E\left[X_i^6\right]\right]},$$

for $u \in \mathbb{R}^S$, where C is a finite constant independent of u.

Proof. Let Y be an independent copy of X. By Lemma 60, it holds

$$|\varphi_X(u)|^2 = \varphi_{X-Y}(u),$$

where $\varphi_{X-Y}(u)$ is a real function, since it is the characteristic function of a random variable with symmetric density. Therefore, by Lemma 59 it holds

$$\varphi_{X-Y}(u) = 1 - \frac{1}{2}E\left[(u^\top Z)^2\right] + \frac{\theta}{6}E\left[|u^\top Z|^3\right],$$

where $Z = X - Y$ and $\theta = \theta(u) \in \mathbb{R}$. Recalling that $e^x \geq 1 + x$ for every x, we have

$$\varphi_{X-Y}(u) \leq e^{-\frac{1}{2}E\left[(u^\top Z)^2\right]+\frac{\theta}{6}E\left[|u^\top Z|^3\right]}.$$

By Lemma 8.8 in Bhattacharya and Rao (2010) it holds

$$E\left[(u^\top Z)^2\right] = 2E\left[(u^\top X)^2\right] = 2u^\top \mathrm{Var}(X)u$$

and

$$E\left[(u^\top Z)^3\right] \leq 4E\left[(u^\top X)^3\right] \leq 4|u|^3 E\left[|X|^3\right].$$

Moreover by Lemma 56 we have

$$E\left[|X|^3\right] \leq 1 + C\sum_{s=1}^S E\left[X_i^6\right].$$

Therefore

$$\varphi_{X-Y}(u) \leq e^{-u^\top \mathrm{Var}(X)u + \frac{2|u|^3\theta}{3}\left[1+C\sum_{s=1}^S E\left[X_i^6\right]\right]}$$

and the result follows from $|\theta| \leq 1$. □

The following lemma is a minor variation of commonly used techniques to prove total variation Central Limit Theorems.

Lemma 61. *Let $(X_J)_{J\geq 1}$ and X be $\mathbb{R}^S$-valued random variables with characteristic functions $(\varphi_J)_{J\geq 1}$ and φ, respectively. Denote by $L^1(\mathbb{R}^S)$ the space of complex-valued integrable functions with domain $\mathbb{R}^S$. If*

(a) X_J converges weakly to X as $J \to \infty$

(b) φ belongs to $L^1(\mathbb{R}^S)$, i.e. $\int_{\mathbb{R}^S} |\varphi(t)|\, dt < \infty$

(c) $\lim_{A\to\infty} \limsup_{J\to\infty} \int_{|t|\geq A} |\varphi_J(t)|\, dt = 0$.

then X_J converges to X in total variation as $J \to \infty$.

Proof. First we prove that $\lim_{J\to\infty} \|\varphi_J - \varphi\|_{L^1} = 0$. By the triangle inequality, for every $A > 0$ we have

$$\|\varphi_J - \varphi\|_{L^1} \leq \int_{|t|<A} |\varphi_J(t) - \varphi(t)| \,\mathrm{d}t + \int_{|t|\geq A} |\varphi_J(t)| \,\mathrm{d}t + \int_{|t|\geq A} |\varphi(t)| \,\mathrm{d}t\,. \tag{70}$$

Since weak convergence implies pointwise convergence of characteristic functions, assumption (a) implies that $\varphi_J(t) \to \varphi(t)$ as $J \to \infty$ for every $t \in \mathbb{R}^S$. Thus by the Dominated Convergence Theorem and $|\varphi_J(t) - \varphi(t)| \leq |\varphi_J(t)| + |\varphi(t)| = 2$, we have $\int_{|t|<A} |\varphi_J(t) - \varphi(t)| \,\mathrm{d}t \to 0$ as $J \to \infty$ for every $A > 0$. It follows by (70) that

$$0 \leq \limsup_{J\to\infty} \|\varphi_J - \varphi\|_{L^1} \leq \int_{|t|\geq A} |\varphi(t)| \,\mathrm{d}t + \limsup_{J\to\infty} \int_{|t|\geq A} |\varphi_J(t)| \,\mathrm{d}t\,, \tag{71}$$

for every $A > 0$. By assumption (b) $\lim_{A\to\infty} \int_{|t|\geq A} |\varphi(t)| \,\mathrm{d}t = 0$. Combining with assumption (c), taking the limit $A \to \infty$ we obtain $\limsup_{J\to\infty} \|\varphi_J - \varphi\|_{L^1} \leq 0$ and thus $\lim_{J\to\infty} \|\varphi_J - \varphi\|_{L^1} = 0$.

Then, note that $\varphi \in L^1(\mathbb{R}^S)$ and $\|\varphi_J - \varphi\|_{L^1} \to 0$ as $J \to \infty$ imply $\varphi_J \in L^1(\mathbb{R}^S)$ eventually as $J \to \infty$, since by the triangle inequality

$$\|\varphi_J\|_{L^1} \leq \|\varphi_J - \varphi\|_{L^1} + \|\varphi\|_{L^1} < \infty$$

for J large enough. Thus, by the Inversion formula, for J large enough X_J and X admit density functions w.r.t. the Lebesgue measure, which can be written as $f_{X_J}(\boldsymbol{t}) = \frac{1}{(2\pi)^S} \int_{\mathbb{R}^S} e^{-i\boldsymbol{t}^\top t} \varphi_J(t) \,\mathrm{d}t$ and $f_X(\boldsymbol{t}) = \frac{1}{(2\pi)^S} \int_{\mathbb{R}^S} e^{-i\boldsymbol{t}^\top t} \varphi(t) \,\mathrm{d}t$. Thus

$$\begin{aligned} |f_{X_J}(\boldsymbol{t}) - f_X(\boldsymbol{t})| &= \left| \frac{1}{(2\pi)^S} \int_{\mathbb{R}^S} e^{-i\boldsymbol{t}^\top t} \varphi_J(t) \,\mathrm{d}t - \frac{1}{(2\pi)^S} \int_{\mathbb{R}^S} e^{-i\boldsymbol{t}^\top t} \varphi(t) \,\mathrm{d}t \right| \\ &\leq \int_{\mathbb{R}^S} \left| e^{-i\boldsymbol{t}^\top t} (\varphi_J(t) - \varphi(t)) \right| \,\mathrm{d}t \leq \|\varphi_J - \varphi\|_{L^1} \to 0 \end{aligned}$$

as $J \to \infty$ for every $\boldsymbol{t} \in \mathbb{R}^S$. By Scheffé Theorem, total variation convergence is implied by pointwise convergence of the densities. □

Proof of Lemma 46. Fix $\tilde{\psi} \in \mathbb{R}^D$ and denote $\mu = C(\psi^*)\tilde{\psi}$ and $\Xi = V(\psi^*)$. We will prove conditions (a), (b) and (c) of Lemma 61 to show that $\mathcal{L}(\mathrm{d}\tilde{\boldsymbol{T}} \mid Y_{1:J}, \tilde{\psi}) \overset{TV}{\to} N(\mu, \Xi)$ for $Q_{\psi^*}^{(\infty)}$-almost every Y as $J \to \infty$.

Condition (a) is shown in Proposition 58. Regarding condition (b), the characteristic function of the limiting distribution $N(\mu, \Xi)$ is $\varphi(t) = e^{i\mu^\top t - \frac{1}{2} t^\top \Xi t}$, which is integrable since Ξ is positive definite by (B4).

We now turn to condition (c). Let

$$\tilde{\varphi}(t \mid Y_{1:J}, \psi) = \mathbb{E}\left[e^{it^\top \tilde{\boldsymbol{T}}} \mid Y_{1:J}, \psi\right] \qquad\qquad t \in \mathbb{R}^S$$

be the characteristic function of $\mathcal{L}\left(\mathrm{d}\tilde{\boldsymbol{T}} \mid Y_{1:J}, \psi\right)$. Using the definition of $\tilde{\boldsymbol{T}}$ in (4.19), and the fact that $T_s(\theta_j)$ are conditionally independent given $\tilde{\psi}$, we can write $\tilde{\varphi}$ as

$$\tilde{\varphi}(t \mid Y_{1:J}, \tilde{\psi}) = e^{-it^\top \alpha_J} \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right),$$

where $\alpha_J = C(\psi^*)\Delta_J + \frac{1}{\sqrt{J}} \sum_{j=1}^{J} M^{(1)}(\psi^* \mid Y_j)$, $\varphi\left(t \mid Y_j, \psi\right) = E\left[e^{it^\top T(\theta_j)} \mid Y_j, \psi\right]$ as in the definition of (B5) and $\psi^{(J)}$ as in (61). Since $\alpha_J \in \mathbb{R}^S$ we have $|e^{-it^\top \alpha_J}| = 1$ and thus

$$|\tilde{\varphi}(t \mid Y_{1:J}, \psi)| = \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi\right) \right|. \tag{72}$$

For every $\epsilon > 0$, by (72) and the subadditivity of lim sup we have

$$\lim_{A\to\infty} \limsup_{J\to\infty} \int_{|t|>A} \left|\tilde{\varphi}(t \mid Y_{1:J}, \tilde{\psi})\right| \mathrm{d}t \leq$$
$$\lim_{A\to\infty} \limsup_{J\to\infty} \int_{A<|t|<\epsilon\sqrt{J}} \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right) \right| \mathrm{d}t + \limsup_{J\to\infty} \int_{|t|>\epsilon\sqrt{J}} \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right) \right| \mathrm{d}t.$$

Lemma 62 shows that the second lim sup in the last line is equal to 0 for every $\epsilon > 0$, while Lemma 63 shows that the $\lim_{A\to\infty} \limsup_{J\to\infty}$ term goes to 0 when ϵ is chosen as in (73). Thus condition (c) follows by taking ϵ as in (73) in the above inequality. □

Lemma 62. *Under the same setting and notation as in the proof of Lemma 46, for every $\epsilon > 0$ we have*

$$\limsup_{J\to\infty} \int_{|t|>\epsilon\sqrt{J}} \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right) \right| dt = 0$$

$Q_{\psi^*}^{(\infty)}$*-almost surely.*

Proof. Consider the change of variables $x = t/\sqrt{J}$. Then

$$\int_{|t|>\epsilon\sqrt{J}} \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right) \right| \mathrm{d}t = J^{S/2} \int_{|x|>\epsilon} \left| \prod_{j=1}^{J} \varphi\left(x \mid Y_j, \psi^{(J)}\right) \right| \mathrm{d}x.$$

Let k and B_{δ_5} be as in $(B5)$ and k' and B_{δ_6} be as in $(B6)$. Take J high enough so that $J \geq 2k$ as well as $\psi^{(J)} \in B := B_{\delta_5} \cap B_{\delta_6}$, so that

$$\int_{|x|>\epsilon} \left| \prod_{j=1}^{J} \varphi\left(x \mid Y_j, \psi^{(J)}\right) \right| \mathrm{d}x \leq \sup_{\psi\in B} \int_{|x|>\epsilon} \left| \prod_{j=1}^{2k} \varphi\left(x \mid Y_j, \psi\right) \right| \left| \prod_{j=2k+1}^{J} \varphi\left(x \mid Y_j, \psi\right) \right| \mathrm{d}x\,.$$

For every $a \in \mathbb{R}_+$ denote its integer part as $\lfloor a \rfloor$. By (B6), for every $\psi \in B$ we have

$$\left| \prod_{j=2k+1}^{J} \varphi\left(x \mid Y_j, \psi\right) \right| \leq \prod_{s=1}^{\lfloor \frac{J-2k}{k'} \rfloor} A_s \leq \phi(\epsilon)^{\lfloor \frac{J-2k}{k'} \rfloor}, \qquad \text{with } A_s = \left| \prod_{j=2k+1+(s-1)k'}^{2k+1+sk'} \varphi\left(x \mid Y_j, \psi\right) \right|$$

almost surely, where we exploited the fact that each A_s is distributed as $\varphi^{(k')}\left(t \mid Y_{1:k'}, \psi\right)$ in (B6). Therefore

$$\int_{|x|>\epsilon}\left|\prod_{j=1}^{J}\varphi\left(x \mid Y_j, \psi^{(J)}\right)\right| \mathrm{d}x \leq \phi(\epsilon)^{\lfloor\frac{J-2k}{k'}\rfloor}\sup_{\psi\in B}\int_{|x|>\epsilon}\left|\prod_{j=1}^{2k}\varphi\left(x \mid Y_j, \psi\right)\right| \mathrm{d}x.$$

almost surely. By Hölder Inequality and $(B5)$, we have

$$c = \sup_{\psi\in B}\int_{|x|>\epsilon}\left|\prod_{j=1}^{2k}\varphi\left(x \mid Y_j, \psi\right)\right| \mathrm{d}x \leq \sup_{\psi\in B}\int_{\mathbb{R}^S}\left|\prod_{j=1}^{2k}\varphi\left(x \mid Y_j, \psi\right)\right| \mathrm{d}x \leq$$
$$\left\{\sqrt{\sup_{\psi\in B}\int_{\mathbb{R}^S}\left|\prod_{j=1}^{k}\varphi\left(x \mid Y_j, \psi\right)\right|^2 \mathrm{d}x}\right\}\left\{\sqrt{\sup_{\psi\in B}\int_{\mathbb{R}^S}\left|\prod_{j=k+1}^{2k}\varphi\left(x \mid Y_j, \psi\right)\right|^2 \mathrm{d}x}\right\} < \infty,$$

almost surely. Therefore it holds

$$\int_{|t|>\epsilon\sqrt{J}}\left|\prod_{j=1}^{J}\varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right)\right| \mathrm{d}t \leq J^{S/2}\phi(\epsilon)^{\lfloor\frac{J-2k}{k'}\rfloor}c,$$

that goes to 0 as $J \to \infty$, since $\phi(\epsilon) < 1$ by $(B6)$. □

Lemma 63. *Under the same setting and notation as in the proof of Lemma 46, let $\lambda > 0$ be such that the matrix $V(\psi^*) - \lambda I$ is positive definite. Such λ can be found, since $V(\psi^*)$ is positive definite by (B4). Then, given*

$$\epsilon = \frac{\lambda}{1 + C\sum_{s=1}^{S} E\left[T_s(\theta_1)^6 \mid \psi^*\right]} \tag{73}$$

we have

$$\lim_{A\to\infty}\limsup_{J\to\infty}\int_{A<|t|<\epsilon\sqrt{J}}\left|\prod_{j=1}^{J}\varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right)\right| dt = 0$$

$Q_{\psi^*}^{(\infty)}$-almost surely.

Proof. By Corollary 18, we have

$$\left|\varphi(u \mid Y_j, \psi)\right|^2 \leq e^{-u^\top \mathrm{Var}\left(T(\theta_j)|Y_j,\psi\right)u + \frac{2|u|^3}{3}\left[1 + C\sum_{s=1}^{S} E\left[T_s(\theta_j)^6|Y_j,\psi\right]\right]},$$

for every $u \in \mathbb{R}^S$ and $\psi \in \mathbb{R}^D$. Therefore

$$\left|\prod_{j=1}^{J}\varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi\right)\right|^2 \leq e^{-t^\top \frac{1}{J}\sum_{j=1}^{J}\mathrm{Var}\left(T(\theta_j)|Y_j,\psi\right)t + \frac{2|t|^3}{3\sqrt{J}}\left[1 + C\frac{1}{J}\sum_{j=1}^{J}\sum_{s=1}^{S} E\left[T_s(\theta_j)^6|Y_j,\psi\right]\right]}. \tag{74}$$

Notice that in the proof of (65) we have shown through (B4) that

$$\frac{1}{J}\sum_{j=1}^{J} E\left[T_s(\theta_j)^6 \mid Y_j, \psi^{(J)}\right] \to E\left[T_s(\theta_1)^6 \mid \psi^*\right] \tag{75}$$

$Q^{(\infty)}_{\psi^*}$-almost surely as $J \to \infty$, for every $s = 1, \dots, S$. Thus, combining (73) and (75), for every $|t| \leq \epsilon\sqrt{J}$ we have

$$\left| e^{\frac{2|t|^3}{3\sqrt{J}}\left[1+C\frac{1}{J}\sum_{j=1}^{J}\sum_{s=1}^{S} E\left[T_s(\theta_j)^6|Y_j,\psi\right]\right]} \right|^2 \leq e^{\lambda t^\top t}, \tag{76}$$

almost surely for J high enough. Finally by (74) and (76)

$$\int_{A<|t|<\epsilon\sqrt{J}} \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right) \right| \mathrm{d}t \leq \int_{|t|>A} e^{-t^\top \Xi^{(J)} t}\,\mathrm{d}t\,, \tag{77}$$

with

$$\Xi^{(J)} = \frac{1}{J}\sum_{j=1}^{J} \mathrm{Var}\left(T(\theta_j) \mid Y_j, \psi^{(J)}\right) - \lambda I\,.$$

Since $\Xi^{(J)} \to V(\psi^*) - \lambda I$ by (64), and $V(\psi^*) - \lambda I$ is positive definite by definition of λ, by Dominated Convergence Theorem

$$\limsup_{J} \int_{A<|t|<\epsilon\sqrt{J}} \left| \prod_{j=1}^{J} \varphi\left(\frac{t}{\sqrt{J}} \mid Y_j, \psi^{(J)}\right) \right| \mathrm{d}t \leq \int_{|t|>A} e^{-t^\top (V(\psi^*)-\lambda I) t}\,\mathrm{d}t\,, \tag{78}$$

Since the right hand side of (78) is integrable the conclusion follows by taking $A \to \infty$. □

Proof of Theorem 20

We first need a technical lemma.

Lemma 64. *Let $\left\{Y^{(n)}\right\}_n$ be a sequence of random elements with state space $\mathcal{Y}^{(n)}$, such that $Y^{(n)} \sim Q^{(n)}$ with $Q^{(n)} \in \mathcal{P}\left(\mathcal{Y}^{(n)}\right)$. Let $\{\pi_n\}_n$ be a sequence of Markov kernels from $\mathcal{Y}^{(n)}$ to $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ and let $\pi \in \mathcal{P}(\mathcal{X})$. If*

$$\|\pi_{n,1}(\cdot) - \pi_1(\cdot)\|_{TV} \to 0 \quad \textit{and} \quad \|\pi_n(\cdot \mid x) - \pi(\cdot \mid x)\|_{TV} \to 0,\ \textit{for } \pi_1\textit{-almost every } x \in \mathcal{X}_1,$$

as $n \to \infty$ in $Q^{(n)}$-probability, where $\pi_{n,1}$ and π_1 are the marginal distributions on $\mathcal{X}_1$ of π_n and π respectively, then

$$\|\pi_n(\cdot) - \pi(\cdot)\|_{TV} \to 0,$$

as $n \to \infty$ in $Q^{(n)}$-probability

Proof. Let $f : \mathcal{X} \to [0, 1]$ be a measurable function. By the triangular inequality we have

$$\begin{aligned}
&\left| \int_{\mathcal{X}} f(x_1, x_2)\pi_n(\mathrm{d}x_1, \mathrm{d}x_2) - \int_{\mathcal{X}} f(x_1, x_2)\pi(\mathrm{d}x_1, \mathrm{d}x_2) \right| \leq \\
&\qquad \left| \int_{\mathcal{X}} f(x_1, x_2)\pi_n(\mathrm{d}x_2 \mid x_1)\pi_{n,1}(\mathrm{d}x_1) - \int_{\mathcal{X}} f(x_1, x_2)\pi_n(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1) \right| + \\
&\qquad \left| \int_{\mathcal{X}} f(x_1, x_2)\pi_n(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1) - \int_{\mathcal{X}} f(x_1, x_2)\pi(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1) \right|.
\end{aligned}$$

Notice that

$$\begin{aligned}\sup_f &\left|\int_{\mathcal{X}} f(x_1,x_2)\pi_n(\mathrm{d}x_2 \mid x_1)\pi_{n,1}(\mathrm{d}x_1) - \int_{\mathcal{X}} f(x_1,x_2)\pi_n(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1)\right| \\ &\leq \left\|\pi_{n,1}(\cdot) - \pi_1(\cdot)\right\|_{TV} \to 0,\end{aligned}$$

as $n \to \infty$ in $Q^{(n)}$-probability, by assumption. Moreover we have

$$\begin{aligned}\sup_f &\left|\int_{\mathcal{X}} f(x_1,x_2)\pi_n(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1) - \int_{\mathcal{X}} f(x_1,x_2)\pi(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1)\right| \leq \\ &\int_{\mathcal{X}_1} \sup_f \left|\int_{\mathcal{X}_2} f(x_1,x_2)\pi_n(\mathrm{d}x_2 \mid x_1) - \int_{\mathcal{X}_2} f(x_1,x_2)\pi(\mathrm{d}x_2 \mid x_1)\right| \pi_1(\mathrm{d}x_1).\end{aligned}$$

The integrand on the right hand side goes to 0 as $n \to \infty$ in $Q^{(n)}$-probability, by assumption. Therefore, by Dominated Convergence Theorem, we have

$$\sup_f \left|\int_{\mathcal{X}} f(x_1,x_2)\pi_n(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1) - \int_{\mathcal{X}} f(x_1,x_2)\pi(\mathrm{d}x_2 \mid x_1)\pi_1(\mathrm{d}x_1)\right| \to 0,$$

as $n \to \infty$ in $Q^{(n)}$-probability, as desired. □

Proof of Theorem 20. Lemma 45 shows that $\tilde{\psi}$ converges to a Normal distribution with zero mean and non-singular covariance matrix $\mathcal{I}^{-1}(\psi^*)$. Similarly, Lemma 46 shows that, conditional to every $\tilde{\psi}$, $\tilde{\boldsymbol{T}}$ converges to a Normal distribution with mean and variance (denoted by $E_\infty[\cdot]$ and $\mathrm{Var}_\infty(\cdot)$) given by

$$E_\infty[\tilde{\boldsymbol{T}} \mid \tilde{\psi}] = C(\psi^*)\tilde{\psi}, \quad \mathrm{Var}_\infty\left(\tilde{\boldsymbol{T}} \mid \tilde{\psi}\right) = V(\psi^*).$$

Therefore, by Lemma 64, we conclude that $\left(\tilde{\boldsymbol{T}}, \tilde{\psi}\right)$ converges in total variation to a $(S+D)$-dimensional Gaussian distribution $\tilde{\pi}$ with zero mean and covariance matrix Σ given by

$$\Sigma = \begin{bmatrix} \Sigma_{\tilde{\boldsymbol{T}}} & \Sigma_{\tilde{\psi}\tilde{\boldsymbol{T}}}^\top \\ \Sigma_{\tilde{\psi}\tilde{\boldsymbol{T}}} & \Sigma_{\tilde{\psi}} \end{bmatrix},$$

where $\Sigma_{\tilde{\psi}} = \mathcal{I}^{-1}(\psi^*) \in \mathbb{R}^{D\times D}$ and $\Sigma_{\tilde{\boldsymbol{T}}} \in \mathbb{R}^{S\times S}$ are the limiting variances of $\tilde{\psi}$ and $\tilde{\boldsymbol{T}}$, while $\Sigma_{\tilde{\psi}\tilde{\boldsymbol{T}}} \in \mathbb{R}^{D\times S}$ is the limiting covariance. Thus, thanks to standard properties of the multivariate Gaussian distribution, the determinant of Σ can be computed as

$$\begin{aligned}\det(\Sigma) &= \det(\Sigma_{\tilde{\psi}})\det\left(\Sigma_{\tilde{\boldsymbol{T}}} - \Sigma_{\tilde{\psi}\tilde{\boldsymbol{T}}}^\top \Sigma_{\tilde{\psi}}^{-1} \Sigma_{\tilde{\psi}\tilde{\boldsymbol{T}}}\right) = \det(\Sigma_{\tilde{\psi}})\det\left(\mathrm{Var}_\infty\left(\tilde{\boldsymbol{T}} \mid \tilde{\psi}\right)\right) \\ &= \det\left(\mathcal{I}^{-1}(\psi^*)\right)\det\left(V(\psi^*)\right),\end{aligned}$$

which implies that Σ is non singular. Indeed, $\det\left(\mathcal{I}^{-1}(\psi^*)\right) > 0$ by $(B3)$, while $\det\left(V(\psi^*)\right) > 0$ by (B4). Therefore, by Theorem 1 in Roberts and Sahu (1997), the Gibbs sampler on the limit Gaussian target has a strictly positive spectral gap. Moreover, since the Gibbs sampler in (4.15) has two blocks, by Lemma 54 we have $\tilde{t}_{mix}(\epsilon, M) < \infty$ for every M and ϵ: thus the result follows by Corollary 8. □

Proof of Proposition 23

Proof. Using the notation $E_\infty[\cdot]$,$\mathrm{Var}_\infty(\cdot)$ and $\mathrm{Cov}_\infty(\cdot,\cdot)$ for the limiting mean, variance and covariance, by Propositions 45 and 46 we have

$$E_\infty[\tilde{\psi}] = \mathbf{0}_D, \quad \mathrm{Var}_\infty(\tilde{\psi}) = \mathcal{I}^{-1}(\psi^*)$$

and

$$E_\infty[\tilde{\boldsymbol{T}} \mid \tilde{\psi}] = C(\psi^*)\tilde{\psi}, \quad \mathrm{Var}_\infty\left(\tilde{\boldsymbol{T}} \mid \tilde{\psi}\right) = V(\psi^*).$$

By standard properties of the multivariate Gaussian distribution we have

$$E_\infty[\tilde{\boldsymbol{T}}] = \mathbf{0}_S, \quad \mathrm{Cov}_\infty\left(\boldsymbol{T}, \tilde{\psi}\right) = C(\psi^*)\mathrm{Var}_\infty(\tilde{\psi}) = C(\psi^*)\mathcal{I}^{-1}(\psi^*)$$

and

$$\begin{aligned}\mathrm{Var}_\infty(\boldsymbol{T}) &= \mathrm{Var}_\infty\left(\tilde{\boldsymbol{T}} \mid \tilde{\psi}\right) + \mathrm{Cov}_\infty\left(\boldsymbol{T}, \tilde{\psi}\right)\mathrm{Var}_\infty^{-1}(\tilde{\psi})\mathrm{Cov}_\infty^\top\left(\boldsymbol{T}, \tilde{\psi}\right)\\ &= V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*),\end{aligned}$$

as desired. □

Proof of Corollary 10

We need three preliminary lemmas. The first one is a special version of well-known results (e.g. Roberts and Sahu (1997)).

Lemma 65. *The Gibbs sampler targeting the distribution in Proposition 23 can be written as*

$$\begin{bmatrix}\tilde{\boldsymbol{T}}^{(t)}\\ \tilde{\psi}^{(t)}\end{bmatrix} = B\begin{bmatrix}\tilde{\boldsymbol{T}}^{(t-1)}\\ \tilde{\psi}^{(t-1)}\end{bmatrix} + \begin{bmatrix}U_1\\ U_2\end{bmatrix},$$

where

$$B = \begin{bmatrix}\boldsymbol{0}_{S\times S} & C(\psi^*)\\ \\ \boldsymbol{0}_{D\times S} & \mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1}C(\psi^*)\end{bmatrix}$$

and

$$\begin{bmatrix}U_1\\ U_2\end{bmatrix} \sim N\left(\boldsymbol{0}_{S+D}, \Sigma - B\Sigma B^\top\right)$$

Proof. By Proposition 46 we have

$$E\left[\tilde{\boldsymbol{T}}^{(t)} \mid \tilde{\boldsymbol{T}}^{(t-1)}, \tilde{\psi}^{(t-1)}\right] = C(\psi^*)\tilde{\psi}^{(t-1)}.$$

Moreover, by Proposition 23 and standard properties of the multivariate Gaussian distribution, we have

$$\begin{aligned}&E\left[\tilde{\psi}^t \mid \tilde{\boldsymbol{T}}^{(t-1)}, \tilde{\psi}^{(t-1)}\right]\\ &\quad= E\left[\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1}\tilde{\boldsymbol{T}}^{(t)} \mid \tilde{\boldsymbol{T}}^{(t-1)}, \tilde{\psi}^{(t-1)}\right]\\ &\quad= \mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1}C(\psi^*)\tilde{\psi}^{(t-1)},\end{aligned}$$

as desired. □

Lemma 66. *Let*

$$M = \begin{bmatrix} \boldsymbol{O}_{S\times S} & A \\ \\ \boldsymbol{O}_{D\times S} & W \end{bmatrix},$$

with $A \in \mathbb{R}^{S\times D}$ and $W \in \mathbb{R}^{D\times D}$. Then M and W have the same non null eigenvalues.

Proof. Let $\mu \neq 0$ be an eigenvalue of M, with eigenvector $x = [x_S^\top, x_D^\top]^\top$. We have

$$Mx = \mu x \quad \Leftrightarrow \quad \begin{bmatrix} Ax_D \\ Wx_D \end{bmatrix} = \begin{bmatrix} \mu x_S \\ \mu x_D \end{bmatrix},$$

so that μ is an eigenvalue of W with eigenvector x_D. Indeed, x_D is different from the null vector, since $\mu \neq 0$.

Let $\lambda \neq 0$ be an eigenvalue of W with eigenvector x_D. Then

$$M \begin{bmatrix} \frac{Ax_D}{\lambda} \\ x_D \end{bmatrix} = \begin{bmatrix} Ax_D \\ Wx_D \end{bmatrix} = \lambda \begin{bmatrix} \frac{Ax_D}{\lambda} \\ x_D \end{bmatrix},$$

so that λ is an eigenvalue of M, with eigenvector

$$\begin{bmatrix} \frac{Ax_D}{\lambda} \\ x_D \end{bmatrix},$$

as desired. □

Lemma 67. *Let $A \in \mathbb{R}^{D\times S}$ and $B \in \mathbb{R}^{S\times D}$. Then the matrices AB and BA have the same non-null eigenvalues.*

Proof. Let $\lambda \neq 0$ be an eigenvalue of AB, with eigenvector $v \in \mathbb{R}^D$. Then

$$\lambda Bv = B(AB)v = (BA)Bv.$$

Since $Bv \neq \mathbf{0}$ we conclude that λ is an eigenvalue of BA with eigenvector Bv. □

Proof of Corollary 10. With B as in Lemma 65, by Theorem 1 in Roberts and Sahu (1997) the spectral gap of the Gibbs sampler with operator $\tilde{P}$ is given by

$$\mathrm{Gap}(\tilde{P}) = \min\left\{1 - |\lambda_i| \,:\, \lambda_i \text{ eigenvalue of } B\right\}$$

Thus, by Lemma 66, with $M := B$ and

$$W = \mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1} C(\psi^*),$$

we have

$$\mathrm{Gap}(\tilde{P}) = \min\left\{1 - |\lambda_i| \,:\, \lambda_i \text{ eigenvalue of } \mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1} C(\psi^*)\right\}.$$

By Lemma 67 with

$$A = \mathcal{I}^{-1}(\psi^*)C^\top(\psi^*), \quad B = \left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1} C(\psi^*)$$

we deduce

$$\mathrm{Gap}(\tilde{P}) = \min\left\{1 - |\lambda_i| \,:\, \lambda_i \text{ eigenvalue of } \left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1} C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}.$$

Notice that

$$\begin{aligned}&\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1} C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\\ &= I - \left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1} V(\psi^*).\end{aligned}$$

Since λ is an eigenvalue of A if and only if $1-\lambda$ is an eigenvalue of $I - A$, it follows that

$$\mathrm{Gap}(\tilde{P}) = \min\left\{1 - |1-\lambda_i|\,;\, \lambda_i \text{ eigenvalue of } \left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}^{-1} V(\psi^*)\right\}.$$

Moreover the eigenvalues of the inverse are the inverse of the eigenvalues, so that the rate of convergence is equal to

$$\mathrm{Gap}(\tilde{P}) = \min\left\{1 - \left|1 - \frac{1}{\lambda_i}\right|\,;\, \lambda_i \text{ eigenvalue of } V^{-1}(\psi^*)\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}\right\}.$$

Since

$$V^{-1}(\psi^*)\left\{V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\} = I + V^{-1}(\psi^*)C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*),$$

we have

$$\mathrm{Gap}(\tilde{P}) = \min\left\{1 - \left|1 - \frac{1}{1+\lambda_i}\right|\,;\, \lambda_i \text{ eigenvalue of } V^{-1}(\psi^*)C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}.$$

Moreover both $V^{-1}(\psi^*)$ and $C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)$ are positive semi-definite, so that also their product is positive semi-definite and has positive eigenvalues. Therefore we conclude

$$\mathrm{Gap}(\tilde{P}) = \min\left\{\frac{1}{1+\lambda_i}\,;\, \lambda_i \text{ eigenvalue of } V^{-1}(\psi^*)C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*)\right\}$$

and the result follows by Corollary 9. □

Proof of Corollary 11

We need a preliminary lemma, that we prove for self-containedness.

Lemma 68. *Let $p(\theta \mid \psi)$ be as in* (4.14)*. Then it holds*

$$E[T(\theta) \mid \psi] = \frac{\partial_\psi A(\psi)}{\partial_\psi \eta(\psi)}, \quad Var(T(\theta) \mid \psi) = \left\{\partial^2_\psi A(\psi) - \frac{\partial^2_\psi \eta(\psi)\partial_\psi A(\psi)}{\partial_\psi \eta(\psi)}\right\}\left[\partial_\psi \eta(\psi)\right]^{-2}.$$

Proof. Differentiating the following equality

$$1 = \int p(\theta \mid \psi)\,\mathrm{d}\theta, \tag{79}$$

by the regularity properties of the exponential family we get

$$0 = \int \partial_\psi p(\theta \mid \psi)\,\mathrm{d}\theta = \partial_\psi \eta(\psi) E[T(\theta) \mid \psi] + \partial_\psi A(\psi),$$

and the formula for the expected value follows. As regards the variance, differentiating (79) twice, we obtain

$$0 = \partial^2_\psi \eta(\psi) E[T(\theta) \mid \psi] - \partial^2_\psi A(\psi) + \Big[\partial_\psi \eta(\psi)\Big]^2 E[T^2(\theta) \mid \psi] - 2\Big[\partial_\psi \eta(\psi)\Big]^2 E^2[T(\theta) \mid \psi] + \Big[\partial_\psi A(\psi)\Big]^2.$$

Noticing that

$$\Big[\partial_\psi \eta(\psi)\Big]^2 E^2[T(\theta) \mid \psi] = \Big[\partial_\psi A(\psi)\Big]^2$$

and rearranging, we get

$$\partial^2_\psi A(\psi) - \partial^2_\psi \eta(\psi) E[T(\theta) \mid \psi] = \Big[\partial_\psi \eta(\psi)\Big]^2 \mathrm{Var}(T(\theta) \mid \psi),$$

from which the result follows. □

Proof of Corollary 11. By Corollary 10, we have

$$\gamma(\psi^*) = \frac{1}{1+\lambda} \quad \text{with } \lambda = \frac{C^2(\psi^*)}{V(\psi^*)\mathcal{I}(\psi^*)},$$

where

$$\begin{aligned} C(\psi) &= E_{Y_j}\Big[\partial_\psi E[T(\theta_j) \mid Y_j, \psi]\Big], \\ V(\psi) &= E_{Y_j}\left[\mathrm{Var}(T(\theta_j) \mid Y_j, \psi)\right], \\ \mathcal{I}(\psi) &= -E_{Y_j}\Big[\partial^2_\psi \log g(Y_j \mid \psi)\Big], \end{aligned}$$

with $g(y \mid \psi)$ as in (4.16). As regards $C(\psi)$, notice that

$$\begin{aligned} \partial_\psi E[T(\theta) \mid Y, \psi] =& \frac{\int T(\theta) f(Y \mid \theta) \partial_\psi p(\theta \mid \psi)\,\mathrm{d}\theta}{g(Y \mid \psi)} - \\ & \frac{\left[\int T(\theta) f(Y \mid \theta) p(\theta \mid \psi)\,\mathrm{d}\theta\right]\Big[\int f(Y \mid \theta) \partial_\psi p(\theta \mid \psi)\,\mathrm{d}\theta\Big]}{g^2(Y \mid \psi)} \\ =& \partial_\psi \eta(\psi) E\Big[T^2(\theta) \mid Y, \psi\Big] - \partial_\psi \eta(\psi) E^2\left[T(\theta) \mid Y, \psi\right] \\ =& \partial_\psi \eta(\psi) \mathrm{Var}\left(T(\theta) \mid Y, \psi\right). \end{aligned}$$

Therefore

$$C^2(\psi^*) = \Big[\partial_\psi \eta(\psi^*)\Big]^2 E^2_{Y_j}\Big[\mathrm{Var}\left(T(\theta_j) \mid Y_j, \psi^*\right)\Big]. \tag{80}$$

As regards $\mathcal{I}(\psi)$, notice that

$$\partial_\psi \log g(Y_j \mid \psi) = \frac{\int f(Y \mid \theta)\partial_\psi p(\theta \mid \psi)\,\mathrm{d}\theta}{g(Y \mid \psi)} = \partial_\psi \eta(\psi)\frac{\int T(\theta) f(Y \mid \theta) p(\theta \mid \psi)\,\mathrm{d}\theta}{g(Y \mid \psi)} - \partial_\psi A(\psi)$$

and

$$\begin{aligned}\partial^2_\psi \log g(Y_j \mid \psi) &= \partial^2_\psi \eta(\psi) E\left[T(\theta) \mid Y, \psi\right] - \partial^2_\psi A(\psi) + \partial_\psi \eta(\psi)\frac{\int T(\theta) f(Y \mid \theta)\partial_\psi p(\theta \mid \psi)\,\mathrm{d}\theta}{g(Y \mid \psi)} \\ &\quad - \partial_\psi \eta(\psi)\frac{\left[\int T(\theta) f(Y \mid \theta) p(\theta \mid \psi)\,\mathrm{d}\theta\right]\left[\int f(Y \mid \theta)\partial_\psi p(\theta \mid \psi)\,\mathrm{d}\theta\right]}{g^2(Y \mid \psi)} \\ &= \partial^2_\psi \eta(\psi) E\left[T(\theta) \mid Y, \psi\right] - \partial^2_\psi A(\psi) + \left[\partial_\psi \eta(\psi)\right]^2 \mathrm{Var}\left(T(\theta) \mid Y, \psi\right).\end{aligned}$$

Noticing that, by Lemma 68, we have

$$\begin{aligned}\partial^2_\psi \eta(\psi) E\left[T(\theta) \mid Y, \psi\right] - \partial^2_\psi A(\psi) &= \left\{\partial^2_\psi A(\psi) - \frac{\partial^2_\psi \eta(\psi)\partial_\psi A(\psi)}{\partial_\psi \eta(\psi)}\right\} \\ &= \left[\partial_\psi \eta(\psi)\right]^2 \mathrm{Var}\left(T(\theta) \mid \psi\right),\end{aligned}$$

we get

$$\begin{aligned}\mathcal{I}(\psi^*) &= \left[\partial_\psi \eta(\psi^*)\right]^2 \mathrm{Var}\left(T(\theta_j) \mid \psi^*\right) - \left[\partial_\psi \eta(\psi^*)\right]^2 E_{Y_j}\left[\mathrm{Var}\left(T(\theta_j) \mid Y_j, \psi^*\right)\right] \\ &= \left[\partial_\psi \eta(\psi^*)\right]^2 \mathrm{Var}_{Y_j}\left(E\left[T(\theta_j) \mid Y_j, \psi^*\right]\right),\end{aligned} \tag{81}$$

by the Law of Total Variance. Combining (80) and (81), it holds

$$\lambda = \frac{E^2_{Y_j}\left[\mathrm{Var}\left(T(\theta_j) \mid Y_j, \psi^*\right)\right]}{V(\psi^*)\mathrm{Var}_{Y_j}\left(E\left[T(\theta_j) \mid Y_j, \psi^*\right]\right)} = \frac{E_{Y_j}\left[\mathrm{Var}\left(T(\theta_j) \mid Y_j, \psi^*\right)\right]}{\mathrm{Var}_{Y_j}\left(E\left[T(\theta_j) \mid Y_j, \psi^*\right]\right)}.$$

The expression for $\gamma(\psi^*)$ follows by rearranging and applying the Law of Total Variance. □

Proof of Proposition 24

First of all notice that, by Bayes' Theorem, we have

$$\theta_j \mid Y_j, \mu, \tau_1 \overset{\text{ind.}}{\sim} N\left(m_j, (m\tau_0 + \tau_1)^{-1}\right), \tag{82}$$

where

$$m_j = \frac{m\tau_0}{m\tau_0 + \tau_1}\bar{Y}_j + \frac{\tau_1}{m\tau_0 + \tau_1}\mu.$$

Recall that by (B1) we have

$$Y_j \overset{\text{iid}}{\sim} g(\cdot \mid \psi^*) = N\left(\mu^*, (\tau_0^*)^{-1} I + (\tau_1^*)^{-1}\mathbb{H}\right),$$

so that

$$\bar{Y}_j = \frac{1}{m}\sum_{i=1}^{m} Y_{j,i} \overset{\text{iid}}{\sim} N\left(\mu^*, \frac{1}{\tau_1^*} + \frac{1}{m\tau_0^*}\right). \tag{83}$$

Moreover we need some preliminary lemmas.

Lemma 69. *Let $X \sim N(\nu, \sigma^2)$. Then*

$$E[X^p] = \sum_{i=0}^{p} \binom{p}{i} \nu^i \sigma^{p-i} E[Z^{p-i}],$$

where $Z \sim N(0,1)$ and

$$E[Z^s] = \begin{cases} 0 & \textit{if } s \textit{ is odd} \\ 2^{-s/2}\frac{s!}{(s/2)!} & \textit{if } s \textit{ is even} \end{cases}$$

Proof. The result follows by noticing $X = \nu + \sigma Z$ and applying Netwon's Binomial Theorem. □

Lemma 70. *Let A be $m \times m$ matrix such that $A = aI + b\mathbb{H}$, with $a \neq b$ and $a \neq (1-m)b$. Then $det(A) = [a+mb]a^{m-1}$ and $A^{-1} = \frac{1}{a}\mathbb{I} - \frac{b}{a(a+mb)}\mathbb{H}$.*

Proof. We start by the determinant

$$\det\begin{pmatrix} c & d & \cdots & d \\ d & c & \cdots & d \\ \vdots & \vdots & \ddots & \vdots \\ d & d & \cdots & c \end{pmatrix} = [c + (m-1)d]\det\begin{pmatrix} 1 & 1 & \cdots & 1 \\ d & c & \cdots & d \\ \vdots & \vdots & \ddots & \vdots \\ d & d & \cdots & c \end{pmatrix}$$

$$= [c + (m-1)d]\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & c-d & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c-d \end{pmatrix} = [c + (m-1)d](c-d)^{m-1},$$

where the first equality comes by adding to the first row all the others, while the second comes by subtracting the first row (scaled by d) from all the others. In our case $c = a + b$ and $d = b$, that is $\det(A) = [a+mb]a^{m-1}$, as desired. With our assumptions we get that the determinant is different from zero.

As regards the inverse we prove $A^{-1} = xI + y\mathbb{H}$ for suitable x and y. Indeed

$$(aI + b\mathbb{H})(xI + y\mathbb{H}) = axI + ay\mathbb{H} + bx\mathbb{H} + by\mathbb{H}^2 = axI + (ay + bx + mby)\mathbb{H}.$$

Setting the above equal to I, we obtain $x = 1/a$ and

$$ay + bx + mby = 0 \quad \Rightarrow \quad y(a+mb) = -\frac{b}{a} \quad \Rightarrow \quad y = -\frac{b}{a(a+mb)}$$

as desired. □

Lemma 71. *Consider the marginal likelihood as in* (4.23), *with* $\psi^* = (\mu^*, \tau_1^*, \tau_0^*)$. *Then we have*

$$\mathcal{I}(\psi^*) = \begin{pmatrix} \frac{m\tau_0^*\tau_1^*}{\tau_1^*+m\tau_0^*} & 0 & 0 \\ 0 & \frac{m^2(\tau_0^*)^2}{2(\tau_1^*)^2(\tau_1^*+m\tau_0^*)^2} & \frac{m}{2(\tau_1^*+m\tau_0^*)^2} \\ 0 & \frac{m}{2(\tau_1^*+m\tau_0^*)^2} & \frac{m-1}{2(\tau_0^*)^2} + \frac{(\tau_1^*)^2}{2(\tau_0^*)^2(\tau_1^*+m\tau_0^*)^2} \end{pmatrix} \tag{84}$$

Proof. The log–likelihood $l(\psi) = \log g(y \mid \psi)$ is given by

$$l(\mu, \tau_0, \tau_1) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\left(\det(\Sigma)\right) - \frac{1}{2}(Y_1 - \mu I)^t\Sigma^{-1}(Y_1 - \mu I),$$

with $\Sigma = \tau_0^{-1}I + \tau_1^{-1}\mathbb{H}$. By Lemma 70 with $a = \tau_0^{-1}$ and $b = \tau_1^{-1}$ we have

$$\det(\Sigma) = [\tau_0^{-1} + m\tau_1^{-1}](\tau_0^{-1})^{m-1}, \quad \Sigma^{-1} = \tau_0 I - \frac{\tau_0^2}{\tau_1 + m\tau_0}\mathbb{H}.$$

Thus, the log–likelihood becomes

$$\begin{aligned} l(\mu, \tau_0, \tau_1) = &-\frac{1}{2}\log 2\pi + \frac{m-1}{2}\log\tau_0 - \frac{1}{2}\log(\tau_0^{-1} + m\tau_1^{-1}) - \frac{\tau_0}{2}\sum_{i=1}^m (Y_{1,i} - \mu)^2 \\ &+ \frac{\tau_0^2}{2(\tau_1 + m\tau_0)}(Y_1 - \mu I)^t\mathbb{H}(Y_1 - \mu I). \end{aligned}$$

Rewriting the last expression we get

$$\begin{aligned} l(\mu, \tau_0, \tau_1) = &-\frac{1}{2}\log 2\pi + \frac{m-1}{2}\log\tau_0 - \frac{1}{2}\log(\tau_0^{-1} + m\tau_1^{-1}) - \frac{\tau_0}{2}\sum_{i=1}^m (Y_{1,i} - \mu)^2 \\ &+ \frac{\tau_0^2}{2(\tau_1 + m\tau_0)}\left(\sum_{i=1}^m (Y_{1,i} - \mu)\right)^2. \end{aligned}$$

The required derivatives are given by

$$\begin{aligned} &\frac{\partial^2 l}{\partial\mu^2} = -\frac{m\tau_0\tau_1}{\tau_1 + m\tau_0}, \quad \frac{\partial^2 l}{\partial\tau_1^2} = -\frac{m\tau_0(2\tau_1 + m\tau_0)}{2\tau_1^2(\tau_1 + m\tau_0)^2} + \frac{\tau_0^2}{(\tau_1 + m\tau_0)^3}\left(\sum_{i=1}^m (Y_{1,i} - \mu)\right)^2, \\ &\frac{\partial^2 l}{\partial\tau_0^2} = -\frac{m-1}{2\tau_0^2} - \frac{\tau_1(\tau_1 + 2m\tau_0)}{2\tau_0^2(\tau_1 + m\tau_0)^2} + \frac{(\tau_1 + m\tau_0)^2 - 2m\tau_0\tau_1 - m^2\tau_0^2}{(\tau_1 + m\tau_0)^3}\left(\sum_{i=1}^m (Y_{1,i} - \mu)\right)^2, \\ &\frac{\partial^2 l}{\partial\mu\partial\tau_0} = \sum_{i=1}^m (Y_{1,i} - \mu) - \frac{2m\tau_0\tau_1 + m^2\tau_0^2}{(\tau_1 + m\tau_0)^2}\sum_{i=1}^m (Y_{1,i} - \mu), \\ &\frac{\partial^2 l}{\partial\mu\partial\tau_1} = \frac{\tau_0^2}{(\tau_1 + m\tau_0)^2}\sum_{i=1}^m (Y_{1,i} - \mu), \quad \frac{\partial^2 l}{\partial\tau_0\partial\tau_1} = \frac{m}{2(\tau_1 + m\tau_0)^2} - \frac{\tau_0\tau_1}{(\tau_1 + m\tau_0)^3}\left(\sum_{i=1}^m (Y_{1,i} - \mu)\right)^2. \end{aligned}$$

The entries of the Fisher Information matrix reported in (84) can then be computed from the above expressions by taking expectations with respect to Y_1 and exploiting that

$$\begin{aligned}
&\mathbb{E}[Y_{1,i}-\mu]=0, \quad \mathbb{E}\left[(Y_{1,i}-\mu)^2\right]=Var(Y_{1,i}-\mu)=\frac{\tau_0+\tau_1}{\tau_0\tau_1},\\
&\mathbb{E}\left[\left(\sum_{i=1}^{m}(Y_{1,i}-\mu)\right)^2\right]=Var\left(\sum_{i=1}^{m}(Y_{1,i}-\mu)\right)=[1,\dots,1]Var(Y_1)[1,\dots,1]^t\\
&\qquad\qquad\qquad\qquad\qquad\quad =[1,\dots,1]\left(\tau_0^{-1}I+\tau_1^{-1}\mathbb{H}\right)[1,\dots,1]^t\\
&\qquad\qquad\qquad\qquad\qquad\quad =m\left(\frac{m\tau_0+\tau_1}{\tau_0\tau_1}\right).
\end{aligned}$$

Thus we can compute the entries of the Fisher Information matrix as

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial^2 l}{\partial\tau_0^2}\right]&=-\frac{m-1}{2\tau_0^2}-\frac{\tau_1(\tau_1+2m\tau_0)}{2\tau_0^2(\tau_1+m\tau_0)^2}+\frac{m(\tau_1+m\tau_0)^2-2m^2\tau_0\tau_1-m^3\tau_0^2}{\tau_0\tau_1(\tau_1+m\tau_0)^2}\\
&=-\frac{m-1}{2\tau_0^2}-\frac{\tau_1^2}{2\tau_0^2(\tau_1+m\tau_0)^2},\\
\mathbb{E}\left[\frac{\partial^2 l}{\partial\tau_1^2}\right]&=-\frac{m\tau_0(2\tau_1+m\tau_0)}{2\tau_1^2(\tau_1+m\tau_0)^2}+\frac{m\tau_0}{\tau_1(\tau_1+m\tau_0)^2}=-\frac{m^2\tau_0^2}{2\tau_1^2(\tau_1+m\tau_0)^2},\\
\mathbb{E}\left[\frac{\partial^2 l}{\partial\mu\partial\tau_0}\right]&=0, \quad \mathbb{E}\left[\frac{\partial^2 l}{\partial\mu\partial\tau_1}\right]=0,\\
\mathbb{E}\left[\frac{\partial^2 l}{\partial\tau_0\partial\tau_1}\right]&=\frac{m}{2(\tau_1+m\tau_0)^2}-\frac{m}{(\tau_1+m\tau_0)^2}=-\frac{m}{2(\tau_1+m\tau_0)^2},
\end{aligned}$$

as desired. □

Lemma 72. *Let $X\sim N(\nu,\sigma^2)$. Then*

$$\left|E\left[e^{i(aX^2+bX)}\right]\right|\le\frac{e^{-\frac{\sigma^2}{2}\frac{(2\nu a+b)^2}{1+4a^2\sigma^4}}}{(1+4a^2\sigma^4)^{1/4}},$$

for every $(a,b)\in\mathbb{R}_2$.

Proof. By definition of expectation we have

$$E\left[e^{i(aX^2+bX)}\right]=\int_{\mathbb{R}}e^{i(az^2+bz)}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(z-\nu)^2}{2\sigma^2}}\,\mathrm{d}z=\frac{e^{-\frac{\nu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\int_{\mathbb{R}}e^{-\frac{1}{2}\left[z^2\left(\frac{1}{\sigma^2}-2ia\right)-2z\left(\frac{\nu}{\sigma^2}+ib\right)\right]}\,\mathrm{d}z$$

Notice that

$$\begin{aligned}
z^2\left(\frac{1}{\sigma^2}-2ia\right)-2z\left(\frac{\nu}{\sigma^2}+ib\right)&=\left(\frac{1-2ia\sigma^2}{\sigma^2}\right)\left[z^2-2z\frac{\nu+ib\sigma^2}{1-2ia\sigma^2}+\left(\frac{\nu+ib\sigma^2}{1-2ia\sigma^2}\right)-\left(\frac{\nu+ib\sigma^2}{1-2ia\sigma^2}\right)^2\right]\\
&=\left(\frac{1-2ia\sigma^2}{\sigma^2}\right)\left(z-\frac{\nu+i\sigma^2 b}{1-2ia\sigma^2}\right)^2-\frac{(\nu+ib\sigma^2)^2}{\sigma^2(1-2ia\sigma^2)},
\end{aligned}$$

so that

$$\frac{1}{\sqrt{2\pi\sigma^2}}\int_{\mathbb{R}} e^{-\frac{1}{2}\left[z^2\left(\frac{1}{\sigma^2}-2ia\right)-2z\left(\frac{\nu}{\sigma^2}+ib\right)\right]}\,\mathrm{d}z = \frac{e^{\frac{(\nu+ib\sigma^2)^2}{2\sigma^2(1-2ia\sigma^2)}}}{\sqrt{1-2ia\sigma^2}}.$$

Finally, we get

$$E\left[e^{i(aX^2+bX)}\right] = e^{-\frac{\nu^2}{2\sigma^2}}\,\frac{e^{\frac{(\nu+ib\sigma^2)^2}{2\sigma^2(1-2ia\sigma^2)}}}{\sqrt{1-2ia\sigma^2}}. \tag{85}$$

With simple computations we obtain

$$\begin{aligned}\frac{(\nu+ib\sigma^2)^2}{2\sigma^2(1-2ia\sigma^2)} &= \frac{(\nu^2+2i\nu b\sigma^2-b^2\sigma^4)(1+2ia\sigma^2)}{2\sigma^2(1+4a^2\sigma^4)}\\ &= \frac{\nu^2+2i\nu b\sigma^2-b^2\sigma^4+2i\nu^2a\sigma^2-4\nu ab\sigma^2-2i\sigma^6ab^2}{2\sigma^2(1+4a^2\sigma^4)}\\ &= \frac{\nu^2+2i(\nu b\sigma^2+\nu^2a\sigma^2-\sigma^6ab^2)-4\nu ab\sigma^4-\sigma^4b^2}{2\sigma^2(1+4a^2\sigma^4)}.\end{aligned}$$

Thus, by (85) we can write

$$E\left[e^{i(aX^2+bX)}\right] = e^{-\frac{\nu^2}{2\sigma^2}}\,\frac{e^{\frac{(\nu+ib\sigma^2)^2}{2\sigma^2(1-2ia\sigma^2)}}}{\sqrt{1-2ia\sigma^2}},$$

that implies

$$\left|E\left[e^{i(aX^2+bX)}\right]\right| \leq \frac{e^{-\frac{4\nu^2a^2\sigma^4+4\nu ab\sigma^4+b^2\sigma^4}{2\sigma^2(1+4a^2\sigma^4)}}}{|\sqrt{1-2ia\sigma^2}|} = \frac{e^{-\frac{\sigma^2}{2}\frac{(2\nu a+b)^2}{1+4a^2\sigma^4}}}{\left(1+4a^2\sigma^4\right)^{1/4}},$$

as desired. □

Define

$$\psi = (\mu, \tau_1) \quad \text{and} \quad \boldsymbol{T} = \boldsymbol{T}(\boldsymbol{\theta}) = \left(\sum_{j=1}^{J}\theta_j, \sum_{j=1}^{J}(\theta_j-\mu^*)^2\right). \tag{86}$$

Next three lemmas show that assumptions $(B1)-(B6)$ are satisfied for $(\boldsymbol{T},\psi)$ as defined above.

Lemma 73. *Consider the setting of Proposition 24. Then assumptions* $(B1)-(B3)$ *are satisfied for* $(\boldsymbol{T},\psi)$ *as in* (86).

Proof. It is easy to show that assumption $(B1)$ is satisfied, with $g(\cdot)$ as in (4.23). As regards $(B2)$, suitable tests can be defined analogously to Lemma 55.

Finally, by Lemma 71, the Fisher Information is given by

$$\frac{m\tau_0^*\tau_1^*}{m\tau_0^*+\tau_1^*}$$

for $l=1$ and by

$$\begin{bmatrix} \frac{m\tau_0^*\tau_1^*}{m\tau_0^*+\tau_1^*} & 0 \\ 0 & \frac{m^2(\tau_0^*)^2}{2(\tau_1^*)^2(\tau_1^*+m\tau_0^*)^2} \end{bmatrix},$$

for $l=2,3$. Therefore $(B3)$ is satisfied for any ψ^*. □

Lemma 74. *Consider the setting of Proposition 24. Then assumption* $(B4)$ *is satisfied for* $(\boldsymbol{T}, \psi)$ *as in* (86).

Proof. Since $T(\theta_j) = (\theta_j, (\theta_j - \mu^*)^2)$ it holds

$$M_s^{(p)}(\mu, \tau_1 \mid Y_j) = E\left[\theta_j^{sp} \mid \mu, \tau_1\right], \quad M_{1,2}^{(1)}(\mu, \tau_1 \mid Y_j) = E\left[\theta_j(\theta_j^* - \mu^*)^2 \mid \mu, \tau_1\right].$$

By Lemma 69 and (82), we obtain

$$E\left[\theta_j^k \mid \mu, \tau_1\right] = \sum_{i=0}^{k} \binom{k}{i} \left(\frac{m\tau_0}{m\tau_0 + \tau_1}\bar{Y}_j + \frac{\tau_1}{m\tau_0 + \tau_1}\mu\right)^i \left(\frac{1}{m\tau_0 + \tau_1}\right)^{(k-i)/2} E[Z^{k-i}].$$

It is a finite sum of infinitely times differentiable terms (with respect to μ and τ_1). Moreover, for every $k \geq 1$, thanks to Lemma 69 and (83), $E_{Y_j}\left[|\bar{Y}_j|^k \mid \mu, \tau_1\right]$ is uniformly bounded over (μ, τ_1) belonging to a bounded set.

Therefore, choosing $\delta_4 < \tau_1^*$, it is easy to find $C < \infty$ that satisfies assumption $(B4)$. □

Lemma 75. *Consider the setting of Proposition 24. Then assumptions* $(B5)$ *and* $(B6)$ *are satisfied for* $(\boldsymbol{T}, \psi)$ *as in* (86).

Proof. Assume $\mu^* = 0$, the general case follows by similar calculations. Recall that the posterior distribution of θ_j is given by $N(m_j, \sigma^2)$, with m_j as in (82) and

$$\sigma^2 = \frac{1}{m\tau_0 + \tau_1}.$$

By Lemma 72 we have

$$\left|E\left[e^{i(t_1\theta_j + t_2\theta_j^2)} \mid Y_j, \mu, \tau_1\right]\right|^2 \leq \frac{e^{-\sigma^2\frac{(2m_jt_2+t_1)^2}{1+4t_2^2\sigma^4}}}{\left(1 + 4t_2^2\sigma^4\right)^{1/2}}. \tag{87}$$

Moreover, notice that

$$\int_{\mathbb{R}} e^{-c\sigma^2\frac{(2m_jt_2+t_1)^2}{1+4t_2^2\sigma^4}} \mathrm{d}t_1 = \sqrt{\frac{\pi}{c\sigma^2}}\sqrt{1 + 4t_2^2\sigma^4},$$

for any $c > 0$. Since θ_j are independent, given μ and τ_1, by Hölder inequality we write

$$\begin{aligned}
&\int_{\mathbb{R}_2} \left|E\left[e^{i(t_1\sum_{j=1}^3\theta_j + t_2\sum_{j=1}^3\theta_j^2)} \mid Y, \mu, \tau_1\right]\right|^2 \mathrm{d}t_1\mathrm{d}t_2 = \int_{\mathbb{R}_2} \prod_{j=1}^{3} \left|E\left[e^{i(t_1\theta_j + t_2\theta_j^2)} \mid Y_j, \mu, \tau_1\right]\right|^2 \mathrm{d}t_1\mathrm{d}t_2 \\
&\leq \int_{\mathbb{R}_2} \prod_{j=1}^{3} \frac{e^{-\sigma^2\frac{(2m_jt_2+t_1)^2}{1+4t_2^2\sigma^4}}}{\left(1 + 4t_2^2\sigma^4\right)^{1/2}} \mathrm{d}t_1\mathrm{d}t_2 = \int_{\mathbb{R}} \frac{1}{\left(1 + 4t_2^2\sigma^4\right)^{3/2}} \left(\int_{\mathbb{R}} \prod_{j=1}^{3} e^{-\sigma^2\frac{(2\nu_jt_2+t_1)^2}{1+4t_2^2\sigma^4}} \mathrm{d}t_1\right) \mathrm{d}t_2 \\
&\leq \int_{\mathbb{R}} \frac{1}{\left(1 + 4t_2^2\sigma^4\right)^{3/2}} \prod_{j=1}^{3} \left(\int_{\mathbb{R}} e^{-3\sigma^2\frac{(2\nu_jt_2+t_1)^2}{1+4t_2^2\sigma^4}} \mathrm{d}t_1\right)^{1/3} \mathrm{d}t_2 \\
&= \sqrt{\frac{\pi}{3\sigma^2}} \int_{\mathbb{R}} \frac{1}{1 + 4t_2^2\sigma^4} \mathrm{d}t_2.
\end{aligned}$$

Therefore

$$\int_{\mathbb{R}^2} \left|\varphi^{(3)}\left(t \mid Y, \psi\right)\right|^2 \mathrm{d}t \leq \sqrt{\frac{\pi}{3\sigma^2}} \int_{\mathbb{R}} \frac{1}{1+4t_2^2\sigma^4} \mathrm{d}t_2 < \infty,$$

where the right hand side does not depend on the data and it is a continuous function of μ and τ_1. This implies $(B5)$ is satisfied with $k = 3$.

As regards $(B6)$, by Lemma 72 if $t_2 \neq 0$ we have

$$|\varphi^{(1)}(t \mid Y_j, \mu, \tau_1)| \leq \frac{1}{\left(1+4t_2^2\sigma^4\right)^{1/4}},$$

while if $t_2 = 0$ then

$$|\varphi^{(1)}(t \mid Y_j, \mu, \tau_1)| \leq e^{-\frac{\sigma^2}{2}t_1^2}.$$

Therefore

$$|\varphi^{(1)}(t \mid Y_j, \mu, \tau_1)| \leq \max\left\{\frac{1}{\left(1+4t_2^2\sigma^4\right)^{1/4}}, e^{-\frac{\sigma^2}{2}t_1^2}\right\},$$

so that

$$\sup_{|t|>\epsilon} |\varphi^{(1)}(t \mid Y_j, \mu, \tau_1)| \leq \max\left\{\frac{1}{\left(1+\epsilon^2\sigma^4\right)^{1/4}}, e^{-\frac{\sigma^2}{8}\epsilon^2}\right\},$$

since at least one between t_1 and t_2 must be larger than $\epsilon/2$. Notice that the right hand side does not depend on Y_j and is strictly smaller than 1 for every triplet (μ, τ_1, τ_0). Since σ^2 is a continuous function of μ and τ_1, assumption $(B6)$ is satisfied by choosing $\delta_6 < \tau_1^*$ and $k' = 1$.

□

Proof of Proposition 24. The result for P_1 follows directly by Theorem 20, whose assumptions are satisfied by Lemmas 73, 74 and 75. As regards P_2 and P_3, they are not particular cases of Theorem 20, since the two operators are different by the one in (4.15). However, the result follows by very similar arguments, that we briefly summarize. Since by construction

$$\mathcal{L}\left(\mathrm{d}\psi \mid \boldsymbol{\theta}, Y_{1:J}\right) = \mathcal{L}\left(\mathrm{d}\psi \mid \boldsymbol{T}(\boldsymbol{\theta}), Y_{1:J}\right)$$

a direct analogue of Lemma 44 holds. Moreover, following the proof of Theorem 20, Lemmas 45, 46 and 64 hold for $\boldsymbol{T}$ in (86). Finally, Corollary 12 proves that the limiting spectral gaps associated to P_2 and P_3 are strictly positive: by Lemma 54 this implies $\tilde{t}_{mix}(\epsilon, M) < \infty$ for P_2, being a two-block Gibbs sampler. The same holds for P_3, since in the limit it can be reduced to a two-block Gibbs sampler, as it will be clear by the proof of Corollary 12. □

Proof of Corollary 12

We split the proof in two different cases.

Proof of Corollary 12 for $\gamma_1(\psi^*)$

Proof. By Corollary 11, the spectral gap is equal to

$$\gamma_1(\psi) = \frac{\mathrm{Var}_{Y_j}\left(E\left[\theta_j \mid \psi, Y_j\right]\right)}{\mathrm{Var}\left(\theta_j \mid \psi\right)}.$$

By (82) and (83) we have

$$\mathrm{Var}_{Y_j}\left(E\left[\theta_j \mid \psi, Y_j\right]\right) = \left(\frac{m\tau_0}{m\tau_0+\tau_1}\right)^2, \quad \mathrm{Var}\left(\bar{Y}_j\right) = \frac{m\tau_0}{\tau_1(m\tau_0+\tau_1)},$$

and $\mathrm{Var}\left(\theta_j \mid \psi\right) = \tau_1^{-1}$, that leads to

$$\gamma_1(\psi^*) = \frac{m\tau_0^*}{m\tau_0^*+\tau_1^*},$$

as desired. □

Proof of Corollary 12 for $\gamma_2(\psi^*)$ and $\gamma_3(\psi^*)$

We need a technical Lemma.

Lemma 76. *Consider the setting of Proposition 24. Then*

$$C(\psi^*) = \begin{bmatrix} \frac{\tau_1^*}{m\tau_0^*+\tau_1^*} & 0 \\ 0 & -\frac{\tau_1^*+2m\tau_0^*}{\tau_1^*(m\tau_0^*+\tau_1^*)^2} \end{bmatrix}, \quad V(\psi^*) = \begin{bmatrix} \frac{1}{m\tau_0^*+\tau_1^*} & 0 \\ 0 & \frac{2\tau_1^*+4m\tau_0^*}{\tau_1^*(m\tau_0^*+\tau_1^*)^2} \end{bmatrix},$$

with $C(\psi^*)$ and $V(\psi^*)$ as in (38).

Proof. Recall that, in the context of Proposition 24, we define $T_1(\theta_j) = \theta_j$ and $T_2(\theta_j) = (\theta_j - \mu^*)^2$. By (82) we have

$$\begin{aligned} E[T_1(\theta_j) \mid Y_j, \psi] &= \frac{m\tau_0}{m\tau_0+\tau_1}\bar{Y}_j + \frac{\tau_1}{m\tau_0+\tau_1}\mu, \\ E[T_2(\theta_j) \mid Y_j, \psi] &= \frac{1}{m\tau_0+\tau_1} + \left(\frac{m\tau_0}{m\tau_0+\tau_1}\bar{Y}_j + \frac{\tau_1}{m\tau_0+\tau_1}\mu - \mu^*\right)^2. \end{aligned}$$

Therefore we can compute $C(\psi^*)$ as

$$
\begin{aligned}
E_{Y_j}\left[\partial_\mu M_1(\psi^* \mid Y_j)\right] &= \frac{\tau_1^*}{m\tau_0^* + \tau_1^*}, \\
E_{Y_j}\left[\partial_\mu M_2(\psi^* \mid Y_j)\right] &= E_{Y_j}\left[\frac{2\tau_1^*}{m\tau_0^* + \tau_1^*}\left(\frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}\bar{Y}_j - \frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}\mu^*\right)\right] = 0, \\
E_{Y_j}\left[\partial_{\tau_1} M_1(\psi^* \mid Y_j)\right] &= E_{Y_j}\left[-\frac{m\tau_0^*}{(m\tau_0^* + \tau_1^*)^2}\bar{Y}_j + \frac{m\tau_0^*}{(m\tau_0^* + \tau_1^*)^2}\mu^*\right] = 0, \\
E_{Y_j}\left[\partial_{\tau_1} M_2(\psi^* \mid Y_j)\right] &= -\frac{1}{(m\tau_0^* + \tau_1^*)^2} + \\
&\quad E_{Y_j}\left[2\left(-\frac{m\tau_0^*}{(m\tau_0^* + \tau_1^*)^2}\bar{Y}_j + \frac{m\tau_0^*}{(m\tau_0^* + \tau_1^*)^2}\mu^*\right)\left(\frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}\bar{Y}_j - \frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}\mu^*\right)\right] \\
&= -\frac{1}{(m\tau_0^* + \tau_1^*)^2} - 2\frac{(m\tau_0^*)^2}{(m\tau_0^* + \tau_1^*)^3}E_{Y_j}\left[(\bar{Y}_j - \mu^*)^2\right] \\
&= -\frac{1}{(m\tau_0^* + \tau_1^*)^2} - 2\frac{m\tau_0^*}{\tau_1^*(m\tau_0^* + \tau_1^*)^2},
\end{aligned}
$$

by (83).

We now consider $V(\psi^*)$. Given $X \sim N(\mu, \sigma^2)$, we have

$$
\operatorname{Cov}(X, X^2) = 2\mu\sigma^2, \quad \operatorname{Var}(X^2) = 2\sigma^4 + 4\mu^2\sigma^2\,,
$$

which can be easily derived by computing the first four moments of X using Lemma 69, which are $E[X] = \mu$, $E[X^2] = \mu^2 + \sigma^2$, $E[X^3] = 3\mu\sigma^2 + \mu^3$ and $E[X^4] = 3\sigma^4 + 6\mu^2\sigma^2 + \mu^4$. By (82) we have

$$
\begin{aligned}
\operatorname{Var}(\theta_j \mid Y_j, \psi^*) &= \frac{1}{m\tau_0^* + \tau_1^*}, \\
\operatorname{Cov}(\theta_j, (\theta_j - \mu^*)^2 \mid Y_j, \psi^*) &= \operatorname{Cov}(\theta_j - \mu^*, (\theta_j - \mu^*)^2 \mid Y_j, \psi^*) = 2\frac{m_j - \mu^*}{m\tau_0^* + \tau_1^*}, \\
\operatorname{Var}((\theta_j - \mu^*)^2 \mid Y_j, \psi^*) &= \frac{2}{(m\tau_0^* + \tau_1^*)^2} + \frac{4}{m\tau_0^* + \tau_1^*}\,(m_j - \mu^*)^2 \\
&= \frac{2}{(m\tau_0^* + \tau_1^*)^2} + \frac{4}{m\tau_0^* + \tau_1^*}\left(\frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}(\bar{Y}_j - \mu^*) + \mu^*\right)^2.
\end{aligned}
$$

Therefore, we conclude

$$
E_{Y_j}\left[\operatorname{Cov}(\theta_j, \theta_j^2 \mid Y_j, \psi^*)\right] = 0
$$

and

$$
\begin{aligned}
E_{Y_j}\left[\operatorname{Var}(\theta_j^2 \mid Y_j, \psi^*)\right] &= \frac{2}{(m\tau_0^* + \tau_1^*)^2} + \frac{4}{m\tau_0^* + \tau_1^*} E_{Y_j}\left[\left(\frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}\right)^2 (\bar{Y_j} - \mu^*)^2\right] \\
&= \frac{2}{(m\tau_0^* + \tau_1^*)^2} + \frac{4m^2(\tau_0^*)^2}{(m\tau_0^* + \tau_1^*)^3} E_{Y_j}\left[(\bar{Y_j} - \mu^*)^2\right] \\
&= \frac{2}{(m\tau_0^* + \tau_1^*)^2} + \frac{4m\tau_0^*}{\tau_1^*(m\tau_0^* + \tau_1^*)^2},
\end{aligned}
$$

as desired. □

Lemma 77. *Consider the same assumptions of Proposition 24. Then*

$$
\left\| \mathcal{L}(d\tilde{\boldsymbol{T}}, d\tilde{\psi} \mid Y_{1:J}) - N(\mathbf{0}, \Sigma) \right\|_{TV} \to 0,
$$

as $J \to \infty$*, in* $Q_{\psi^*}^{(\infty)}$*-probability, where* $(\tilde{\boldsymbol{T}}, \tilde{\psi})$ *are derived by* (86) *with transformations* (4.17) *and* (4.19) *and where*

$$
\Sigma = \begin{bmatrix}
2\frac{\tau_1^*+2m\tau_0^*}{m^2(\tau_0^*)^2\tau_1^*} & 0 & -2\frac{\tau_1^*(\tau_1^*+2m\tau_0^*)}{m^2(\tau_0^*)^2} & 0 \\
0 & \frac{1}{m\tau_0^*} & 0 & \frac{1}{m\tau_0^*} \\
-2\frac{\tau_1^*(\tau_1^*+2m\tau_0^*)}{m^2(\tau_0^*)^2} & 0 & 2\frac{(\tau_1^*)^2(\tau_1^*+m\tau_0^*)^2}{m^2(\tau_0^*)^2} & 0 \\
0 & \frac{1}{m\tau_0^*} & 0 & \frac{m\tau_0^*+\tau_1^*}{m\tau_0^*\tau_1^*}
\end{bmatrix} \tag{88}
$$

Proof. The result follows by an argument similar to the proof of Proposition 23, where

$$
\Sigma = \begin{bmatrix}
V(\psi^*) + C(\psi^*)\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*) & C(\psi^*)\mathcal{I}^{-1}(\psi^*) \\
\mathcal{I}^{-1}(\psi^*)C^\top(\psi^*) & \mathcal{I}^{-1}(\psi^*)
\end{bmatrix}
$$

The entries of Σ can be computed through Lemmas 71 and 76. □

Proof of Corollary 12 for $\gamma_2(\psi^*)$ *and* $\gamma_3(\psi^*)$*.* Recall that P_2 is the transition kernel of the Gibbs sampler that alternates updates from $\mathcal{L}\left(\mathrm{d}\mu, \mathrm{d}\boldsymbol{\theta} \mid \tau_1, Y_{1:J}\right)$ and $\mathcal{L}\left(\mathrm{d}\tau_1 \mid \boldsymbol{\theta}, \mu, Y_{1:J}\right)$. Through the same reasoning of Lemma 44, the mixing times of P_2 are the same of the Gibbs sampler targeting $\mathcal{L}\left(\mathrm{d}\mu, \mathrm{d}\tau_1, \mathrm{d}\boldsymbol{T} \mid Y_{1:J}\right)$ by alternating updates from $\mathcal{L}\left(\mathrm{d}\mu, \mathrm{d}\boldsymbol{T} \mid \tau_1, Y_{1:J}\right)$ and $\mathcal{L}\left(\mathrm{d}\tau_1 \mid \mu, \boldsymbol{T}, Y_{1:J}\right)$. Indeed

$$
\mathcal{L}\left(\mathrm{d}\tau_1 \mid \mu, \boldsymbol{\theta}, Y_{1:J}\right) = \mathcal{L}\left(\mathrm{d}\tau_1 \mid \mu, \boldsymbol{T}(\boldsymbol{\theta}), Y_{1:J}\right).
$$

Therefore, by Corollary 9 $\gamma_2(\psi^*)$ is the spectral gap of the Gibbs sampler alternating updates from $\tilde{\mathcal{L}}\left(\mathrm{d}\tilde{\mu}, \mathrm{d}\tilde{\boldsymbol{T}}_1, \mathrm{d}\tilde{\boldsymbol{T}}_2 \mid \tilde{\tau}_1\right)$ and $\tilde{\mathcal{L}}\left(\mathrm{d}\tilde{\tau}_1 \mid \tilde{\mu}, \tilde{\boldsymbol{T}}_1, \tilde{\boldsymbol{T}}_2\right)$, where $\tilde{\mathcal{L}}(\cdot)$ is the law identified in Lemma 77. By inspection of the matrix (88), $\left(\tilde{\mu}, \tilde{\boldsymbol{T}}_1\right)$ is independent from $\tilde{\tau}_1$ and $\tilde{\boldsymbol{T}}_2$ according to $\tilde{\mathcal{L}}$, so that $\left(\tilde{\mu}, \tilde{\boldsymbol{T}}_1\right)$ is sampled independently from everything else at each iteration. Therefore by the same arguments of the proof of Corollary 10 we have

$$
\gamma_2(\psi^*) = 1 - \frac{\Sigma_{24}^2}{\Sigma_{22}\Sigma_{44}} = \left(\frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}\right)^2.
$$

Instead, recall that P_3 is the transition kernel of the Gibbs sampler that alternates updates from $\mathcal{L}\left(\mathrm{d}\boldsymbol{\theta} \mid \tau_1, Y_{1:J}\right)$, $\mathcal{L}\left(\mathrm{d}\mu \mid \boldsymbol{\theta}, \tau_1, Y_{1:J}\right)$ and $\mathcal{L}\left(\mathrm{d}\tau_1 \mid \boldsymbol{\theta}, \mu, Y_{1:J}\right)$. Reasoning as before, by Corollary 9 $\gamma_3(\psi^*)$ is the spectral gap of the Gibbs sampler alternating updates from $\tilde{\mathcal{L}}\left(\mathrm{d}\tilde{\boldsymbol{T}} \mid \tilde{\mu}, \tilde{\tau}_1\right)$, $\tilde{\mathcal{L}}\left(\mathrm{d}\tilde{\mu} \mid \tilde{\tau}_1, \tilde{\boldsymbol{T}}\right)$ and $\tilde{\mathcal{L}}\left(\mathrm{d}\tilde{\tau}_1 \mid \tilde{\mu}, \tilde{\boldsymbol{T}}\right)$, where $\tilde{\mathcal{L}}(\cdot)$ is the law identified in Lemma 77. By inspection of the matrix (88), the pair $(\tilde{\mu}, \tilde{\boldsymbol{T}}_1)$ is independent from $(\tilde{\tau}_1, \tilde{\boldsymbol{T}}_2)$, according to $\tilde{\mathcal{L}}$. By standard properties of the Gibbs samplers (e.g. Lemma 2 in Papaspiliopoulos et al. (2020)), the spectral gap is given by the minimum of the spectral gaps of the Gibbs samplers associated to the two pairs, i.e.

$$\gamma_3(\psi^*) = \min\left\{1 - \frac{\Sigma_{24}^2}{\Sigma_{22}\Sigma_{44}}, 1 - \frac{\Sigma_{13}^2}{\Sigma_{11}\Sigma_{33}}\right\} = \left(\frac{m\tau_0^*}{m\tau_0^* + \tau_1^*}\right)^2 .$$

Notice that the result of Lemma 54 holds even if P_3 has three blocks: indeed, by inspection of the matrix (88), $\tilde{\mu}$ and $\tilde{\tau}_1$ are independent according to $\tilde{\mathcal{L}}$, so that the updates $\tilde{\mathcal{L}}\left(\mathrm{d}\tilde{\mu} \mid \tilde{\tau}_1, \tilde{\boldsymbol{T}}\right)$ and $\tilde{\mathcal{L}}\left(\mathrm{d}\tilde{\tau}_1 \mid \tilde{\mu}, \tilde{\boldsymbol{T}}\right)$ can be equivalently seen as a single one. □

Proof of Lemma 47

Since it will be useful in the following, we denote

$$c(\mu, \tau) = \min_{r \in \{0,\dots,m\}} g(y_r \mid \mu, \tau) \,,$$

with $g(y_r \mid \mu, \tau)$ defined in (4.28). Notice that by construction, see e.g. (4.26), we have $0 < c(\mu, \tau) \leq 1$. Also, $g(y_r \mid \mu, \tau)$ is continuous w.r.t. (μ, τ) since it is defined in (4.28) as the integral of a bounded function, $\theta \mapsto f(y \mid \theta)$, with respect to the normal kernel which is continuous w.r.t. (μ, τ). It follows that also $c(\mu, \tau)$ is continuous, since it is the minimum of a finite number of continuous functions. Define

$$c := \inf_{(\mu,\tau) \in B} c(\mu, \tau) > 0 \tag{89}$$

where B is the largest of the three balls – namely B_{δ_4}, B_{δ_5} and B_{δ_6} – centered at $\psi^* = (\mu^*, \tau^*)$ defined in (B4), (B5) and (B6), respectively. The positivity of c follows from the continuity of $c(\mu, \tau)$ and the compactness of B.

Recall that $T(\theta_j) = \left(\theta_j, \theta_j^2\right)$. Thus we need three lemmas.

Lemma 78. *Consider the setting of Lemma 47. Then assumption* ($B4$) *is satisfied.*

Proof. First of all, consider $V(\psi^*)$, as defined in (38). For every $y = 0, \dots, m$, we have that the posterior distribution of θ_j admits a density with respect to the Lebesgue measure of the form

$$p(\theta_j \mid y, \mu, \tau) \propto f(y_r \mid \theta_j) N(\theta_j \mid \mu, \tau),$$

which implies that

$$\mathrm{Var}(\theta_j \mid y, \psi^*) > 0, \quad \mathrm{Var}(\theta_j^2 \mid y, \psi^*) > 0, \quad |\mathrm{Corr}(\theta_j, \theta_j^2 \mid y, \psi^*)| < 1.$$

Consequently $V(\psi^*)$ is a sum of positive definite matrices and is therefore non singular.

Secondly, let $s, p = 1, 2$. Then by Bayes' Theorem it follows

$$M_s^{(p)}(y_r \mid \mu, \tau) = \frac{\int_{\mathbb{R}} \theta^{sp} f(y_r \mid \theta) N(\theta \mid \mu, \tau^{-1}) \,\mathrm{d}\theta}{\int_{\mathbb{R}} f(y_r \mid \theta) N(\theta \mid \mu, \tau^{-1}) \,\mathrm{d}\theta}, \quad r = 0, \dots, m.$$

Therefore

$$|\partial_\mu M_1^{(p)}(y_r \mid \mu, \tau)| \leq \left| \frac{\int_{\mathbb{R}} \theta^p f(y_r \mid \theta) \partial_\mu N(\theta \mid \mu, \tau^{-1})\, \mathrm{d}\theta}{\int_{\mathbb{R}} f(y_r \mid \theta) N(\theta \mid \mu, \tau^{-1})\, \mathrm{d}\theta} \right| + \\ \left| \frac{\left(\int_{\mathbb{R}} \theta^p f(y_r \mid \theta) N(\theta \mid \mu, \tau^{-1})\, \mathrm{d}\theta\right) \left(\int_{\mathbb{R}} f(y_r \mid \theta) \partial_\mu N(\theta \mid \mu, \tau^{-1})\, \mathrm{d}\theta\right)}{\left(\int_{\mathbb{R}} f(y_r \mid \theta) N(\theta \mid \mu, \tau^{-1})\, \mathrm{d}\theta\right)^2} \right|.$$

By definition of c we have

$$\begin{aligned} |\partial_\mu M_1^{(p)}(y_r \mid \mu, \tau)| \leq & \frac{1}{c} \int_{\mathbb{R}} |\theta|^p \left|\partial_\mu N(\theta \mid \mu, \tau^{-1})\right| \mathrm{d}\theta + \\ & \frac{1}{c^2} \left(\int_{\mathbb{R}} |\theta|^p N(\theta \mid \mu, \tau^{-1})\, \mathrm{d}\theta\right) \left(\int_{\mathbb{R}} |\theta|^p \left|\partial_\mu N(\theta \mid \mu, \tau^{-1})\right| \mathrm{d}\theta\right) \\ = & \frac{\tau}{c} \int_{\mathbb{R}} |(\theta - \mu)\theta^p| N(\theta \mid \mu, \tau^{-1})\, \mathrm{d}\theta + \\ & \frac{\tau}{c^2} \left(\int_{\mathbb{R}} |(\theta - \mu)\theta|^p N(\theta \mid \mu, \tau^{-1})\, \mathrm{d}\theta\right) \left(\int_{\mathbb{R}} |\theta|^p f \left|N(\theta \mid \mu, \tau^{-1})\right| \mathrm{d}\theta\right). \end{aligned}$$

The right hand side does not depend on the data, so that

$$E_{Y_j}\left[|\partial_\mu M_1^{(p)}(y_r \mid \mu, \tau)|\right] \leq m\frac{\tau}{c} E[|(\theta_j - \mu)\theta_j^p| \mid \mu, \tau] + m\frac{\tau}{c^2} E[|(\theta_j - \mu)\theta_j^p| \mid \mu, \tau] E[|\theta_j|^p \mid \mu, \tau].$$

By the specification of model (4.27), the prior absolute moments are all finite and continuous function of μ and τ: therefore the right hand side is uniformly bounded for every bounded neighborhood of (μ^*, τ^*). Using a similar argument for all the other quantities involved, it is easy to see that assumption $(B4)$ holds for every $\delta_4 < \tau^*$. □

Lemma 79. *Consider the setting of Lemma 47. Then assumption* $(B5)$ *is satisfied with* $k = 5$.

Proof. Consider the random vector $X = (X_1, X_2) = (\sum_{j=1}^5 \theta_j, \sum_{j=1}^5 \theta_j^2)$. First of all we prove that X admits a density function with respect to the Lebesgue measure on $\mathbb{R}^2$, conditional to (μ, τ). By Lemma 72 and conditional independence of θ_j we have

$$\left|E\left[e^{i(t_1 X_1 + t_2 X_2)} \mid \mu, \tau_1\right]\right| \leq \frac{e^{-5\frac{\sigma^2}{2}\frac{(2\mu t_2 + t_1)^2}{1 + 4t_2^2\sigma^4}}}{\left(1 + 4t_2^2\sigma^4\right)^{5/4}},$$

where we denote $\sigma^2 = \tau^{-1}$, so that we can write

$$\begin{aligned} \int_{\mathbb{R}^2} |\varphi_X(t \mid \mu, \tau)|\, \mathrm{d}t = & \int_{\mathbb{R}^2} \left|E\left[e^{i(t_1 X_1 + t_2 \sum_{j=1}^3 X_2)} \mid Y, \mu, \tau_1\right]\right| \mathrm{d}t_1 \mathrm{d}t_2 \\ \leq & \int_{\mathbb{R}} \frac{1}{\left(1 + 4t_2^2\sigma^4\right)^{5/4}} \left(\int_{\mathbb{R}} e^{-5\frac{\sigma^2}{2}\frac{(2\mu t_2 + t_1)^2}{1 + 4t_2^2\sigma^4}} \mathrm{d}t_1\right) \mathrm{d}t_2 \\ = & \sqrt{\frac{2\pi}{5\sigma^2}} \int_{\mathbb{R}} \frac{1}{\left(1 + 4t_2^2\sigma^4\right)^{3/4}} \mathrm{d}t_2 < \infty. \end{aligned} \tag{90}$$

Therefore, by the Inversion Formula we have that X admits a density $p(x \mid \mu, \tau)$ with respect to the Lebesgue measure on $\mathbb{R}^2$. Thus, by Bayes' Theorem we can write

$$p(x \mid Y_{1:5}, \mu, \tau) = \frac{f(Y_{1:5} \mid x, \mu, \tau) p(x \mid \mu, \tau)}{\int_{\mathbb{R}^2} f(Y_{1:5} \mid x, \mu, \tau) p(x \mid \mu, \tau) \, \mathrm{d}x},$$

where $f(Y_{1:5} \mid x, \mu, \tau) = \int \prod_{j=1}^5 f(Y_j \mid \theta_j) \mathcal{L}(\mathrm{d}\theta_{1:5} \mid x, \mu, \tau)$. It is easy to see that $f(Y_{1:5} \mid x, \mu, \tau) \leq 1$ and

$$\int_{\mathbb{R}^2} f(Y_{1:5} \mid x, \mu, \tau) p(x \mid \mu, \tau) \, \mathrm{d}x = \prod_{j=1}^5 g(Y_j \mid \mu, \tau) \geq c^5,$$

for every $(\mu, \tau) \in B_{\delta_5}$, with δ_5 to be fixed. We can therefore conclude that

$$p(x \mid Y_{1:5}, \mu, \tau) \leq \frac{p(x \mid \mu, \tau)}{c^5}.$$

We can now apply the Plancherel identity to get

$$\int_{\mathbb{R}^2} \left| \varphi^{(5)}(t \mid Y, \mu, \tau) \right|^2 \mathrm{d}t = \int_{\mathbb{R}^2} p^2(x_1, x_2 \mid Y, \mu, \tau) \, \mathrm{d}x \leq \frac{1}{c^{10}} \int_{\mathbb{R}^2} p^2(x_1, x_2 \mid \mu, \tau) \, \mathrm{d}x.$$

Applying again the Plancherel identity we obtain

$$\int_{\mathbb{R}^2} \left| \varphi^{(5)}(t \mid Y, \mu, \tau) \right|^2 \mathrm{d}t \leq \frac{1}{c^{10}} \int_{\mathbb{R}^2} \left| \varphi_X(t \mid \mu, \tau) \right|^2 \mathrm{d}t \leq \frac{1}{c^{10}} \int_{\mathbb{R}^2} \left| \varphi_X(t \mid \mu, \tau) \right| \mathrm{d}t < \infty,$$

by (90) for every $\tau > 0$. Therefore assumption $(B5)$ follows with $\delta_5 < \tau^*$. □

Lemma 80. *Consider the setting of Lemma 47. Then assumption $(B6)$ is satisfied with $k' = 5$.*

Proof. As shown in the proof of Lemma 79, the vector $(\sum_{j=1}^5 \theta_j, \sum_{j=1}^5 \theta_j^2)$ admits a density with respect to the Lebesgue measure on $\mathbb{R}^2$, conditional to Y and (μ^*, τ^*). Therefore, by Lemma 4 in Chapter 15 of Feller (1970), $|\varphi^{(5)}(t \mid Y, \mu^*, \tau^*)| < 1$ for every $t = (t_1, t_2)$. Moreover, by Riemann-Lebesgue Lemma we have

$$|\varphi^{(5)}(t \mid Y, \mu^*, \tau^*)| \quad \to \quad 0,$$

as $|t| \to \infty$. We conclude

$$\sup_{|t| \geq \epsilon} \left| \varphi^{(5)}(t \mid Y, \mu^*, \tau^*) \right| < 1.$$

Let $\delta_6 > 0$ to be chosen later and $(\mu, \tau) \in B_{\delta_6}$. Then by Taylor formula we get

$$|\varphi^{(5)}(t \mid Y, \mu, \tau)|^2 = |\varphi^{(5)}(t \mid Y, \mu^*, \tau^*)|^2 + (\mu^* - \mu) \partial_\mu |\varphi^{(5)}(t \mid Y, \bar{\mu}, \bar{\tau})|^2 + (\tau^* - \tau) \partial_\tau |\varphi^{(5)}(t \mid Y, \bar{\mu}, \bar{\tau})|^2, \tag{91}$$

where $(\bar{\mu}, \bar{\tau}) \in B_{\delta_6}$. Notice that

$$|\varphi^{(5)}(t \mid Y, \mu, \tau)|^2 = \left(\int_{\mathbb{R}^3} \cos\left(t_1 \sum_{j=1}^{5} \theta_j + t_2 \sum_{j=1}^{5} \theta_j^2 \right) \left\{ \prod_{j=1}^{5} \frac{f(Y_j \mid \theta_j) N(\theta_j \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(Y_j \mid \psi_j) N(\psi_j \mid \mu, \tau^{-1}) \mathrm{d}\psi_j} \right\} \mathrm{d}\theta_{1:5} \right)^2$$
$$+ \left(\int_{\mathbb{R}^5} \sin\left(t_1 \sum_{j=1}^{5} \theta_j + t_2 \sum_{j=1}^{5} \theta_j^2 \right) \left\{ \prod_{j=1}^{5} \frac{f(Y_j \mid \theta_j) N(\theta_j \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(Y_j \mid \psi_j) N(\psi_j \mid \mu, \tau^{-1}) \mathrm{d}\psi_j} \right\} \mathrm{d}\theta_{1:5} \right)^2,$$

which implies

$$\left| \partial_\mu |\varphi^{(5)}(t \mid Y, \mu, \tau)|^2 \right| \leq 2 \left| \int_{\mathbb{R}^5} \cos\left(t_1 \sum_{j=1}^{5} \theta_j + t_2 \sum_{j=1}^{5} \theta_j^2 \right) \partial_\mu \left\{ \prod_{j=1}^{5} \frac{f(Y_j \mid \theta_j) N(\theta_j \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(Y_j \mid \psi_j) N(\psi_j \mid \mu, \tau^{-1}) \mathrm{d}\psi_j} \right\} \mathrm{d}\theta_{1:5} \right|$$
$$+ 2 \left| \int_{\mathbb{R}^5} \sin\left(t_1 \sum_{j=1}^{5} \theta_j + t_2 \sum_{j=1}^{5} \theta_j^2 \right) \partial_\mu \left\{ \prod_{j=1}^{5} \frac{f(Y_j \mid \theta_j) N(\theta_j \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(Y_j \mid \psi_j) N(\psi_j \mid \mu, \tau^{-1}) \mathrm{d}\psi_j} \right\} \mathrm{d}\theta_{1:5} \right|$$

and therefore

$$\begin{aligned} \left| \partial_\mu |\varphi^{(5)}(t \mid Y, \mu, \tau)|^2 \right| &\leq 4 \int_{\mathbb{R}^5} \left| \partial_\mu \left\{ \prod_{j=1}^{5} \frac{f(Y_j \mid \theta_j) N(\theta_j \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(Y_j \mid \psi_j) N(\psi_j \mid \mu, \tau^{-1}) \mathrm{d}\psi_j} \right\} \right| \mathrm{d}\theta_{1:5} \\ &= 4 \sum_{j=1}^{5} \int_{\mathbb{R}} \left| \partial_\mu \left\{ \frac{f(Y_j \mid \theta_j) N(\theta_j \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(Y_j \mid \psi_j) N(\psi_j \mid \mu, \tau^{-1}) \mathrm{d}\psi_j} \right\} \right| \mathrm{d}\theta_j. \end{aligned} \tag{92}$$

Moreover, for every $r = 0, \dots, m$, we have

$$\begin{aligned} &\left| \partial_\mu \left\{ \frac{f(y_r \mid \theta) N(\theta \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(y_r \mid \psi) N(\psi \mid \mu, \tau^{-1}) \mathrm{d}\psi} \right\} \right| \leq \left| \left\{ \frac{f(y_r \mid \theta) \partial_\mu N(\theta \mid \mu, \tau^{-1})}{\int_{\mathbb{R}} f(y_r \mid \psi) N(\psi \mid \mu, \tau^{-1}) \mathrm{d}\psi} \right\} \right| \\ &+ \left| \left\{ \frac{f(y_r \mid \theta) \partial_\mu N(\theta \mid \mu, \tau^{-1}) \left(\int_{\mathbb{R}} f(y_r \mid \psi) \partial_\mu N(\psi \mid \mu, \tau^{-1}) \mathrm{d}\psi \right)}{\left(\int_{\mathbb{R}} f(y_r \mid \psi) N(\psi \mid \mu, \tau^{-1}) \mathrm{d}\psi \right)^2} \right\} \right| \\ &\leq \frac{|\partial_\mu N(\theta \mid \mu, \tau^{-1})|}{c} + \frac{1}{c^2} |\partial_\mu N(\theta \mid \mu, \tau^{-1})| \left(\int_{\mathbb{R}} |\partial_\mu N(\psi \mid \mu, \tau^{-1})| \mathrm{d}\psi \right) \\ &= 2\tau \frac{|\theta - \mu| N(\theta \mid \mu, \tau)}{c} + \frac{4\tau^2}{c^2} |\theta - \mu| N(\theta \mid \mu, \tau^{-1}) \left(\int_{\mathbb{R}} |\psi - \mu| N(\psi \mid \mu, \tau^{-1}) \mathrm{d}\psi \right). \end{aligned}$$

Therefore, by (92) there exists $C(\delta_6) < \infty$ which does not depend on μ and τ such that

$$\begin{aligned} \left| \partial_\mu |\varphi^{(5)}(t \mid Y, \mu, \tau)|^2 \right| &\leq 40\tau \frac{\int_{\mathbb{R}} |\theta - \mu| N(\theta \mid \mu, \tau^{-1}) \, \mathrm{d}\theta}{c} + 80\tau^2 \left(\frac{\int_{\mathbb{R}} |\theta - \mu| N(\theta \mid \mu, \tau^{-1}) \, \mathrm{d}\theta}{c} \right)^2 \\ &\leq C(\delta_6), \end{aligned}$$

for every $(\mu, \tau) \in B_{\delta_6}$ Notice that $C(\delta_6)$ becomes smaller as δ_6 decreases. Similarly holds for

$\partial_\tau|\varphi^{(3)}(t \mid Y, \mu, \tau)|^2$, so that by (91) we have

$$\begin{aligned}|\varphi^{(5)}(t \mid Y, \mu, \tau)|^2 &\leq |\varphi^{(5)}(t \mid Y, \mu^*, \tau^*)|^2 + |\mu^* - \mu|C(\delta_6) + |\tau^* - \tau|C(\delta_6)\\ &\leq |\varphi^{(5)}(t \mid Y\mu^*, \tau^*)|^2 + 2\delta_6 C(\delta_6).\end{aligned}$$

Since $\sup_{|t|\geq\epsilon} |\varphi^{(5)}(t \mid Y, \mu^*, \tau^*)|^2 < 1$, by choosing δ_6 small enough we have

$$\sup_{(\mu,\tau)\in B_{\delta_6}} \sup_{|t|\geq\epsilon} |\varphi^{(5)}(t \mid Y, \mu, \tau)|^2 \leq \sup_{|t|\geq\epsilon} |\varphi^{(5)}(t \mid Y, \mu^*, \tau^*)|^2 + 2\delta_6 C(\delta_6) < 1,$$

and $(B6)$ is satisfied. □

Proof of Lemma 47. Assumption (B4) is satisfied by Lemma 78, assumption (B5) by Lemma 79 and assumption (B6) by Lemma 80. □

Proof of Proposition 25

Proof. Requirements $(B1)-(B3)$ of Theorem 20 are satisfied by assumption, while $(B4)-(B6)$ hold by Lemma 47. □

Proof of Corollary 13

Proof. The result is a direct consequence of Corollary 10. □

Statement and proof of Lemma 81

Let

$$f(y \mid \theta) = \binom{m}{y} \frac{e^{y\theta}}{(1+e^\theta)^m}, \tag{93}$$

where $y = 0, \dots, m$. It means that for each group, conditional to θ, m independent Bernoulli trials are performed, with probability of success given by $e^\theta/(1+e^\theta)$. The following Section is devoted to the proof of the following lemma.

Lemma 81. *Consider the setting of Proposition 25 with likelihood* (93)*. The Fisher Information Matrix $I(\mu, \tau)$ is non-singular if and only if $m \geq 2$, for every (μ, τ).*

First of all we need few preliminary results.

Lemma 82. *Consider the setting of Proposition 25 with likelihood* (93) *and fix* (μ, τ)*. Let $h(y \mid \mu, \tau) = \log g(y \mid \mu, \tau)$, with $g(\cdot)$ as in* (4.28)*. Then it holds*

$$E_Y\left[\frac{\partial}{\partial\mu} h(Y \mid \mu, \tau)\right] = E_Y\left[\frac{\partial}{\partial\tau} h(Y \mid \mu, \tau)\right] = 0$$

and

$$E_Y\left[\left(\frac{\partial}{\partial\mu} h(Y \mid \mu, \tau)\right)^2\right] < \infty, \quad E_Y\left[\left(\frac{\partial}{\partial\tau_1} h(Y \mid \mu, \tau)\right)^2\right] < \infty.$$

Moreover, for every $y = 0, \dots, m$ we have

$$\frac{\partial}{\partial \mu} g(y \mid, \mu, \tau) = \binom{m}{y} \int \frac{e^{y\theta}\left[y + ye^{\theta} - me^{\theta}\right]}{(1+e^{\theta})^{m+1}} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\, d\theta$$

and

$$\frac{\partial}{\partial \tau} g(y \mid, \mu, \tau) = -\binom{m}{y} \frac{1}{2\tau} \int (\theta - \mu) \frac{e^{y\theta}\left[y + ye^{\theta} - me^{\theta}\right]}{(1+e^{\theta})^{m+1}} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\, d\theta.$$

Proof. Through Dominated Convergence Theorem it is easy to verify that

$$\frac{\partial}{\partial \mu} g(y \mid \mu, \tau) = \binom{m}{y} \int \frac{e^{y\theta}}{(1+e^{\theta})^{m}} \frac{\partial}{\partial \mu} \left\{ \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2} \right\} \mathrm{d}\theta$$

and

$$\frac{\partial}{\partial \tau} g(y \mid \mu, \tau) = \binom{m}{y} \int \frac{e^{y\theta}}{(1+e^{\theta})^{m}} \frac{\partial}{\partial \tau} \left\{ \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2} \right\} \mathrm{d}\theta,$$

that is integrals and derivatives can be exchanged. Therefore

$$\frac{\partial}{\partial \mu} h(y \mid, \mu, \tau) = E\left[\theta - \mu \mid y, \mu, \tau\right], \quad \frac{\partial}{\partial \mu} h(y \mid, \mu, \tau) = \frac{1}{2\tau} - \frac{1}{2} E\left[(\theta - \mu)^2 \mid y, \mu, \tau\right]$$

and the statements on $h(y \mid \mu, \tau)$ easily follow. Moreover

$$\begin{aligned} \frac{\partial}{\partial \mu} g(y \mid \mu, \tau) &= \binom{m}{y} \int \frac{e^{y\theta}}{(1+e^{\theta})^{m}} (\theta - \mu) \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\, \mathrm{d}\theta \\ &= \binom{m}{y} \int \frac{e^{y\theta}\left[y + ye^{\theta} - me^{\theta}\right]}{(1+e^{\theta})^{m+1}} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\, \mathrm{d}\theta \end{aligned}$$

integrating by parts. Similarly

$$\begin{aligned} \frac{\partial}{\partial \tau} g(y \mid \mu, \tau) &= \binom{m}{y} \frac{1}{2\tau} \int \frac{e^{y\theta}}{(1+e^{\theta})^{m}} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\, \mathrm{d}\theta \\ &\quad - \binom{m}{y} \frac{1}{2} \int \frac{e^{y\theta}}{(1+e^{\theta})^{m}} (\theta - \mu)^2 \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\, \mathrm{d}\theta \\ &= -\binom{m}{y} \frac{1}{2\tau} \int (\theta - \mu) \frac{e^{y\theta}\left[y + ye^{\theta} - me^{\theta}\right]}{(1+e^{\theta})^{m+1}} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2}\, \mathrm{d}\theta. \end{aligned}$$

□

Lemma 83. *Consider the setting of Proposition 25 with likelihood* (93) *and let $y, y' \in \{0, 1, \dots, m\}$ be such that $y < y'$ and $m \geq 1$. Then*

$$E\left[\theta \mid y, \mu, \tau\right] < E\left[\theta \mid y', \mu, \tau\right]$$

for every (μ, τ_1).

Proof. Fix (μ, τ). Consider the function

$$r(x) = \frac{\int \theta \frac{e^{x\theta}}{(1+e^{\theta})^m} \sqrt{\frac{\tau_1}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2} \, \mathrm{d}\theta}{\int \frac{e^{x\theta}}{(1+e^{\theta})^m} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2} \, \mathrm{d}\theta}.$$

with $x \in (0, m)$. Notice that

$$r(y) = E\left[\theta \mid y, \mu, \tau\right] \quad \text{and} \quad r(y') = E\left[\theta \mid y', \mu, \tau\right].$$

Notice that

$$\frac{\mathrm{d}}{\mathrm{d}x} r(x) = \frac{\int \theta^2 \frac{e^{x\theta}}{(1+e^{\theta})^m} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2} \, \mathrm{d}\theta}{\int \frac{e^{x\theta}}{(1+e^{\theta})^m} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2} \, \mathrm{d}\theta} - \left[\frac{\int \theta \frac{e^{x\theta}}{(1+e^{\theta})^m} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2} \, \mathrm{d}\theta}{\int \frac{e^{x\theta}}{(1+e^{\theta})^m} \sqrt{\frac{\tau}{2\pi}} e^{-\frac{\tau}{2}(\theta-\mu)^2} \, \mathrm{d}\theta} \right]^2 > 0$$

for every $x \in (0, m)$ by Jensen inequality. Therefore $r(x)$ is strictly increasing and $r(y) < r(y')$. □

Lemma 84. *Consider the setting of Proposition 25 with likelihood* (93)*. Then the Fisher Information Matrix $I(\mu, \tau)$ is non-singular in (μ, τ) if and only if there exists $\alpha = \alpha(\mu, \tau) \neq 0$ such that*

$$\frac{\partial}{\partial \mu} g(y \mid \mu, \tau) = \alpha \frac{\partial}{\partial \tau} g(y \mid \mu, \tau)$$

for every $y = 0, \ldots, m$.

Proof. Fix a pair (μ, τ). By Lemma 82 the matrix $I(\mu, \tau)$ is well-defined. The determinant is given by

$$E_Y\left[\left(\frac{\partial}{\partial \mu} h(Y \mid \mu, \tau)\right)^2\right] E_Y\left[\left(\frac{\partial}{\partial \tau} h(Y \mid \mu, \tau)\right)^2\right] - E^2\left[\left(\frac{\partial}{\partial \mu} h(Y \mid \mu, \tau)\right)\left(\frac{\partial}{\partial \tau} h(Y \mid \mu, \tau)\right)\right].$$

By Cauchy–Schwartz inequality, the above formula is always non-negative and it is equal to 0 if and only if $\frac{\partial}{\partial \mu} h(Y \mid \mu, \tau)$ and $\frac{\partial}{\partial \tau} h(Y \mid \mu, \tau)$ are linearly dependent, that is

$$\frac{\partial}{\partial \mu} h(y \mid \mu, \tau) = \alpha \frac{\partial}{\partial \tau} h(y \mid \mu, \tau) + \beta \tag{94}$$

for every $y \in \{0, 1, \ldots, m\}$ and for constants α and β. By Lemma 82 it is immediate to prove $\beta = 0$. Moreover, by Lemma 83, we deduce that $\alpha \neq 0$. Multiplying by $g(y \mid \mu, \tau)$ on both sides of (94) we get the final result. □

Proof of Lemma 81. Fix (μ, τ) and let $m = 1$. Define

$$\alpha := \frac{\frac{\partial}{\partial \mu} g(0 \mid \mu, \tau)}{\frac{\partial}{\partial \tau} g(0 \mid \mu, \tau)}.$$

Notice that α is well defined, since $\frac{\partial}{\partial \tau} g(0 \mid \mu, \tau) \neq 0$ for every (μ, τ). Then by construction

$$\frac{\partial}{\partial \mu} g(0 \mid \mu, \tau) = \alpha \frac{\partial}{\partial \tau} g(0 \mid \mu, \tau)$$

and

$$\frac{\partial}{\partial\mu}g(1 \mid \mu,\tau) = -\frac{\partial}{\partial\mu}g(0 \mid \mu,\tau) = -\alpha\frac{\partial}{\partial\tau}g(0 \mid \mu,\tau) = \alpha\frac{\partial}{\partial\tau}g(1 \mid \mu,\tau),$$

so that the Fisher Information matrix is singular by Lemma 84.

Let $m \geq 2$ and fix (μ,τ). Assume by contradiction that $I(\mu,\tau)$ is singular. By Lemma 84 we have that there exists $\alpha \neq 0$ such that

$$\frac{\partial}{\partial\mu}g(y \mid \mu,\tau) = \alpha\frac{\partial}{\partial\tau}g(y \mid \mu,\tau)$$

for every $y \in \{0,1,\ldots,m\}$. By the second part of Lemma 82 for $y = 0$ and $y = m$ it implies

$$-m\int\frac{e^{\theta}}{(1+e^{\theta})^{m+1}}\sqrt{\frac{\tau_1}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta = \alpha\frac{m}{2\tau}\int(\theta-\mu)\frac{e^{\theta}}{(1+e^{\theta})^{m+1}}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta$$

and

$$m\int\frac{e^{m\theta}}{(1+e^{\theta})^{m+1}}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta = -\alpha\frac{m}{2\tau}\int(\theta-\mu)\frac{e^{m\theta}}{(1+e^{\theta})^{m+1}}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta.$$

Since $\alpha \neq 0$, we conclude

$$\frac{\int(\theta-\mu)\frac{e^{m\theta}}{(1+e^{\theta})^{m+1}}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta}{\int\frac{e^{m\theta}}{(1+e^{\theta})^{m+1}}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta} = \frac{\int(\theta-\mu)\frac{e^{\theta}}{(1+e^{\theta})^{m+1}}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta}{\int\frac{e^{\theta}}{(1+e^{\theta})^{m+1}}\sqrt{\frac{\tau}{2\pi}}e^{-\frac{\tau}{2}(\theta-\mu)^2}\,\mathrm{d}\theta},$$

that means

$$E[\theta \mid m,\mu,\tau] = E[\theta \mid 1,\mu,\tau].$$

Since $m > 1$, the above equality directly contradicts Lemma 83. Therefore the Fisher Information matrix is non singular. □

Proof of Proposition 26

Define a one-to-one transformation of $\psi = (\mu,\tau_1,\tau_0)$ as

$$\tilde{\psi} = \sqrt{J}\,(\psi-\psi^*) - \Delta_J, \quad \Delta_J = \frac{1}{\sqrt{J}}\sum_{j=1}^{J}\mathcal{I}^{-1}(\psi^*)\nabla\log g(Y_j \mid \psi^*), \tag{95}$$

with $g(\cdot)$ as in (4.23) and $\mathcal{I}(\psi^*)$ as in (84).

Lemma 85. *Consider the assumptions of Proposition 26. Then it holds*

$$\left|\left|\mathcal{L}(d\tilde{\psi} \mid Y_{1:J}) - N\left(\mathbf{0},\mathcal{I}^{-1}(\psi^*)\right)\right|\right|_{TV} \to 0,$$

as $J \to \infty$ in $Q_{\psi^*}^{(\infty)}$-probability, with $\mathcal{I}(\psi^*)$ non singular matrix as in (84).

Proof. The result follows by Theorem 19. Indeed, the map $\psi \to g(y \mid \psi)$ clearly satisfies identifiability and smoothness requirements. Moreover, by Lemma 71 we have

$$\det\left(\mathcal{I}(\psi^*)\right) = \frac{m^3(m-1)\tau_0^*}{4\tau_1^*(\tau_1^*+m\tau_0^*)^3},$$

that is strictly positive for every ψ^*, with $m \geq 2$. As regards the testing conditions, analogously to Lemma 55 define

$$\Psi = \Psi_1 \times \Psi_2 \times \Psi_3 = [\mu^*-1, \mu^*+1] \times \left[\frac{\tau_1^*}{2}, 2\tau_1^*\right] \times \left[\frac{\tau_0^*}{2}, 2\tau_0^*\right]$$

compact neighborhood of ψ^* and

$$u_J(Y_{1:J}) = 1 - \mathbb{1}_{g_1(Y_{1:J})\leq c_1}\,\mathbb{1}_{g_2(Y_{1:J})\leq c_2}\,\mathbb{1}_{g_3(Y_{1:J})\leq c_3},$$

where (c_1, c_2, c_3) are positive constants to be fixed and

$$g_1(Y_{1:J}) = \left|\bar{Y} - \mu^*\right|, \quad g_2(Y_{1:J}) = \left|\frac{1}{J}\sum_{j=1}^J \left(\bar{Y}_j - \bar{Y}\right)^2 - \frac{1}{\tau_1^*} - \frac{1}{m\tau_0^*}\right|,$$
$$g_3(Y_{1:J}) = \left|\frac{1}{J}\sum_{j=1}^J \left(Y_{j,1} - \hat{Y}_1\right)\left(Y_{j,2} - \hat{Y}_2\right) - \frac{1}{\tau_1^*}\right|,$$

where

$$\bar{Y} = \frac{1}{J}\sum_{j=1}^J \bar{Y}_j, \quad \hat{Y}_i = \frac{1}{J}\sum_{j=1}^J Y_{j,i}.$$

By definition of $g(\cdot)$ in (4.23), by the Law of Large numbers we have

$$\int u_J(y_{1:J}) \prod_{j=1}^J g(\mathrm{d}y_j \mid \psi^*)$$
$$\leq P\left(g_1(Y_{1:J}) > c_1\right) + P\left(g_2(Y_{1:J}) > c_2\right) + P\left(g_3(Y_{1:J}) > c_3\right) \to 0,$$

as $J \to \infty$ for every strictly positive constants (c_1, c_2, c_3). Moreover, notice that

$$\sup_{\psi\notin\Psi} \int [1 - u_J(y_{1:J})] \prod_{j=1}^J g(\mathrm{d}y_j \mid \psi)$$
$$\leq \sup_{\tau_1\notin\Psi_2} P\left(g_3(Y_{1:J}) \leq c_3\right) + \sup_{\tau_1\in\Psi_2,\,\tau_0\notin\Psi_3} P\left(g_2(Y_{1:J}) \leq c_2\right) + \sup_{\mu\notin\Psi_1,\,\tau_0\in\Psi_3,\,\tau_1\in\Psi_2} P\left(g_1(Y_{1:J}) > c_1\right).$$

With the same reasoning of the proof of Lemma 55, we can find (c_1, c_2, c_3) such that the three suprema goes to 0 as $J \to \infty$. □

We need another technical Lemma.

Lemma 86. *Consider the setting of Proposition 26. Then we have*

$$E\left[(\theta_j-\mu)^2 \mid Y,\psi\right] = \frac{1}{m\tau_0+\tau_1} + \left(\frac{m\tau_0}{m\tau_0+\tau_1}\right)^2 (\bar{Y}_j-\mu)^2,$$
$$E\left[(\theta_j-\bar{Y}_j)^2 \mid Y,\psi\right] = \frac{1}{m\tau_0+\tau_1} + \left(\frac{\tau_1}{m\tau_0+\tau_1}\right)^2 (\bar{Y}_j-\mu)^2$$

and

$$Var\Big((\theta_j-\mu)^2 \mid Y,\psi\Big) = \frac{2}{(m\tau_0+\tau_1)^2} + 4\frac{m^2\tau_0^2}{(m\tau_0+\tau_1)^3}(\bar{Y}_j-\mu)^2,$$
$$Var\Big((\theta_j-\bar{Y}_j)^2 \mid Y,\psi\Big) = \frac{2}{(m\tau_0+\tau_1)^2} + 4\frac{\tau_1^2}{(m\tau_0+\tau_1)^3}(\bar{Y}_j-\mu)^2$$

and

$$Cov\Big((\theta_j-\mu)^2, (\theta_j-\bar{Y}_j)^2 \mid Y,\psi\Big) = \frac{2}{(m\tau_0+\tau_1)^2} - 4\frac{m\tau_0\tau_1}{(m\tau_0+\tau_1)^3}(\bar{Y}_j-\mu)^2.$$

Proof. Notice that by (82) we have

$$(\theta_j-\mu) \mid Y_j,\psi \sim N\left(\frac{m\tau_0}{m\tau_0+\tau_1}(\bar{Y}_j-\mu), (m\tau_0+\tau_1)^{-1}\right)$$

and

$$(\theta_j-\bar{Y}_j) \mid Y_j,\psi \sim N\left(\frac{\tau_1}{m\tau_0+\tau_1}(\mu-\bar{Y}_j), (m\tau_0+\tau_1)^{-1}\right).$$

Therefore we have

$$E\left[(\theta_j-\mu)^2 \mid Y,\psi\right] = \frac{1}{m\tau_0+\tau_1} + \left(\frac{m\tau_0}{m\tau_0+\tau_1}\right)^2 (\bar{Y}_j-\mu)^2,$$

and similarly for the other case. If $X \sim N(\mu,\sigma^2)$, by Lemma 69 we have $E[X^4] = 3\sigma^4 + 6\mu^2\sigma^2 + \mu^4$. In our case, considering $\sigma = (m\tau_0+\tau_1)^{-1/2}$ and $\mu = \frac{m\tau_0}{m\tau_0+\tau_1}(\bar{Y}_j-\mu)$, we have

$$E\left[(\theta_j-\mu)^4 \mid Y,\psi\right] = \frac{3}{(m\tau_0+\tau_1)^2} + 6\frac{m^2\tau_0^2}{(m\tau_0+\tau_1)^3}(\bar{Y}_j-\mu)^2 + \left(\frac{m\tau_0}{m\tau_0+\tau_1}\right)^4 (\bar{Y}_j-\mu)^4$$

and

$$E^2\left[(\theta_j-\mu)^2 \mid Y,\psi\right] = \frac{1}{(m\tau_0+\tau_1)^2} + 2\frac{m^2\tau_0^2}{(m\tau_0+\tau_1)^3}(\bar{Y}_j-\mu)^2 + \left(\frac{m\tau_0}{m\tau_0+\tau_1}\right)^4 (\bar{Y}_j-\mu)^4.$$

Therefore

$$\mathrm{Var}\Big((\theta_j-\mu)^2 \mid Y,\psi\Big) = \frac{2}{(m\tau_0+\tau_1)^2} + 4\frac{m^2\tau_0^2}{(m\tau_0+\tau_1)^3}(\bar{Y}_j-\mu)^2,$$

and similarly for the other one. Finally, again by Lemma 69, if $Z \sim N(0,1)$ we have

$$E[(\sigma Z+\mu_1)^2\,(\sigma Z+\mu_2)^2] = 3\sigma^4 + \sigma^2(\mu_1^2 + 4\mu_1\mu_2 + \mu_2^2) + \mu_1^2\mu_2^2.$$

In our case, considering $\sigma = (m\tau_0+\tau_1)^{-1/2}$, $\mu_1 = \frac{m\tau_0}{m\tau_0+\tau_1}(\bar{Y}_j-\mu)$ and $\mu_2 = \frac{\tau_1}{m\tau_0+\tau_1}(\mu-\bar{Y}_j)$, we have

$$\begin{aligned} E\left[(\theta_j-\mu)^2(\theta_j-\bar{Y}_j)^2 \mid Y,\psi\right] =& \frac{3}{(m\tau_0+\tau_1)^2}+\frac{m^2\tau_0^2}{(m\tau_0+\tau_1)^3}(\bar{Y}_j-\mu)^2+\frac{\tau_1^2}{(m\tau_0+\tau_1)^3}(\bar{Y}_j-\mu)^2 \\ &-4\frac{m\tau_0\tau_1}{(m\tau_0+\tau_1)^3}(\bar{Y}_j-\mu)^2+\frac{m^2\tau_0^2\tau_1^2}{(m\tau_0+\tau_1)^4}(\bar{Y}_j-\mu)^4 \end{aligned}$$

and

$$\begin{aligned} &E\left[(\theta_j-\mu)^2 \mid Y,\psi\right]E\left[(\theta_j-\bar{Y}_j)^2 \mid Y,\psi\right] = \\ &\frac{1}{(m\tau_0+\tau_1)^2}+\frac{m^2\tau_0^2}{(m\tau_0+\tau_1)^3}(\bar{Y}_j-\mu)^2+\frac{\tau_1^2}{(m\tau_0+\tau_1)^3}(\bar{Y}_j-\mu)^2+\frac{m^2\tau_0^2\tau_1^2}{(m\tau_0+\tau_1)^4}(\bar{Y}_j-\mu)^4. \end{aligned}$$

Therefore

$$\operatorname{Cov}\left((\theta_j-\mu)^2,(\theta_j-\bar{Y}_j)^2 \mid Y,\psi\right)=\frac{2}{(m\tau_0+\tau_1)^2}-4\frac{m\tau_0\tau_1}{(m\tau_0+\tau_1)^3}(\bar{Y}_j-\mu)^2,$$

as desired. □

Define

$$C(\psi)=\begin{bmatrix}0 & \frac{1}{(m\tau_0+\tau_1)^2} & \frac{m}{(m\tau_0+\tau_1)^2}\\ 0 & \frac{1}{(m\tau_0+\tau_1)^2} & \frac{m}{(m\tau_0+\tau_1)^2}\end{bmatrix},\quad V(\psi)=\begin{bmatrix}\frac{2}{(m\tau_0+\tau_1)^2}+4\frac{m\tau_0(\tau_1)^{-1}}{(m\tau_0+\tau_1)^2} & -\frac{2}{(m\tau_0+\tau_1)^2}\\ -\frac{2}{(m\tau_0+\tau_1)^2} & \frac{2}{(m\tau_0+\tau_1)^2}+4\frac{\tau_1(m\tau_0)^{-1}}{(m\tau_0+\tau_1)^2}\end{bmatrix}. \tag{96}$$

Now we define a linear rescaling of $\boldsymbol{T}=\left(\sum_{j=1}^J(\theta_j-\bar{Y}_j)^2,\sum_{j=1}^J(\theta_j-\mu)^2\right)$ as

$$\tilde{\boldsymbol{T}}=\frac{1}{\sqrt{J}}\sum_{j=1}^J\begin{bmatrix}(\theta_j-\bar{Y}_j)^2-\frac{1}{m\tau_0^*+\tau_1^*}-\left(\frac{\tau_1^*}{m\tau_0^*+\tau_1^*}\right)^2\left(\bar{Y}_j-\mu^*\right)^2\\ (\theta_j-\mu)^2-\frac{1}{m\tau_0^*+\tau_1^*}-\left(\frac{m\tau_0^*}{m\tau_0^*+\tau_1^*}\right)^2\left(\bar{Y}_j-\mu^*\right)^2\end{bmatrix}-C(\psi^*)\Delta_J, \tag{97}$$

with Δ_J as in (95). The next lemma shows the asymptotic distribution of $\tilde{\boldsymbol{T}}$ using the weak topology.

Lemma 87. *Define $\tilde{\psi}$ and $\tilde{\boldsymbol{T}}$ as in* (95) *and* (97)*, respectively. For every $\tilde{\psi}\in\mathbb{R}^D$ it holds*

$$\left\|\mathcal{L}(d\tilde{\boldsymbol{T}} \mid Y_{1:J},\tilde{\psi})-N\left(C(\psi^*)\tilde{\psi},V(\psi^*)\right)\right\|_W\to 0,$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J\to\infty$.

Proof. The result follows by arguments similar to the proof of Lemma 58. First of all notice that $C(\psi)$ defined in (96) is such that

$$C(\psi)=\begin{bmatrix}E_{Y_j}\left[\partial_\mu E[(\theta_j-\bar{Y}_j)^2 \mid Y_j,\psi\right] & E_{Y_j}\left[\partial_{\tau_1}E[(\theta_j-\bar{Y}_j)^2 \mid Y_j,\psi\right] & E_{Y_j}\left[\partial_{\tau_0}E[(\theta_j-\bar{Y}_j)^2 \mid Y_j,\psi\right]\\ E_{Y_j}\left[\partial_\mu E[(\theta_j-\mu)^2 \mid Y_j,\psi\right] & E_{Y_j}\left[\partial_{\tau_1}E[(\theta_j-\mu)^2 \mid Y_j,\psi\right] & E_{Y_j}\left[\partial_{\tau_0}E[(\theta_j-\mu)^2 \mid Y_j,\psi\right]\end{bmatrix},$$

since by Lemma 86 we have

$$\begin{aligned}
&E_{Y_j}\left[\partial_\mu E[(\theta_j-\bar{Y}_j)^2 \mid Y_j,\psi\right] = E_{Y_j}\left[\partial_\mu E[(\theta_j-\mu)^2 \mid Y_j,\psi\right] = 0,\\
&E_{Y_j}\left[\partial_{\tau_0} E[(\theta_j-\bar{Y}_j)^2 \mid Y_j,\psi\right] = E_{Y_j}\left[\partial_{\tau_0} E[(\theta_j-\mu)^2 \mid Y_j,\psi\right] = \frac{m}{(m\tau_0+\tau_1)^2},\\
&E_{Y_j}\left[\partial_{\tau_1} E[(\theta_j-\bar{Y}_j)^2 \mid Y_j,\psi\right] = E_{Y_j}\left[\partial_{\tau_1} E[(\theta_j-\mu^*)^2 \mid Y_j,\psi\right] = \frac{1}{(m\tau_0+\tau_1)^2}.
\end{aligned}$$

By the same reasoning in the proofs of (63) and (64) we get

$$E_{Y_j}\left[\tilde{T} \mid Y_{1:J},\psi^* + \frac{\tilde{\psi}+\Delta_J}{\sqrt{J}}\right] \quad\to\quad C(\psi^*)\tilde{\psi}$$

and

$$\left|\operatorname{Cov}\left(\tilde{\boldsymbol{T}} \mid Y_{1:J},\psi^* + \frac{\tilde{\psi}+\Delta_J}{\sqrt{J}}\right) - \operatorname{Cov}\left(\tilde{\boldsymbol{T}} \mid Y_{1:J},\psi^*\right)\right| \quad\to\quad 0,$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J\to\infty$. Then by (83), Lemma 86 and the Law of Large Numbers we have

$$\begin{aligned}
\operatorname{Var}\left(\frac{1}{\sqrt{J}}\sum_{j=1}^J(\theta_j-\bar{Y}_j)^2 \mid Y_{1:J},\psi^*\right) &= \frac{2}{(m\tau_0^*+\tau_1^*)^2} + 4\frac{(\tau_1^*)^2}{(m\tau_0^*+\tau_1^*)^3}\frac{1}{J}\sum_{j=1}^J(\bar{Y}_j-\mu^*)^2\\
&\to \frac{2}{(m\tau_0^*+\tau_1^*)^2} + 4\frac{(m\tau_0^*)^{-1}\tau_0^*}{(m\tau_0^*+\tau_1^*)^2}
\end{aligned}$$

and

$$\begin{aligned}
\operatorname{Var}\left(\frac{1}{\sqrt{J}}\sum_{j=1}^J(\theta_j-\mu^*)^2 \mid Y_{1:J},\psi^*\right) &= \frac{2}{(m\tau_0^*+\tau_1^*)^2} + 4\frac{(m\tau_0^*)^2}{(m\tau_0^*+\tau_1^*)^3}\frac{1}{J}\sum_{j=1}^J(\bar{Y}_j-\mu^*)^2\\
&\to \frac{2}{(m\tau_0^*+\tau_1^*)^2} + 4\frac{m\tau_0^*(\tau_1^*)^{-1}}{(m\tau_0^*+\tau_1^*)^2}
\end{aligned}$$

and

$$\begin{aligned}
&\operatorname{Cov}\left(\frac{1}{\sqrt{J}}\sum_{j=1}^J(\theta_j-\bar{Y}_j)^2, \frac{1}{\sqrt{J}}\sum_{j=1}^J(\theta_j-\mu^*)^2 \mid Y_{1:J},\psi^*\right) = \frac{2}{(m\tau_0+\tau_1)^2} - 4\frac{m\tau_0\tau_1}{(m\tau_0+\tau_1)^3}\frac{1}{J}\sum_{j=1}^J(\bar{Y}_j-\mu)^2\\
&\quad\to -\frac{2}{(m\tau_0^*+\tau_1^*)^2},
\end{aligned}$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J\to\infty$. Finally, by the Law of Large Numbers and calculations similar to Lemma 86, we have

$$E\left[(\theta_j-\bar{Y}_j)^{12} \mid Y_J,\psi\right] < \infty, \quad E\left[(\theta_j-\mu)^{12} \mid Y_J,\psi\right] < \infty$$

for every ψ. Therefore, with the same arguments in the proof of (65) we conclude that

$$\frac{1}{J^{3/2}}\sum_{j=1}^{J} E\left[(\theta_j-\bar{Y}_j)^{12} \mid Y_j, \psi^* + \frac{\tilde{\psi}+\Delta_J}{\sqrt{J}}\right] \to 0, \quad \frac{1}{J^{3/2}}\sum_{j=1}^{J} E\left[(\theta_j-\mu^*)^{12} \mid Y_j, \psi^* + \frac{\tilde{\psi}+\Delta_J}{\sqrt{J}}\right] \to 0,$$

$Q^{(\infty)}_{\psi^*}$-almost surely, as $J \to \infty$. The result then follows by Lyapunov version of Central Limit Theorem. □

We need another technical Lemma.

Lemma 88. *Consider the assumptions of Proposition 26. Then it holds*

$$\left|E\left[e^{it_1(\theta_j-\mu)^2+it_2(\theta_j-\bar{Y}_j)^2} \mid Y_j,\psi\right]\right| \leq \frac{e^{-\frac{2\sigma^2\left[\nu_j(t_1+t_2)-(t_1\mu+t_2\bar{Y}_j)\right]^2}{1+4\sigma^4(t_1+t_2)^2}}}{\left[1+4(t_1+t_2)^2\sigma^4\right]^{1/4}},$$

with $(t_1,t_2)\in\mathbb{R}_2$ *and*

$$\nu_j = \frac{m\tau_0}{m\tau_0+\tau_1}\mu + \frac{\tau_1}{m\tau_0+\tau_1}\bar{Y}_j, \quad \sigma^2 = \frac{1}{m\tau_0+\tau_1}.$$

Proof. By simple computations we get

$$t_1(\theta_j-\mu)^2 + t_2(\theta_j-\bar{Y}_j)^2 = (t_1+t_2)\theta_j^2 - 2\theta_j(t_1\mu+t_2\bar{Y}_j) + t_1\mu^2 + t_2\bar{Y}_j^2.$$

Therefore

$$\left|E\left[e^{it_1(\theta_j-\mu)^2+it_2(\theta_j-\bar{Y}_j)^2}\right]\right| \leq \left|E\left[e^{i\left((t_1+t_2)\theta_j^2-2\theta_j(\mu+\bar{Y}_j)\right)}\right]\right|.$$

Then we can apply Lemma 72, with

$$a=t_1+t_2, \quad b=-2(t_1\mu+t_2\bar{Y}_j), \quad \nu=\frac{m\tau_0}{m\tau_0+\tau_1}\mu+\frac{\tau_1}{m\tau_0+\tau_1}\bar{Y}_j, \quad \sigma^2=\frac{1}{m\tau_0+\tau_1}.$$

□

Consistently with the previous Sections, we denote

$$\varphi(t \mid Y_j,\psi) = E\left[e^{it_1(\theta_j-\bar{Y}_j)^2+it_2(\theta_j-\mu)^2} \mid Y_j,\psi\right], \quad \tilde{\varphi}(t \mid Y_{1:J},\psi) = \mathbb{E}\left[e^{it^\top\tilde{\boldsymbol{T}}} \mid Y_{1:J},\psi\right]$$

for every ψ and $t=(t_1,t_2)\in\mathbb{R}^2$. The next lemma proves the same convergence of Lemma 87 using the total variation distance.

Lemma 89. *Define* $\tilde{\psi}$ *and* $\tilde{\boldsymbol{T}}$ *as in* (95) *and* (97)*, respectively. For every* $\tilde{\psi}\in\mathbb{R}^D$ *it holds*

$$\left\|\mathcal{L}(d\tilde{\boldsymbol{T}} \mid Y_{1:J},\tilde{\psi}) - N\left(C(\psi^*)\tilde{\psi}, V(\psi^*)\right)\right\|_{TV} \to 0,$$

$Q^{(\infty)}_{\psi^*}$*-almost surely as* $J\to\infty$.

Proof. Since the result holds under the weak metric by Lemma 87, with the same reasoning of

Lemma 61 it suffices to prove

$$\lim_{A\to\infty}\lim_{B\to\infty}\limsup_{J\to\infty}\int_{\left((t_1+t_2)^2\le A, t_1^2\le B\right)^c}\left|\tilde{\varphi}(t \mid Y_{1:J},\psi^{(J)})\right| \,\mathrm{d}t = 0$$

$Q_{\psi^*}^{(\infty)}$-almost surely as $J\to\infty$, where

$$\psi^{(J)} = \psi^* + \frac{\tilde{\psi}+\Delta_J}{\sqrt{J}}$$

Analogously, denote also

$$\mu^{(J)} = \mu^* + \frac{\tilde{\mu}+\Delta_{J,1}}{\sqrt{J}}, \quad \tau_1^{(J)} = \tau_1^* + \frac{\tilde{\tau}_1+\Delta_{J,2}}{\sqrt{J}}, \quad \tau_0^{(J)} = \tau_0^* + \frac{\tilde{\tau}_0+\Delta_{J,3}}{\sqrt{J}}.$$

As in (72) we have

$$\left|\tilde{\varphi}(t \mid Y_{1:J},\psi)\right| = \left|\prod_{j=1}^{J}\varphi\left(\frac{t}{\sqrt{J}} \mid Y_j,\psi\right)\right|.$$

Therefore, with the change of variables $u = t_1+t_2$ and $v = t_1$, we have

$$\int_{\left((t_1+t_2)^2\le A, t_1^2\le B\right)^c}\left|\tilde{\varphi}(t \mid Y_{1:J},\psi^{(J)})\right| \,\mathrm{d}t \\ = \int_{\left(u^2\le A, v^2\le B\right)^c}\prod_{j=1}^{J}\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}} \mid Y_j,\psi^{(J)}\right)\right| \,\mathrm{d}u\mathrm{d}v$$

Moreover it is easy to see that

$$\Big\{(u,v)\mid u^2\le A \text{ and } v^2\le B\Big\}^c \subset \Big\{(u,v)\mid u^2>A\Big\}\cup\Big\{(u,v)\mid u^2\le A \text{ and } v^2>B\Big\},$$

so that

$$\int_{\left(u^2\le A, v^2\le B\right)^c}\prod_{j=1}^{J}\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}} \mid Y_j,\psi^{(J)}\right)\right| \,\mathrm{d}u\mathrm{d}v \le \int_{u^2>A}\prod_{j=1}^{J}\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}} \mid Y_j,\psi^{(J)}\right)\right| \,\mathrm{d}u\mathrm{d}v \\ + \int_{\left(u^2\le A, v^2>B\right)}\prod_{j=1}^{J}\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}} \mid Y_j,\psi^{(J)}\right)\right| \,\mathrm{d}u\mathrm{d}v. \tag{98}$$

For every ψ, by Lemma 88 with

$$\nu_j = \frac{m\tau_0}{m\tau_0+\tau_1}\mu + \frac{\tau_1}{m\tau_0+\tau_1}\bar{Y}_j, \quad \sigma^2 = \frac{1}{m\tau_0+\tau_1}$$

we have

$$\prod_{j=1}^{J}\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}} \mid Y_j,\psi\right)\right| \le \frac{e^{-\frac{2\sigma^2\frac{1}{J}\sum_{j=1}^{J}\left[u(\nu_j-\bar{Y}_j)-v(\mu-\bar{Y}_j)\right]^2}{1+4\sigma^4u^2}}}{\left[1+4u^2\sigma^4\right]^{J/4}}.$$

Notice that

$$\frac{1}{J}\sum_{j=1}^{J}\Big[u(\nu_j-\bar{Y}_j)-v(\mu-\bar{Y}_j)\Big]^2 =$$

$$= v^2\left[\frac{1}{J}\sum_{j=1}^{J}(\mu-\bar{Y}_j)^2\right] - 2uv\left[\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)(\mu-\bar{Y}_j)\right] + u^2\left[\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)^2\right]$$

$$= \left[\frac{1}{J}\sum_{j=1}^{J}(\mu-\bar{Y}_j)^2\right]\left[v-u\frac{\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)(\mu-\bar{Y}_j)}{\frac{1}{J}\sum_{j=1}^{J}(\mu-\bar{Y}_j)^2}\right]^2$$

$$+ u^2\left[\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)^2 - \frac{\left\{\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)(\mu-\bar{Y}_j)\right\}^2}{\frac{1}{J}\sum_{j=1}^{J}(\mu-\bar{Y}_j)^2}\right].$$

As regards the first element in (98), by integrating with respect to v we get

$$\int_{u^2>A}\prod_{j=1}^{J}\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}}\mid Y_j,\psi^{(J)}\right)\right|\mathrm{d}u\mathrm{d}v \le \int_{u^2>A}\frac{e^{-\frac{2\sigma_J^2\frac{1}{J}\sum_{j=1}^{J}\left[u(\nu_j-\bar{Y}_j)-v(\mu^{(J)}-\bar{Y}_j)\right]^2}{1+4\sigma_J^4u^2}}}{\left[1+4u^2\sigma_J^4\right]^{J/4}}\mathrm{d}u\mathrm{d}v$$

$$\le\sqrt{\frac{\pi}{2\sigma_J^2\frac{1}{J}\sum_{j=1}^{J}(\mu^{(J)}-\bar{Y}_j)^2}}\int_A^\infty\frac{e^{-\frac{2\sigma_J^2}{1+4\sigma_J^4u^2}u^2\left[\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)^2-\frac{\left\{\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)(\mu_J-\bar{Y}_j)\right\}^2}{\frac{1}{J}\sum_{j=1}^{J}(\mu^{(J)}-\bar{Y}_j)^2}\right]}}{\left[1+4u^2\sigma_J^4\right]^{J/4-1/2}}\mathrm{d}u,$$

where

$$\sigma_J^2=\frac{1}{m\tau_0^{(J)}+\tau_1^{(J)}},\quad \nu_j=\frac{m\tau_0^{(J)}}{m\tau_0^{(J)}+\tau_1^{(J)}}\mu^{(J)}+\frac{\tau_1^{(J)}}{m\tau_0^{(J)}+\tau_1^{(J)}}\bar{Y}_j.$$

By the Law of Large Numbers we have

$$\liminf\frac{1}{J}\sum_{j=1}^{J}(\mu^{(J)}-\bar{Y}_j)^2=\liminf\frac{1}{J}\sum_{j=1}^{J}(\mu^*-\bar{Y}_j)^2=c_1>0$$

$Q_{\psi^*}^{(\infty)}$-almost surely and similarly

$$\liminf\left\{\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)^2-\frac{\left\{\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)(\mu^{(J)}-\bar{Y}_j)\right\}^2}{\frac{1}{J}\sum_{j=1}^{J}(\mu^{(J)}-\bar{Y}_j)^2}\right\}=c_2>0,$$

by Cauchy-Schwartz inequality, $Q_{\psi^*}^{(\infty)}$-almost surely. Moreover, by Lemma 57

$$\sigma_J^2\in\left(\frac{1}{2}\frac{1}{m\tau_0^*+\tau_1^*},\frac{2}{m\tau_0^*+\tau_1^*}\right)=(\sigma_1^2,\sigma_2^2)$$

$Q_{\psi^*}^{(\infty)}$-almost surely, for J high enough. Therefore

$$\begin{aligned}
&\lim_{A\to\infty}\lim_{B\to\infty}\limsup_{J\to\infty}\int_{u^2>A}\prod_{j=1}^{J}\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}}\mid Y_j,\psi^{(J)}\right)\right|\mathrm{d}u\mathrm{d}v \\
&\qquad \leq \lim_{A\to\infty}\sqrt{\frac{\pi}{2\sigma_1^2 c_1}}\int_A^{\infty}\frac{e^{-\frac{2c_2\sigma_1^2}{1+4\sigma_2^4u^2}u^2}}{\left[1+4u^2\sigma_1^4\right]^{J/4-1/2}}\mathrm{d}u=0
\end{aligned}$$

$Q_{\psi^*}^{(\infty)}$-almost surely. As regards the second addend in (98) we get

$$\begin{aligned}
&\limsup_{J\to\infty}\int_{(u^2\leq A,v^2>B)}\prod_{j=1}^{J}\left|\varphi\left(\frac{(v,u-v)}{\sqrt{J}}\mid Y_j,\psi^{(J)}\right)\right|\mathrm{d}u\mathrm{d}v \\
&\qquad \leq \int_{(u^2\leq A,v^2>B)} e^{-\frac{2\sigma_1^2}{1+\sigma_2^4A^2}\left[v-u\frac{\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)(\mu^{(J)}-\bar{Y}_j)}{\frac{1}{J}\sum_{j=1}^{J}(\mu^{(J)}-\bar{Y}_j)^2}\right]^2}\mathrm{d}u\mathrm{d}v,
\end{aligned}$$

$Q_{\psi^*}^{(\infty)}$-almost surely. Fix $A>0$ and notice that for every u we have

$$\lim_{B\to\infty}\int_B^{\infty} e^{-\frac{2\sigma_1^2}{1+\sigma_2^4A^2}\left[v-u\frac{\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)(\mu^{(J)}-\bar{Y}_j)}{\frac{1}{J}\sum_{j=1}^{J}(\mu^{(J)}-\bar{Y}_j)^2}\right]^2}\mathrm{d}v=0.$$

Moreover

$$\int_{u^2\leq A} e^{-\frac{2\sigma_1^2}{1+\sigma_2^4A^2}\left[v-u\frac{\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)(\mu-\bar{Y}_j)}{\frac{1}{J}\sum_{j=1}^{J}(\mu-\bar{Y}_j)^2}\right]^2}\mathrm{d}u\mathrm{d}v<\infty,$$

so that, by Dominated Convergence Theorem we get

$$\lim_{B\to\infty}\int_{(u^2\leq A,v^2>B)} e^{-\frac{2\sigma_1^2}{1+\sigma_2^4A^2}\left[v-u\frac{\frac{1}{J}\sum_{j=1}^{J}(\nu_j-\bar{Y}_j)(\mu^{(J)}-\bar{Y}_j)}{\frac{1}{J}\sum_{j=1}^{J}(\mu^{(J)}-\bar{Y}_j)^2}\right]^2}\mathrm{d}u\mathrm{d}v=0,$$

for every $A>0$ and the result follows. □

Proof of Proposition 26. The result follows by arguments similar to the proof of Theorem 20, that we briefly summarize. Since by construction

$$\mathcal{L}\left(\mathrm{d}\psi\mid\boldsymbol{\theta},Y_{1:J}\right)=\mathcal{L}\left(\mathrm{d}\psi\mid\boldsymbol{T},Y_{1:J}\right)$$

a direct analogue of Lemma 44 holds. Moreover, by Lemmas 85 and 89, we can use Lemma 64 to prove that $\mathcal{L}\left(\mathrm{d}\tilde{\boldsymbol{T}},\mathrm{d}\tilde{\psi}\mid Y_{1:J}\right)$, as in (95), converges to a Gaussian vector with non singular covariance matrix. Finally, Lemma 54 holds for P, being a two-block Gibbs sampler. Therefore the Gibbs sampler on the limit Gaussian target has a strictly positive spectral gap: thus the result follows by Corollary 8. □

Proof of Corollary 14

Let $\phi = (\tau_1, \tau_0)$ and define

$$\mathcal{I}(\phi^*) = \begin{bmatrix} \frac{m^2(\tau_0^*)^2}{2(\tau_1^*)^2(\tau_1^*+m\tau_0^*)^2} & \frac{m}{2(\tau_1^*+m\tau_0^*)^2} \\ \frac{m}{2(\tau_1^*+m\tau_0^*)^2} & \frac{m-1}{2(\tau_0^*)^2} + \frac{(\tau_1^*)^2}{2(\tau_0^*)^2(\tau_1^*+m\tau_0^*)^2} \end{bmatrix}, \quad C(\phi^*) = \begin{bmatrix} \frac{1}{(m\tau_0^*+\tau_1^*)^2} & \frac{m}{(m\tau_0^*+\tau_1^*)^2} \\ \frac{1}{(m\tau_0^*+\tau_1^*)^2} & \frac{m}{(m\tau_0^*+\tau_1^*)^2} \end{bmatrix}$$

and

$$V(\phi^*) = \begin{bmatrix} \frac{2}{(m\tau_0^*+\tau_1^*)^2} + 4\frac{m\tau_0^*(\tau_1^*)^{-1}}{(m\tau_0^*+\tau_1^*)^2} & -\frac{2}{(m\tau_0^*+\tau_1^*)^2} \\ -\frac{2}{(m\tau_0^*+\tau_1^*)^2} & \frac{2}{(m\tau_0^*+\tau_1^*)^2} + 4\frac{\tau_1^*(m\tau_0^*)^{-1}}{(m\tau_0^*+\tau_1^*)^2} \end{bmatrix}.$$

We have a preliminary Lemma.

Lemma 90. *Consider the setting of Proposition 26. Then we have*

$$\gamma(\psi^*) = \min\left\{\frac{1}{1+\lambda_i}\,;\, \lambda_i \textit{ eigenvalue of } V^{-1}\left(\phi^*\right) C(\phi^*)\mathcal{I}^{-1}(\phi^*)C^\top(\phi^*)\right\}.$$

Proof. With the same reasoning of Corollary 10, $\gamma(\psi^*)$ is the spectral gap on the limiting Gaussian distribution of $\left(\tilde{\psi}, \tilde{\boldsymbol{T}}\right)$, given by by Lemmas 85 and 89. By inspecting $\mathcal{I}(\psi^*)$ in (84) and $C(\psi^*)$ in (96), we have that $\tilde{\mu}$ is asymptotically independent from everything else, therefore it suffices to study the Gibbs sampler that alternates updates of $(\tilde{\tau}_1, \tilde{\tau}_0)$ and $\tilde{\boldsymbol{T}}$. Then the result follows by the same arguments of Corollary 10. □

Proof of Corollary 14. By Lemma 90 we have to study the eigenvalues of

$$V^{-1}\left(\phi^*\right) C(\phi^*)\mathcal{I}^{-1}(\phi^*)C^\top(\phi^*). \tag{99}$$

Notice that

$$\mathcal{I}(\phi^*) = \frac{1}{(m\tau_0^*+\tau_1^*)^2}\begin{bmatrix} \frac{m^2(\tau_0^*)^2}{2(\tau_1^*)^2} & \frac{m}{2} \\ \frac{m}{2} & \frac{(m-1)(m\tau_0^*+\tau_1^*)^2+(\tau_1^*)^2}{2(\tau_0^*)^2} \end{bmatrix}, \quad C(\phi^*) = \frac{1}{(m\tau_0^*+\tau_1^*)^2}\begin{bmatrix} 1 & m \\ 1 & m \end{bmatrix}$$

and

$$V(\phi^*) = \frac{1}{(m\tau_0^*+\tau_1^*)^2}\begin{bmatrix} 2+4\frac{m\tau_0^*}{\tau_1^*} & -2 \\ -2 & 2+4\frac{\tau_1^*}{m\tau_0^*} \end{bmatrix}$$

Notice that

$$\begin{aligned}\left((m\tau_0^*+\tau_1^*)^2V(\phi^*)\right)^{-1} &= \frac{m\tau_0^*\tau_1^*}{8(m\tau_0^*+\tau_1^*)^2}\begin{bmatrix} 2+4\frac{\tau_1^*}{m\tau_0^*} & 2 \\ 2 & 2+4\frac{m\tau_0^*}{\tau_1^*} \end{bmatrix} \\ &= \frac{1}{4(m\tau_0^*+\tau_1^*)^2}\begin{bmatrix} m\tau_0^*\tau_1^*+2(\tau_1^*)^2 & m\tau_0^*\tau_1^* \\ m\tau_0^*\tau_1^* & m\tau_0^*\tau_1^*+2(m\tau_0^*)^2 \end{bmatrix}\end{aligned}$$

and

$$\Big((m\tau_0^*+\tau_1^*)^2\mathcal{I}(\phi^*)\Big)^{-1} = \frac{2(\tau_1^*)^2}{m^2(m-1)(m\tau_0^*+\tau_1^*)^2}\begin{bmatrix} \frac{(m-1)(m\tau_0^*+\tau_1^*)^2+(\tau_1^*)^2}{(\tau_0^*)^2} & -m \\ -m & \frac{(m\tau_0^*)^2}{(\tau_1^*)^2} \end{bmatrix}$$

Therefore

$$\begin{aligned}\frac{m^2(m-1)(m\tau_0^*+\tau_1^*)^4}{2(\tau_1^*)^2}C(\phi^*)\mathcal{I}^{-1}(\phi^*)C^\top(\phi^*) &= \begin{bmatrix} -m^2+\frac{(m-1)(m\tau_0^*+\tau_1^*)^2+(\tau_1^*)^2}{(\tau_0^*)^2} & \frac{m^3(\tau_0^*)^2}{(\tau_1^*)^2}-m \\ -m^2+\frac{(m-1)(m\tau_0^*+\tau_1^*)^2+(\tau_1^*)^2}{(\tau_0^*)^2} & \frac{m^3(\tau_0^*)^2}{(\tau_1^*)^2}-m \end{bmatrix}\begin{bmatrix} 1 & 1 \\ m & m \end{bmatrix} \\ &= \left(\frac{m^4(\tau_0^*)^2}{(\tau_1^*)^2}-2m^2+\frac{(m-1)(m\tau_0^*+\tau_1^*)^2+(\tau_1^*)^2}{(\tau_0^*)^2}\right)\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \\ &= \left(\frac{m^4(\tau_0^*)^4-2m^2(\tau_0^*)^2(\tau_1^*)^2+(m-1)(\tau_1^*)^2(m\tau_0^*+\tau_1^*)^2+(\tau_1^*)^4}{(\tau_0^*)^2(\tau_1^*)^2}\right)\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}V^{-1}\left(\phi^*\right)C(\phi^*)\mathcal{I}^{-1}(\phi^*)C^\top(\phi^*) = &\left(\frac{m^4(\tau_0^*)^4-2m^2(\tau_0^*)^2(\tau_1^*)^2+(m-1)(\tau_1^*)^2(m\tau_0^*+\tau_1^*)^2+(\tau_1^*)^4}{2m^2(m-1)(\tau_0^*)^2(m\tau_0^*+\tau_1^*)^4}\right) \\ &\begin{bmatrix} 2m\tau_0^*\tau_1^*+2(\tau_1^*)^2 & 2m\tau_0^*\tau_1^*+2(\tau_1^*)^2 \\ 2m\tau_0^*\tau_1^*+2(m\tau_0^*)^2 & 2m\tau_0^*\tau_1^*+2(m\tau_0^*)^2 \end{bmatrix}\end{aligned}$$

Notice that the matrix on the right hand side admits 0 as an eigenvalue, so that the highest eigenvalue in absolute value is given by its trace, that is

$$4m\tau_0^*\tau_1^*+2(\tau_1^*)^2+2(m\tau_0^*)^2 = 2(m\tau_0^*+\tau_1^*)^2,$$

so that the highest eigenvalue of (99) is given by

$$\frac{m^4(\tau_0^*)^4-2m^2(\tau_0^*)^2(\tau_1^*)^2+(m-1)(\tau_1^*)^2(m\tau_0^*+\tau_1^*)^2+(\tau_1^*)^4}{m^2(m-1)(\tau_0^*)^2(m\tau_0^*+\tau_1^*)^2}.$$

The result follows by noticing

$$\begin{aligned}m^4(\tau_0^*)^4-2m^2(\tau_0^*)^2(\tau_1^*)^2+(\tau_1^*)^4 &= \Big[m^2(\tau_0^*)^2-(\tau_1^*)^2\Big]^2 \\ &= (m\tau_0^*-\tau_1^*)^2(m\tau_0^*+\tau_1^*)^2.\end{aligned}$$

□

Proof of Lemma 48

Proof. The proof follows the same lines of Lemma 44, that we briefly summarize. Since

$$\mathcal{L}\Big(\mathrm{d}\theta,\mathrm{d}\tau_\beta \mid \boldsymbol{\beta},Y^{(n)}) = \mathcal{L}(\mathrm{d}\theta,\mathrm{d}\tau_\beta \mid \boldsymbol{T}(\boldsymbol{\beta}),Y^{(n)}\Big) \tag{100}$$

holds by definition of $\boldsymbol{T}$, reasoning as in (59) we can conclude

$$\mathcal{L}\left(\mathrm{d}\boldsymbol{T}^{(t)}, \mathrm{d}\theta^{(t)}, \mathrm{d}\tau_\beta^{(t)} \,|\boldsymbol{T}^{(t-1)}, \theta^{(t-1)}, \tau_\beta^{(t-1)}\right) \\ = \hat{\pi}_n\left(\mathrm{d}\boldsymbol{T}^{(t)} \mid \theta^{(t-1)}, \tau_\beta^{(t-1)}\right) \hat{\pi}_n\left(\mathrm{d}\theta^{(t)}, \mathrm{d}\tau_\beta^{(t)} \mid \boldsymbol{T}^{(t)}\right),$$

which proves that the transition kernel of the induced chain $\left(\boldsymbol{T}^{(t)}, \theta^{(t)}, \tau_\beta^{(t)}\right)_{t\geq 1}$ coincides with $\hat{P}_n$. The second part of the Lemma follows by the same reasoning used in (60). □

Proof of Corollary 15

Proof. By Lemma 48 we have

$$t_{mix}^{(n)}(\epsilon, M) = \sup_{\nu\in\mathcal{N}(\hat{\pi}_n, M)} \hat{t}_{mix}^{(n)}(\epsilon, \nu).$$

The result then follows by Corollary 8, whose conditions hold by assumption. □

Proof of Corollary 16

Proof. It is easy to show that an analogue of Lemma 48 holds, with $\psi = (\theta, \tau_\beta, \tau_\epsilon)$ and $\boldsymbol{T} = \left(T_\theta, T_{\tau_\beta}, T_{\tau_\epsilon}\right)$. Thus the result follows with the same reasoning of Corollary 15. □

Proof of Theorem 21

Denote with $\tilde{\mu}_J$ the push-forward measure of μ_J according to transformations (4.17) and (4.19). The next theorem shows that the rescaled version of μ_J is a warm start for the limiting distribution in Proposition 23.

Lemma 91. *Let $\mu_J \in \mathcal{P}\left(\mathbb{R}^{lJ+D}\right)$ be as in* (4.35). *Then under assumptions* $(B1) - (B3)$ *there exists a positive constant $M = M(c)$ such that*

$$Q_{\psi^*}^{(J)}\left(\tilde{\mu}_J \in \mathcal{N}\left(N(\mathbf{0}, \Sigma), M\right)\right) \quad \to \quad 1,$$

as $J \to \infty$, with Σ as in Proposition 23.

Proof. According to transformations (4.17), we have

$$\tilde{\mu}_J^{(-1)} = \mathrm{Unif}\left(\sqrt{J}\left(\hat{\psi}_J - \psi^*\right) - \Delta_J, c\right).$$

Denote with $B_r(\mathbf{x})$ the closed ball of radius $r > 0$ and center $\mathbf{x} \in \mathbb{R}^D$. By Theorem 5.39 in Van der Vaart (2000) it holds

$$Q_{\psi^*}^{(J)}\left(\left(\sqrt{J}\left(\hat{\psi}_J - \psi^*\right) - \Delta_J\right) \in B_1(\mathbf{0})\right) \quad \to \quad 1, \tag{101}$$

as $J \to \infty$. Define now

$$M = \max_{\mathbf{x} \in B_{c+1}(\mathbf{0})} \frac{\mathrm{Vol}\left(B_{c+1}(\mathbf{0})\right)}{N(\mathbf{x} \mid \mathbf{0}, \Sigma_D)}, \tag{102}$$

where $\mathrm{Vol}(A)$ is the volume of set A and $N(\mathbf{0}, \Sigma_D)$ is the marginal distribution of $N(\mathbf{0}, \Sigma)$ over the last D components. It is easy to see that $M < \infty$ and it does not depend on J. Therefore, by (101), we conclude

$$\begin{aligned} Q_{\psi^*}^{(J)}\left(\tilde{\mu}_J \in \mathcal{N}\left(N(\mathbf{0}, \Sigma), M\right)\right) &\leq Q_{\psi^*}^{(J)}\left(\max_{\mathbf{x} \in B_{c+1}(\mathbf{0})} \frac{\mathrm{d}\tilde{\mu}_J^{(-1)}}{\mathrm{d}N(\mathbf{0}, \Sigma_D)}(\mathbf{x}) \leq M\right) \\ &\leq Q_{\psi^*}^{(J)}\left(\left(\sqrt{J}\left(\hat{\psi}_J - \psi^*\right) - \Delta_J\right) \in B_1(\mathbf{0})\right) \quad \to \quad 1, \end{aligned}$$

as $J \to \infty$. □

Proof of Theorem 21. Let $\mu_J \in \mathcal{P}\left(\mathbb{R}^{lJ+D}\right)$ be as in (4.35). Thus, by Lemma 91 the event $\{\tilde{\mu}_J \in \mathcal{N}\left(\tilde{\pi}, M\right)\}$ with M as in (102) holds with probability converging to 1, with respect to the law $Q_{\psi^*}^{(J)}$. Then, by Lemma 43, there exists $\tilde{\nu}_J \in \mathcal{N}(\tilde{\pi}_J, M)$ such that

$$\left\|\tilde{\nu}_J - \tilde{\mu}_J\right\|_{TV} \leq M \left\|\tilde{\pi}_J - \tilde{\pi}\right\|_{TV}.$$

Therefore, by the above facts, the triangle inequality and Lemma 44 we have

$$\begin{aligned} \left\|\mu_J P_J^t - \pi_J\right\|_{TV} &= \left\|\tilde{\mu}_J \tilde{P}_J^t - \tilde{\pi}_J\right\|_{TV} \\ &\leq \left\|\tilde{\mu}_J \tilde{P}_J^t - \tilde{\nu}_J \tilde{P}_J^t\right\|_{TV} + \left\|\tilde{\nu}_J \tilde{P}_J^t - \tilde{\pi}_J\right\|_{TV} \\ &\leq \left\|\tilde{\mu}_J - \tilde{\nu}_J\right\|_{TV} + \left\|\tilde{\nu}_J \tilde{P}_J^t - \tilde{\pi}_J\right\|_{TV} \\ &\leq M \left\|\tilde{\pi}_J - \tilde{\pi}\right\|_{TV} + \sup_{\tilde{\nu}_J \in \mathcal{N}(\tilde{\pi}_J, M)} \left\|\tilde{\nu}_J \tilde{P}_J^t - \tilde{\pi}_J\right\|_{TV} \\ &= M \left\|\tilde{\pi}_J - \tilde{\pi}\right\|_{TV} + \sup_{\nu_J \in \mathcal{N}(\pi_J, M)} \left\|\nu_J P_J^t - \pi_J\right\|_{TV}. \end{aligned}$$

Thus the result follows by Theorem 20. □

References

Amit, Y. (1991). On Rates of Convergence of Stochastic Relaxation for Gaussian and Non-Gaussian Distributions. *J. Multivar. Anal.*, 38:82–99.

Andrieu, C., Lee, A., Power, S., and Wang, A. Q. (2022). Explicit convergence bounds for Metropolis Markov chains: isoperimetry, spectral gaps and profiles. *arXiv preprint arXiv:2211.08959.*

Atchadé, Y. F. (2021). Approximate Spectral Gaps for Markov Chain Mixing Times in High Dimensions. *SIAM. J. MATH. DATA SCI.*, 3:854–872.

Bally, V. and Caramellino, L. (2015). Asymptotic development for the CLT in total variation distance. *Bernoulli*, 22:2442–2485.

Bass, M. R. and Sahu, S. K. (2016). A comparison of centring parameterisations of Gaussian process-based models for Bayesian computation using MCMC. *Stat. Comput.*, 27:1491–1512.

Belloni, A. and Chernozhukov, V. (2009). On the computational complexity of MCMC-based estimators in large samples. *Ann. Statist.*, 37:2011–2055.

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J., and Stuart, A. (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, 19:1501–1534.

Bhattacharya, R. N. and Rao, R. R. (2010). *Normal Approximations and Asymptotic Expansions*. Society for Industrial and Applied Mathematics.

Bobkov, S. G., Chistyakov, G. P., and Götze, F. (2014). Berry-Essen bounds in the entropic central limit theorem. *Probab. Theory Relat. Fields*, 159:435–478.

Brooks, S., Gelman, A., Jones, G. L., and Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall.

Caprio, R. and Johansen, A. (2023). A calculus for Markov chain Monte Carlo: studying approximations in algorithms. *arXiv preprint arXiv:2310.03853*.

Casella, G. and george, E. I. (1992). Explaining the Gibbs Sampler. *Am. Stat.*, 46:167–174.

Chlebicka, I., Latuszynski, K., and Miasojedow, B. (2023). Solidarity of Gibbs Samplers: the spectral gap. *arXiv preprint arXiv:2304.02109*.

Dalalyan, A. S. (2017). Theoretical Guarantees for Approximate Sampling from Smooth and Log-Concave Densities. *J. R. Stat. Soc. Ser. B.*, 79:651–676.

Diaconis, P., Khare, K., and Saloff-Coste, L. (2008). Gibbs Sampling, Exponential Families and Orthogonal Polynomials. *Stat. Sci.*, 23:151–178.

Diebolt, J. and Robert, C. P. (1994). Estimation of Finite Mixture Distributions through Bayesian Sampling. *J. R. Stat. Soc. Ser. B.*, 56:363–375.

Durante, D. (2019). Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika*, 106:765–779.

Durmus, A. and Moulines, E. (2017). Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *Ann. Appl. Probab.*, 27:1551–1587.

Dwivedi, R., Chen, Y., Wainwright, M. J., and Yu, B. (2019). Log–concave sampling: Metropolis–Hastings algorithms are fast! *J. Mach. Learn. Res.*, 20:1–42.

Feller, W. (1970). *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons.

Flegal, J. M., Hughes, J., Vats, D., Gupta, K., and Maji, U. (2021). mcmcse: Monte Carlo Standard Errors for MCMC. *R package*.

Gelfand, A. E., Kim, H. J., Sirmans, C., and Banerjee, S. (2003). Spatial Modelling With Spatially Varying Coefficient Processes. *J. Am. Stat. Assoc.*, 98:387–396.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient Parametrisations for Normal Linear Mixed Models. *Biometrika*, 82:479–488.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC press.

Gelman, A. and Hill, J. L. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gilks, W. R. and Wild, P. (1992). Adaptive Rejection Sampling for Gibbs Sampling. *J. R. Stat. Soc. Ser. C*, 41:337–348.

Gong, L. and Flegal, J. M. (2015). A Practical Sequential Stopping Rule for High-Dimensional Markov Chain Monte Carlo. *J. Comput. Graph. Stat.*, 25:684–700.

Green, P. J., Latuszynski, K., Pereyra, M., and Robert, C. P. (2015). Bayesian computation: a summary of the current state, and samples backwards and forwards. *Stat. Comput.*, 25:835–862.

Hobert, J. P. (2011). The data augmentation algorithm: Theory and methodology. *Handbook of Markov chain Monte Carlo*, pages 253–293.

Jin, Z. and Hobert, J. P. (2022). Dimension free convergence rates for Gibbs samplers for Bayesian linear mixed models. *Stoch. Process. Their Appl.*, 148:25–67.

Johndrow, J. E., Smith, A., Pillai, N., and Dunson, D. B. (2019). MCMC for Imbalanced Categorical Data. *J. Am. Stat. Assoc.*, 114:1394–1403.

Kamatani, K. (2014). Local consistency of Markov chain Monte Carlo methods. *Ann. Inst. Stat. Math.*, 66:63–74.

Khare, K. and Zhou, H. (2009). Rates of convergence of some multivariate Markov chains with polynomial eigenfunctions. *Ann. Appl. Probab.*, 2:737–777.

Kleijn, B. J. K. and van der Vaart, A. W. (2012). The Bernstein-Von-Mises theorem under misspecification. *Electron. J. Stat.*, 6:353–381.

Liu, J. S. (1994). Fraction of Missing Information and Convergence Rate for Data Augmentation. In *Computationally Intensive Statistical Methods: Proceedings of the 26th Symposium Interface*.

Lovász, L. and Simonovits, M. (1993). Random Walks in a Convex Body and an Improved Volume Algorithm. *Random Struct. and Alg.*, 4:359–412.

Martin, G. M., Frazier, D. T., and Robert, C. P. (2023). Computing Bayes: From Then 'Til Now. *Stat. Sci.*, In press.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, pages 113–162.

Negrea, J., Yang, J., Feng, H., Roy, D. M., and Huggins, J. H. (2022). Statistical Inference with Stochastic Gradient Algorithms. *arXiv preprint arXiv:2207.12395*.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.*, 41:370–400.

Nickl, R. and Wang, S. (2022). On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms. *J. Eur. Math. Soc.*

Papaspiliopoulos, O., Roberts, G., and Zanella, G. (2020). Scalable inference for crossed random effects models. *Biometrika*, 107:25–40.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2003). Non-Centered Parameterizations for Hierarchical Models and Data Augmentation (with discussion). In *Bayesian Statistics (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.)*, pages 307–326.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007a). A General Framework for the Parametrization of Hierarchical Models. *Stat. Sci.*, pages 59–73.

Papaspiliopoulos, O., Roberts, G. O. R., and Sköld, M. (2007b). A General Framework for the Parametrization of Hierarchical Models. *Statistical Science*, pages 59–73.

Papaspiliopoulos, O., Stumpf-Fétizon, T., and Zanella, G. (2023). Scalable computation for Bayesian hierarchical models. *arXiv preprint arXiv:2103.10875.*

Petrov, V. V. (1956). A local theorem for densities of sums of independent random variables. *Theory Probab. Appl.*, 84:316–322.

Qin, Q. and Hobert, J. P. (2019). Convergence complexity analysis of Albert and Chib's algorithm for Bayesian probit regression. *Ann. Statist.*, 47:2320–2347.

Qin, Q. and Hobert, J. P. (2022). Wasserstein-based methods for convergence complexity analysis of MCMC with applications. *Ann, Appl. Prob.*, 32:124–166.

Rajaratnam, B. and Sparks, D. (2015). MCMC-Based Inference in the Era of Big Data: A Fundamental Analysis of the Convergence Complexity of High-Dimensional Chains. *arXiv preprint arXiv:1508.00947.*

Roberts, G. O. and Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B*, 60:255–268.

Roberts, G. O. and Rosenthal, J. S. (2001). Markov Chains and De-Initializing Processes. *Scand. J. Stat.*, 28:489–504.

Roberts, G. O. and Rosenthal, J. S. (2004). General state space Markov chains and MCMC algorithms. *Probab. Surv.*, 60:255–268.

Roberts, G. O. and Sahu, S. H. (1994). Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stoch. Process. Their Appl.*, 49:207–216.

Roberts, G. O. and Sahu, S. H. (1997). Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *J. R. Stat. Soc. Ser. B*, 59:291–317.

Roberts, G. O. and Sahu, S. H. (2001). Approximate Predetermined Convergence Properties of the Gibbs Sampler. *J. Comput. Graph. Statist.*, 10:216–229.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2:341–363.

Rosenthal, J. S. (1995). Minorization Conditions and Convergence Rates for Markov Chain Monte Carlo. *J. Am. Stat. Assoc*, 90:558–566.

Rosenthal, J. S. and Rosenthal, P. (2015). Spectral bounds for certain two-factor non-reversible MCMC algorithms. *Electron. Commun. Probab..*, 20:1–10.

Ross, N. (2011). Fundamentals of Stein's method. *Probab. Surv.*, 8:210–293.

Smith, W. L. (1953). A frequency-function form of the central limit theorem. *Math. Proc. Camb. Philos. Soc.*, 49:462–472.

Tang, R. and Yang, Y. (2022). Computational Complexity of Metropolis-Adjusted Langevin Algorithms for Bayesian Posterior Sampling. *arXiv preprint arXiv:2206.06491.*

Thompson, M. (2010). A Comparison of Methods for Computing Autocorrelation Time. *Technical Report No. 1007, Department of Statistics, University of Toronto.*

Van der Vaart, A. W. (2000). *Asymptotic Statistics.* Cambridge University Press.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian Processes for Machine Learning.* Cambridge MA: MIT press.

Wu, K., Schmidler, S., and Chen, Y. (2022). Minimax Mixing Time of the Metropolis-Adjusted Langevin Algorithm for Log-Concave Sampling. *J. Mach. Learn. Res.*, 23:1–63.

Yang, J. and Rosenthal, J. S. (2022). Complexity results for MCMC derived from quantitative bounds. *Ann. Appl. Prob.*, 33:1459–1500.

Yang, J., Wainwright, M. J., and Jordan, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *Ann. Statist.*, 44:2497–2532.

Zhou, Q., Yang, J., Vats, D., Roberts, G. O., and Rosenthal, J. S. (2022). Dimension-free mixing for high-dimensional Bayesian variable selection. *J. R. Stat. Soc. Ser. B*, 84:1751–1784.