Università Commerciale "Luigi Bocconi"
PhD School

PhD program in: Statistics
Cycle: XXXII
Disciplinary Field: secs-s/01

Bayesian dimensionality reduction

Advisor: Daniele DURANTE

PhD Thesis by:
Sirio LEGRAMANTI
ID Number: 3032129

Year 2021

Contents

Acknowledgements

First of all, I want to thank my advisor for his continuous support and advice, which definitely made the difference.

Thanks to David B. Dunson, for the priceless opportunity to work with him: it has been a great honor and such a stimulating experience.

I also want to thank the whole faculty at Bocconi Department of Decision Sciences for being always helpful to students.

A special thank goes to all my fellow Bocconi PhD students in Statistics, for sharing this journey in a spirit of mutual help.

Last but not least, thanks to my parents, my brother and Francesca, for always supporting me with their unconditional love.

Abstract

We are currently witnessing an explosion in the amount of available data. Such growth involves not only the number of data points but also their dimensionality. This poses new challenges to statistical modeling and computations, thus making dimensionality reduction more central than ever. In the present thesis, we provide methodological, computational and theoretical advancements in Bayesian dimensionality reduction via novel structured priors. Namely, we develop a new increasing shrinkage prior and illustrate how it can be employed to discard redundant dimensions in Gaussian factor models. In order to make it usable for larger datasets, we also investigate variational methods for posterior inference under this proposed prior. Beyond traditional models and parameter spaces, we also provide a different take on dimensionality reduction, focusing on community detection in networks. For this purpose, we define a general class of Bayesian nonparametric priors that encompasses existing stochastic block models as special cases and includes promising unexplored options. Our Bayesian approach allows for a natural incorporation of node attributes and facilitates uncertainty quantification as well as model selection.

Chapter 1

Introduction

1.1 Bayesian dimensionality reduction

Even if the expression *big data* is sometimes abused, it is a fact that we are facing an unprecedented availability of data. This abundance is two-fold: massive samples, comprising millions or even billions of individuals, and high dimensionality, with a large number of features being measured for each individual. These huge datasets are costly, both in terms of space (memory required to store them) and time (needed to run algorithms on them). Moreover, even if we had access to unlimited resources, high-dimensional data are troublesome per se. In fact, besides being difficult to visualize and to interpret, they give rise to inherent issues, often referred to as *curse of dimensionality*. This expression, arguably coined by Bellman (1961), refers to a variety of adverse and often counter-intuitive phenomena that arise in high-dimensional spaces while not occurring in low-dimensional ones. Among such phenomena, the one that affects nonparametric statistical inference the most is the inherent sparsity of high-dimensional data (Scott, 2015), which intuitively follows from the fact that volumes increase exponentially with dimensions. This implies that also the sample size required to obtain a given accuracy grows exponentially with dimensions, soon becoming unfeasible. Preprocessing the data to reduce their dimensionality is then key to most nonparametric methods.

In its traditional meaning, *dimensionality reduction* refers to a family of techniques transforming the data $y_i \in \mathbb{R}^p$ $(i = 1, \dots, n)$ into lower-dimensional representations lying in $\mathbb{R}^H$, with $H \ll p$. Such a mapping may be linear or not, with traditional methods as *principal component analysis* (PCA) and *factor analysis* falling within the linear case. These are still two of the most used dimensionality reduction techniques, despite dating back to Pearson (1901) and Spearman (1904) respectively, and until recently very few methods departed from the linear approach. However, in the late 1990s, techniques for nonlinear dimensionality reduction, also called *manifold learning*, started being proposed. These include *kernel PCA* (Schölkopf et al., 1998), *isomap* (Tenenbaum et al., 2000) and

local linear embedding (Roweis & Saul, 2000), just to name a few; see Lee & Verleysen (2007) and Van Der Maaten et al. (2009) for a review and comparison. Such nonlinear methods have been shown to outperform their linear counterparts on complex artificial datasets, such as a Swiss-roll set of points lying on a spiral-like two-dimensional manifold within $\mathbb{R}^3$. However, while performing better on simulated data sampled from smooth manifolds, nonlinear techniques often perform worse than PCA on real-world datasets (Van Der Maaten et al., 2009). This highlights that linear methods are still worthy of analysis, even after almost 120 years since their first appearance. In fact, Bayesian nonparametric versions of traditional linear methods have been recently proposed; e.g. see Bishop (1999) for Bayesian PCA and Ghosh & Dunson (2009) and Bhattacharya & Dunson (2011) for Bayesian factor models.

As mentioned at the beginning, the current data abundance concerns not only the number of features, but also sample sizes. While dimensionality reduction addresses the first issue, *clustering* is a way of dealing with the second one. Though these are two distinct problems, they both deal with compressing the data while preserving information, and they are closely intertwined. For a start, feature extraction may serve as a preprocessing step to alleviate the curse of dimensionality that affects clustering (Assent, 2012). Alternatively, the two tasks can be even pursued simultaneously (Tadesse et al., 2005; Bouveyron et al., 2007; Murphy et al., 2020). A further link among these problems is represented by latent class analysis (Lazarsfeld & Henry, 1968; Goodman, 1974), which models the features through categorical latent variables and may hence be regarded as a bridge between feature extraction and clustering.

Another notable connection between these two tasks is the problem of inferring the number H of latent dimensions or clusters. A traditional approach (Fraley & Raftery, 1998; Handcock et al., 2007) consists in performing inference for multiple values of H and then selecting the value that maximizes the Bayesian information criterion (BIC) (Schwarz, 1978). Such two-step procedures not only require multiple runs, but also ignore the uncertainty on H within each single run, thus potentially affecting inference. An alternative is optimizing the integrated completed likelihood (ICL) proposed by Biernacki et al. (2000) and refined, among others, by Bertoletti et al. (2015), Ryan et al. (2017) and Rastelli et al. (2018). The ICL-optimal clustering is fast to obtain and automatically includes an optimal number of clusters, but lacks uncertainty quantification.

Uncertainty quantification is instead provided by Bayesian nonparametric methods, which avoid fixing H by either considering infinite mixtures, where H grows with the data, or by placing a prior on H, giving rise to the so-called mixtures of finite mixtures (MFM). The prime example of infinite mixture is represented by Dirichlet process (DP) mixtures, for which simple Markov chain Monte Carlo (MCMC) algorithms (Neal, 1992,

2000) are available thanks to the various tractable representations of the DP. Many of these algorithms can be applied also to MFMs, as pointed out in Miller & Harrison (2018), thus avoiding reversible jump MCMC (Green, 1995), which is not trivial to adapt to new formulations, especially in high dimensions.

In the present thesis we develop new Bayesian nonparametric methodologies for both dimensionality reduction and clustering. For the first task, we focus on latent-variable models with an inherent ordering, where latent dimensions can be assumed to have a decreasing impact. Recalling Rousseau & Mengersen (2011), in such contexts it is useful to define overcomplete models endowed with shrinkage priors that favor the deletion of unnecessary dimensions. This avoids overfitting and allows to estimate the model dimension without recurring to reversible jump (Green, 1995) as in Lopes & West (2004), or to other computationally intensive techniques. Although classical shrinkage has proven useful in this context (West, 2003; Carvalho et al., 2008), increasing shrinkage priors can provide a more effective option that exploits the inherent ordering of dimensions to progressively penalize expansions with growing complexity. A notable example of this approach is the *multiplicative gamma process*, proposed by Bhattacharya & Dunson (2011) to induce increasing shrinkage of the loadings in infinite factor models. This prior exhibits appealing theoretical properties such as large support and weak consistency, admits a simple adaptive Gibbs sampler, and outperforms several competitors including lasso (Tibshirani, 1996), elastic-net (Zou & Hastie, 2005) and banding approaches (Bickel & Levina, 2008) in covariance matrix estimation. However, as discussed in Durante (2017), the multiplicative gamma process induces shrinkage only in expectation and for some values of the hyperparameters. Moreover, such hyperparameters determine both the rate of shrinkage and the prior for the retained dimensions. Motivated by these issues, we propose a novel increasing shrinkage prior named *cumulative shrinkage process* that induces increasing shrinkage in probability for any hyperparameter value, preserves large support and weak consistency for factor models, admits a simple adaptive Gibbs sampler, matches the performance of the multiplicative gamma process in covariance matrix estimation and outperforms it in recovering the latent dimension and in running time. Factor models are chosen here as a notable illustrative example, but decreasing-order structures can be found in many other settings, such as Poisson factorization (Dunson & Herring, 2005), mixture models (De Blasi et al., 2020) and latent space models for networks (Durante & Dunson, 2014, 2018; Durante et al., 2017a,b).

As for clustering, we focus on *community detection* in networks, which aims at grouping nodes with respect to connectivity patterns. Among the probabilistic models inducing such a community structure, the *stochastic block model* (SBM) stands out for its balance between simplicity and flexibility. Like the first feature extraction models, the first

SBM formulations (Holland et al., 1983; Nowicki & Snijders, 2001) required to set the number of communities to a fixed and finite value. The evolution of SBMs then kind of reflected that of mixture models, with the Dirichlet process serving as a building block for the first Bayesian nonparametric SBM, the infinite relational model (Kemp et al., 2006), while the mixture-of-finite-mixtures (MFM) approach of Miller & Harrison (2018) inspired the MFM-SBM by Geng et al. (2019). All these SBM formulations rely on priors within the Gibbs-type class (De Blasi et al., 2013a) and what makes them analytically and computationally tractable is not specific of the instances that are currently exploited, but rather underlies the whole Gibbs-type class. Leveraging this observation, we propose a general SBM framework based on Gibbs-type priors, the *extended stochastic block model* (ESBM), that not only encompasses existing formulations but also facilitates the proposal of new ones. Moreover, the ESBM allows to naturally include node attributes, quantify clustering uncertainty and compare models in a principled yet simple manner.

1.2 Factor analysis and stochastic block models

Even if apparently far apart, factor models and SBMs share profound similarities. In fact, they both induce low-rank factorizations of data-related matrices and highlight hidden block structures. In order to illustrate this structural affinity, we first consider Gaussian factor models, which assume that each individual data $y_i \in \mathbb{R}^p$ is generated from $y_i = \Lambda\eta_i + \epsilon_i$ $(i = 1, \dots, n)$, where $\Lambda = [\lambda_{jh}] \in \mathbb{R}^{p\times H}$ is the loadings matrix, $\eta_i \sim N_H(0, I_H)$ are the factor scores for individual i and $\epsilon_i \sim N_p(0, \Sigma)$ are the idiosyncratic errors. This induces the following factorization of the data covariance matrix $\Omega \in \mathbb{R}^{p\times p}$:

$$\Omega = \Lambda\Lambda^T + \Sigma. \tag{1.1}$$

Ideally, the latent dimensionality H is much smaller than p and each variable $j = 1, \dots, p$ is explained mainly by few latent factors. After appropriately reordering the variable indexes, this results in a block structure within the loadings matrix Λ and consequently, via (1.1), within the data covariance matrix Ω and its transformations. Such a block structure is visible for example in Figure 1.1, which is extracted from Chapter 2.

Under SBMs, block structures emerge directly in the data matrix after sorting rows and columns according to the underlying node partition. In this case, the data is represented by the $V \times V$ symmetric adjacency matrix Y of a binary undirected network with V nodes, with elements $y_{vu} = y_{uv} = 1$ if nodes v and u are connected and $y_{vu} = y_{uv} = 0$ otherwise. SBMs partition the nodes into H exhaustive and mutually exclusive communities, with nodes in the same community sharing a common connectivity pattern. Namely, for binary undirected networks without self-loops, each sub-diagonal entry y_{vu} of the adjacency

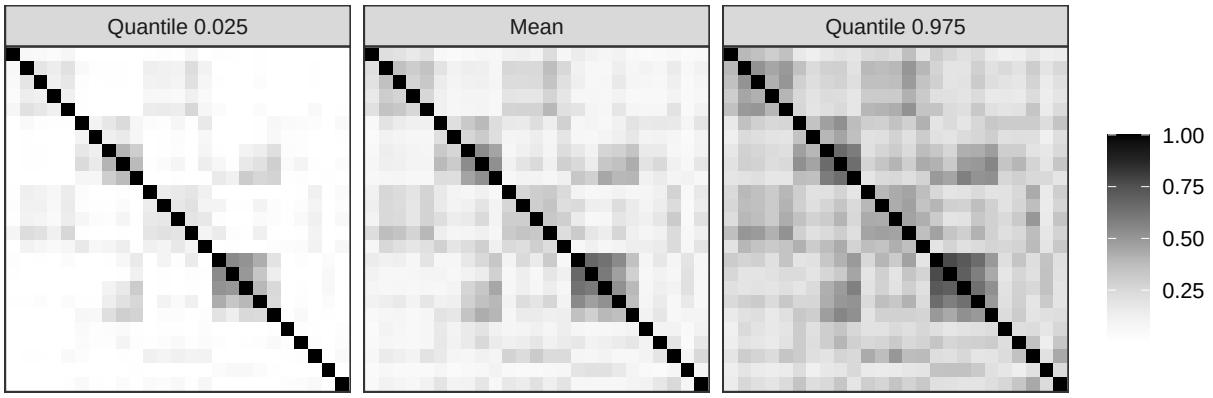

Figure 1.1: Posterior mean and credible intervals for each element of the absolute correlation matrix for the application in § 2.5.

matrix is independently modelled as a Bernoulli random variable with probability depending only on the community memberships of the involved nodes v and u. This corresponds to saying that, after reordering rows and columns with respect to the node clustering, the observed adjacency matrix Y is partitioned in blocks of conditionally i.i.d. entries. This is clearly visible in Figure 1.2, which is extracted from Chapter 3. Just like factor models, SBMs induce a factorization, in this case of the expected adjacency matrix. In fact, collecting cluster memberships in a $V \times H$ matrix Z with entries $z_{vh} = 1$ if node v is in cluster h and $z_{vh} = 0$ otherwise, and gathering probabilities θ_{hk} of connecting a node in cluster h with a node in cluster k in the $H \times H$ matrix Θ, we get a factorization of the expected adjacency matrix that closely reminds equation (1.1):

$$\mathbb{E}[Y \mid Z, \Theta] = Z\Theta Z^{T}. \tag{1.2}$$

A bridge between factor models and SBMs is represented by *random dot product graphs* (Young & Scheinerman, 2007), a very general class that encompasses SBMs and successfully approximates more complex latent position random graphs (Hoff et al., 2002); see Athreya et al. (2017) for a comprehensive survey. In random dot product graphs, the probability of a connection among nodes v and u is given by the dot product among their latent positions x_v and x_u in $\mathbb{R}^H$; a further generalization by Hoff (2008) substitutes the dot product with more general quadratic forms. SBMs are a special case of random dot product graphs in which all the nodes of the same community have the same latent position, thus discretizing the latent space. Defining a $V \times H$ matrix X where each row is the latent position of a node, random dot product graphs induce a factorization of the expected adjacency matrix that is similar to equations (1.1) and (1.2):

$$\mathbb{E}[Y \mid X] = XX^{T}. \tag{1.3}$$

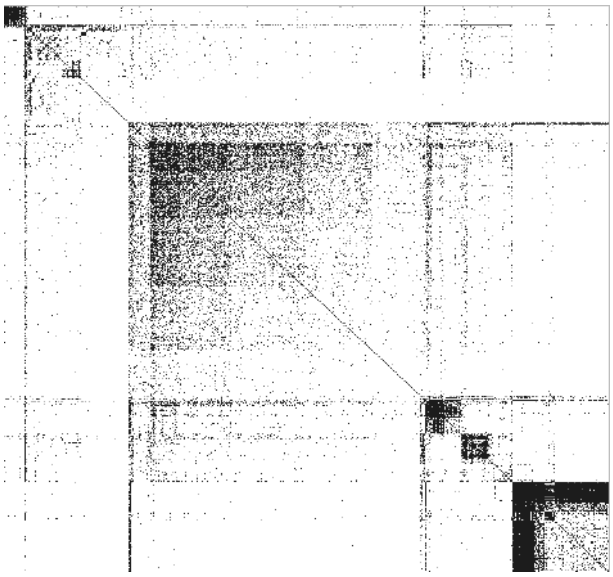

Figure 1.2: Observed adjacency matrix for the application in § 3.5, after sorting rows and columns according to the estimated node clustering.

An additional connection is represented by the fact that increasing shrinkage priors originally designed for factor models, such as the multiplicative gamma process by Bhattacharya & Dunson (2011), have then been employed in latent space models for networks (Durante & Dunson, 2014, 2018; Durante et al., 2017a,b).

In conclusion, the similarities among factor models and SBMs reflect the affinity between feature selection and clustering, which both aim at reducing the complexity of big data, even if from two different sides: the large-p and the large-n side, respectively. In the present thesis, we develop original Bayesian nonparametric methodologies for both problems, as detailed in the following section.

1.3 Summary of the specific contributions

Motivated by the above, we develop novel Bayesian nonparametric methodologies for feature selection (Chapter 2) and community detection in networks (Chapters 3–4).

Namely, in Chapter 2 we propose a novel increasing shrinkage prior named *cumulative shrinkage process*. Our construction has broad applicability, simple interpretation, and is based on a sequence of spike and slab distributions with increasing mass assigned to the spike as model complexity grows. Using factor analysis as an illustrative example, we show that this formulation has theoretical and practical advantages over current competitors, including an improved ability to recover the latent model dimension. We also propose an adaptive Gibbs sampler which tunes the latent dimensionality H as it progresses. This algorithm, together with the ability of the prior to favor the recovery of the number of active latent factors, allows for reduced running times with respect to

the non-adaptive Gibbs sampler. However, the increasing availability of large datasets demands for even faster algorithms. This need for scalability has pushed Bayesian statisticians towards approximate methods for posterior inference, including Laplace approximation, variational Bayes and expectation propagation (Bishop, 2006). In order to offer better scalability, here we propose a mean-field variational algorithm for Gaussian factor models endowed with a cumulative shrinkage process. Such a strategy provides comparable inference with respect to the adaptive Gibbs sampler and further reduces the running time. The derivation of the variational algorithm is facilitated by placing the cumulative shrinkage process directly on the loadings rather than on the loadings variance, as done to derive the adaptive Gibbs sampler. This simpler specification models both active and inactive loadings via Gaussians, which is slightly suboptimal with respect to the original formulation, that instead induced Student-t and Gaussian marginals on active and inactive loadings, respectively, thus facilitating the separation of active and inactive factors. However, this simpler version facilitates the derivation of a faster variational algorithm, while preserving the increasing shrinkage property.

In Chapter 3 we propose the *extended stochastic block model* (ESBM), a general framework for SBMs that relies on Gibbs-type priors (Gnedin & Pitman, 2005) and encompasses various existing SBM formulations (Nowicki & Snijders, 2001; Kemp et al., 2006; Geng et al., 2019) as special cases. It also includes unexplored options, among which we focus on a SBM based on the Gnedin process (Gnedin, 2010), that we show to have theoretical and practical advantages over current models. The proposed ESBM also allows to inform the block assignment mechanism with node attributes. In fact, leveraging the approach of Müller et al. (2011), we obtain modified full conditionals for a generic ESBM that favors the formation of clusters which are homogeneous with respect to node attributes. We also provide a collapsed Gibbs sampler which holds for the whole ESBM class, with or without covariates, and we outline methods for estimation and uncertainty quantification adapting the decision-theoretic approach of Wade & Ghahramani (2018) to the network setting. The performance of the ESBM is assessed in simulations and in an application to a bill co-sponsorship network in the Italian parliament, where we found interesting hidden blocks and core–periphery patterns.

In this bill co-sponsorship dataset, politicians are pre-clustered in parties. This kind of additional information is typically included within SBMs to supervise the community assignment mechanism (e.g. Tallberg, 2004; White & Murphy, 2016; Newman & Clauset, 2016; Stanley et al., 2019) or improve the inference on edge probabilities (e.g. Mariadassou et al., 2010; Choi et al., 2012; Sweet, 2015; Roy et al., 2019). Although these solutions are routinely implemented, there is a lack of formal approaches to test whether exogenous node partitions are coherent with the latent node communities inferred from the observed

connectivity patterns via SBMs. To address this gap, in Chapter 4 we develop a Bayesian testing procedure which relies on the calculation of the Bayes factor (e.g. Kass & Raftery, 1995) between a SBM with known community structure coinciding with the exogenous node partition and an infinite relational model (Kemp et al., 2006) that allows the endogenous community assignments to be unknown and random. A simple MCMC method for computing the Bayes factor and quantifying uncertainty in the endogenous communities is proposed. This routine is evaluated in simulations and in an application to brain networks of Alzheimer's patients.

The contents of Chapters 2–4 correspond to the following papers of ours:

- S. Legramanti, D. Durante & D. B. Dunson (2020), Bayesian Cumulative Shrinkage for Infinite Factorizations. *Biometrika, 107(3), 745-752*;
- S. Legramanti (2020), Variational Bayes for Gaussian Factor Models under the Cumulative Shrinkage Process. *Book of short papers SIS 2020, 416-420*;
- S. Legramanti, T. Rigon, D. Durante & D. B. Dunson (2020+), Extended Stochastic Block Models. *Submitted*;
- S. Legramanti, T. Rigon & D. Durante (2020+), Bayesian Testing for Exogenous Partition Structures in Stochastic Block Models. *Sankhya A, in press*;

and codes are available at `https://github.com/sirio legramanti`.

A discussion of possible extensions of the ideas above is provided in Chapter 5.

Chapter 2

Bayesian cumulative shrinkage for infinite factorizations

2.1 Introduction

There has been a considerable interest in shrinkage priors for high dimensional parameters (e.g., Ishwaran & Rao, 2005; Carvalho et al., 2010) but most of the focus has been on regression, where there is no natural ordering in the coefficients. There are several settings, however, where an order is present and desirable. Indeed, in statistical models relying on low-rank factorizations or basis expansions, such as factor models and tensor factorizations, it is natural to expect that additional dimensions play a progressively less important role in characterizing the data or model structure, and hence the associated parameters should have a stochastically decreasing effect. Such a behavior can be induced through increasing shrinkage priors. For instance, in the context of Bayesian factor models an example of this approach can be found in the multiplicative gamma process developed by Bhattacharya & Dunson (2011) to penalize the effect of additional factor loadings via a cumulative product of gamma priors for their precision. Although this prior has been widely applied, there are practical disadvantages that motivate consideration of alternative solutions (Durante, 2017). In general, despite the importance of increasing shrinkage priors in many factorization models, the methods, theory and computational strategies for these priors remain under-developed.

Motivated by the above considerations, we propose a novel increasing shrinkage prior, the cumulative shrinkage process, which is broadly applicable, while having simple and parsimonious structure. The proposed prior induces increasing shrinkage via a sequence of spike and slab distributions assigning growing mass to the spike as model complexity grows. In Definition 2.1, we present this prior for the general case in which the effect of the hth dimension is controlled by a scalar parameter $\theta_h \in \mathbb{R}$, so that redundant terms can be essentially deleted by progressively shrinking the sequence

$\theta = \{\theta_h \in \Theta \subseteq \mathbb{R} : h = 1, 2, \ldots\}$ towards an appropriate value $\theta_\infty \in \mathbb{R}$. For example, in factor models $\theta_h \in \mathbb{R}_+$ may denote the variance of the loadings for the hth factor, and the goal is to define a prior on these terms which favors stochastically decreasing impact of the factors via increasing concentration of the loadings near zero as h grows.

Definition 2.1. Let $\theta = \{\theta_h \in \Theta \subseteq \mathbb{R} : h = 1, 2, \ldots\}$ denote a countable sequence of parameters. We say that θ is distributed according to a cumulative shrinkage process with parameter $\alpha > 0$, starting slab distribution P_0 and target spike δ_{θ_∞} if, conditionally on $\pi = \{\pi_h \in (0, 1) : h = 1, 2, \ldots\}$, each θ_h is independent and has the following spike and slab distribution:

$$(\theta_h \mid \pi_h) \sim P_h = (1 - \pi_h)P_0 + \pi_h \delta_{\theta_\infty}, \quad \pi_h = \sum_{l=1}^{h} \omega_l, \quad \omega_l = \nu_l \prod_{m=1}^{l-1} (1 - \nu_m), \quad (2.1)$$

where $\nu_1, \nu_2, \ldots$ are independent Beta$(1, \alpha)$ variables, P_0 is a diffuse continuous distribution and δ_{θ_∞} is a point mass at θ_∞.

Equation (2.1) exploits the stick-breaking construction of the Dirichlet process (Ishwaran & James, 2001). This implies that the probability π_h assigned to the spike δ_{θ_∞} increases with the model dimension h, and that $\lim_{h\to\infty} \pi_h = 1$ almost surely. Hence, as complexity grows, P_h increasingly concentrates around θ_∞, which is specified to facilitate the deletion of redundant terms, while the slab P_0 corresponds to the prior on the active parameters. Note that definition 2.1 can be extended to sequences in $\mathbb{R}^p$. Moreover, δ_{θ_∞} can be replaced with a continuous distribution, without affecting the key properties of the prior presented in § 2.2. A continuous spike can be particularly convenient in some contexts; for example, in § 2.6.1 we illustrate how this can facilitate variational inference. As we will discuss in § 2.2 and in § 2.3.1, it is also possible to restrict Definition 2.1 to finitely many terms $(\theta_1, \ldots, \theta_H)$ by letting $\nu_H = 1$. In practical implementations, this truncated version typically ensures full flexibility if H is set to a conservative upper bound, but this value can be extremely large in several high dimensional settings, thus motivating our initial focus on the infinite expansion and its theoretical properties.

2.2 General properties of the cumulative shrinkage process

We first motivate our cumulative stick-breaking construction for the sequence π that controls the mass assigned to the spike in (2.1) as a function of model dimension. Indeed, one could alternatively consider pre-specified non-decreasing functions bounded between 0 and 1. However, we have found that such specifications are overly-restrictive and have worse practical performance. Definition 2.1 is purposely chosen to be effectively

nonparametric, with Proposition 2.1 showing that the prior has large support on the space of non-decreasing sequences taking values in $(0,1)$.

Proposition 2.1. *Let Π be the probability measure induced on $\pi = \{\pi_h \in (0,1) : h = 1,2,\ldots\}$ by (2.1), then Π has large support on the whole space of non-decreasing sequences taking values in $(0,1)$.*

Proof. Since the mapping from the sequence $w = \{w_h \in (0,1) : h = 1,2,\ldots\}$ to $\pi = \{\pi_h \in (0,1) : h = 1,2,\ldots\}$ is one-to-one, it is sufficient to ensure that the stick-breaking prior for w has full support on the infinite dimensional simplex. This result is proved by Bissiri & Ongaro (2014) in § 3.2. □

Besides being fully flexible, our construction for π also has simple interpretation and allows control over shrinkage via an interpretable parameter α, as stated in Proposition 2.2 and in the subsequent results.

Proposition 2.2. *Each π_h in (2.1) coincides with the proportion of the total variation distance between the slab and the spike covered up to step h, in the sense that $\pi_h = d_{TV}(P_0, P_h)/d_{TV}(P_0, \delta_{\theta_\infty})$.*

Proof. The proof adapts the one of Theorem 1 in Canale et al. (2018). In fact, under Definition 2.1, the distance $d_{TV}(P_0, P_h)$ on the Borel σ-algebra in $\mathbb{R}$ is

$$\begin{aligned} d_{TV}(P_0, P_h) &= \sup_{A \in \mathcal{B}(\mathbb{R})} |P_0(A) - P_h(A)| = \sup_{A \in \mathcal{B}(\mathbb{R})} |P_0(A) - (1-\pi_h)P_0(A) - \pi_h \delta_{\theta_\infty}(A)| = \\ &= \pi_h \sup_{A \in \mathcal{B}(\mathbb{R})} |P_0(A) - \delta_{\theta_\infty}(A)| = \pi_h d_{TV}(P_0, \delta_{\theta_\infty}). \end{aligned}$$

□

Using similar arguments, we can obtain analogous expressions for ω_h and ν_h, which represent the proportions of the total $d_{TV}(P_0, \delta_{\theta_\infty})$ and the remaining $d_{TV}(P_{h-1}, \delta_{\theta_\infty})$, respectively, covered between steps $h-1$ and h. Specifically, $\omega_h = d_{TV}(P_{h-1}, P_h)/d_{TV}(P_0, \delta_{\theta_\infty})$ and $\nu_h = d_{TV}(P_{h-1}, P_h)/d_{TV}(P_{h-1}, \delta_{\theta_\infty})$ for every h. The expectations of these quantities are explicitly available as

$$E(\nu_h) = \frac{1}{1+\alpha}, \quad E(\omega_h) = \frac{\alpha^{h-1}}{(1+\alpha)^h}, \quad E(\pi_h) = 1 - \frac{\alpha^h}{(1+\alpha)^h} \quad (h = 1,2,\ldots). \tag{2.2}$$

Moreover, combining (2.2) with Definition 2.1, the expectation of θ_h $(h = 1,2,\ldots)$ is

$$E(\theta_h) = E\{E(\theta_h \mid \pi_h)\} = \{1 - E(\pi_h)\}\theta_0 + E(\pi_h)\theta_\infty = \theta_\infty + \{\alpha(1+\alpha)^{-1}\}^h(\theta_0 - \theta_\infty), \tag{2.3}$$

where θ_0 defines the expected value under the slab P_0. Hence, as h grows, the prior expectation of θ_h converges exponentially towards the spike location θ_∞. As stated in

Lemma 2.1, a stronger notion of cumulative shrinkage in distribution, beyond simple concentration in expectation, also holds under (2.1).

Lemma 2.1. *Let* $\mathbb{B}_\varepsilon(\theta_\infty) = \{\theta_h \in \Theta \subseteq \mathbb{R} : |\theta_h - \theta_\infty| \leqslant \varepsilon\}$ *denote an* ε*-neighborhood around* θ_∞ *with radius* $\varepsilon > 0$*, and define with* $\overline{\mathbb{B}}_\varepsilon(\theta_\infty)$ *the complement of* $\mathbb{B}_\varepsilon(\theta_\infty)$*. Then, for any* $h = 1, 2, \ldots$ *and* $\varepsilon > 0$*,*

$$\text{pr}(|\theta_h - \theta_\infty| > \varepsilon) = P_0\{\overline{\mathbb{B}}_\varepsilon(\theta_\infty)\}\{\alpha(1+\alpha)^{-1}\}^h. \tag{2.4}$$

Therefore, $\text{pr}(|\theta_{h+1} - \theta_\infty| \leqslant \varepsilon) > \text{pr}(|\theta_h - \theta_\infty| \leqslant \varepsilon)$ *for any* $\alpha > 0$*,* $h = 1, 2, \ldots$ *and* $\varepsilon > 0$*.*

Proof. Notice that, for each h, $\text{pr}(|\theta_h - \theta_\infty| > \varepsilon)$ can be expressed as

$$E[P_h\{\overline{\mathbb{B}}_\varepsilon(\theta_\infty)\}] = E[(1-\pi_h)P_0\{\overline{\mathbb{B}}_\varepsilon(\theta_\infty)\} + \pi_h\delta_{\theta_\infty}\{\overline{\mathbb{B}}_\varepsilon(\theta_\infty)\}] = P_0\{\overline{\mathbb{B}}_\varepsilon(\theta_\infty)\}\{1 - E(\pi_h)\}.$$

Therefore, replacing $E(\pi_h)$ with its expression in equation (2.2) leads to (2.4). To prove that $\text{pr}(|\theta_{h+1} - \theta_\infty| \leqslant \varepsilon) > \text{pr}(|\theta_h - \theta_\infty| \leqslant \varepsilon)$ it is sufficient to note that $\{\alpha(1+\alpha)^{-1}\}^{h+1} < \{\alpha(1+\alpha)^{-1}\}^h$. □

Equations (2.2)–(2.4) highlight how the rate of increasing shrinkage is controlled by α. In particular, lower values of α induce faster concentration around θ_∞ and hence more rapid deletion of the redundant terms. This control over the rate of increasing shrinkage via α is separated from the specification of the slab P_0, thereby allowing flexible modelling of the active terms. As discussed in Durante (2017), such a separation does not hold, for example, in the multiplicative gamma process (Bhattacharya & Dunson, 2011) whose hyperparameters control both the rate of shrinkage and the prior for the active factors. This creates a trade-off between the need to maintain diffuse priors for the active terms and the attempt to shrink the redundant ones. Moreover, increasing shrinkage holds only in expectation and for specific hyperparameters.

Instead, our prior ensures increasing shrinkage in distribution for any α, and can model any prior expectation on the number of active terms. In fact, α is equal to the prior mean of the number of terms in θ modelled via the slab P_0. This result follows after noticing that $(\theta_h \mid \pi_h)$ in (2.1) can be alternatively obtained by marginalizing out the augmented indicator $c_h \sim \text{Bern}(1-\pi_h)$ in $(\theta_h \mid c_h) \sim c_h P_0 + (1-c_h)\delta_{\theta_\infty}$. Hence, $H^* = \sum_{h=1}^\infty c_h$ counts the number of active elements in θ, and its prior mean is

$$E(H^*) = \sum_{h=1}^\infty E(c_h) = \sum_{h=1}^\infty E\{E(c_h \mid \pi_h)\} = \sum_{h=1}^\infty E(1-\pi_h) = \sum_{h=1}^\infty \{\alpha(1+\alpha)^{-1}\}^h = \alpha.$$

It follows that α should be set to the expected number of active terms, while P_0 should be sufficiently diffuse to model active components, and θ_∞ should be chosen to facilitate the deletion of redundant ones.

Recalling Bhattacharya & Dunson (2011) and Rousseau & Mengersen (2011), it is useful to define models with more than enough components and then choose shrinkage priors which favor effective deletion of the unnecessary ones. This choice protects against over-fitting and allows estimation of model dimension, bypassing the need for reversible jump (Lopes & West, 2004) or other computationally intensive strategies. Our cumulative shrinkage process in (2.1) provides a useful prior for this purpose. As discussed in § 2.1, it is straightforward to modify Definition 2.1 to instead restrict to H components, by letting $v_H = 1$, with H a conservative upper bound. Theorem 2.1 shows that truncating θ to H components, by setting $\theta_h = 0$ for $h > H$, provides a sequence $\theta^{(H)}$ that accurately approximates the infinite sequence in (2.1) for H large.

Theorem 2.1. *If the countable sequence θ has prior (2.1), then, for any index H and $\varepsilon \geqslant |\theta_\infty|$,*

$$\text{pr}\{d_\infty(\theta, \theta^{(H)}) > \varepsilon\} = \text{pr}\{\sup(|\theta_h| : h = H+1, H+2, \ldots) > \varepsilon\} \leqslant P_0\{\overline{\mathbb{B}}_\varepsilon(0)\}\alpha\{\alpha(1+\alpha)^{-1}\}^H,$$

where d_∞ is the sup-norm distance and $\overline{\mathbb{B}}_\varepsilon(0) = \{\theta_h \in \Theta \subseteq \mathbb{R} : |\theta_h| > \varepsilon\}$.

Proof. The proof follows after noting that $\text{pr}(\sup_{h>H} |\theta_h| > \varepsilon) = \text{pr}\{\cup_{h>H}(|\theta_h| > \varepsilon)\}$, and that $\delta_{\theta_\infty}\{\overline{\mathbb{B}}_\varepsilon(0)\} = 0$ for any $\varepsilon \geqslant |\theta_\infty|$. Hence, adapting the proof of Lemma 2.1, we obtain that

$$\text{pr}\{\cup_{h>H}(|\theta_h| > \varepsilon)\} \leqslant \sum_{h=H+1}^{\infty} \text{pr}(|\theta_h| > \varepsilon) = P_0\{\overline{\mathbb{B}}_\varepsilon(0)\} \sum_{h=H+1}^{\infty} \{\alpha(1+\alpha)^{-1}\}^h.$$

To conclude the proof, notice that $\sum_{h=H+1}^{\infty}\{\alpha(1+\alpha)^{-1}\}^h = \alpha\{\alpha(1+\alpha)^{-1}\}^H$. □

Hence, the prior probability of $\theta^{(H)}$ being close to θ converges to one at a rate which is exponential in H, thus justifying posterior inference under finite sequences based on a conservative H. Although the above bound holds for $\varepsilon \geqslant |\theta_\infty|$, in general θ_∞ is set close to zero. Hence, Theorem 2.1 is valid also for small ε.

2.3 Cumulative shrinkage process for Gaussian factor models

2.3.1 Model formulation and prior specification

Definition 2.1 provides a general prior which can be used in different models (e.g., Gopalan et al., 2014) under appropriate choices of P_0 and θ_∞. Here, we focus on Gaussian sparse factor models as an important special case to illustrate our approach. We will compare primarily to the multiplicative gamma process, which has been devised specifically for this class of models and was shown to have practical gains in this context relative to several competitors, including the use of lasso (Tibshirani, 1996), elastic-net

(Zou & Hastie, 2005) and banding approaches (Bickel & Levina, 2008). Although there are other priors for sparse factor models (e.g., Carvalho et al., 2008; Knowles & Ghahramani, 2011), these choices have practical disadvantages relative to the multiplicative gamma process, so they will not be considered further here.

The focus will be on learning the structure of the $p \times p$ covariance matrix $\Omega = \Lambda\Lambda^T + \Sigma$ for the data $y_i = (y_{i1}, \ldots, y_{ip})^T \in \mathbb{R}^p$ generated from the Gaussian factor model $y_i = \Lambda\eta_i + \epsilon_i$, with $\eta_{ih} \sim N(0,1)$, $(i = 1, \ldots, n; h = 1, 2, \ldots)$, $\epsilon_i \sim N_p(0, \Sigma)$ $(i = 1, \ldots, n)$ and $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. To perform Bayesian inference for this model, Bhattacharya & Dunson (2011) assume $\sigma_j^2 \sim \mathrm{InvGa}(a_\sigma, b_\sigma)$ $(j = 1, \ldots, p)$, and $(\lambda_{jh} \mid \phi_{jh}, \theta_h) \sim N(0, \phi_{jh}\theta_h)$ $(j = 1, \ldots, p; h = 1, 2, \ldots)$ with scales ϕ_{jh} from independent $\mathrm{InvGa}(\nu/2, \nu/2)$ priors and global variances θ_h having multiplicative gamma process prior

$$\theta_h = \prod_{l=1}^{h} \vartheta_l^{-1} \quad (h = 1, 2, \ldots), \quad \vartheta_1 \sim \mathrm{Ga}(a_1, 1), \quad \vartheta_l \sim \mathrm{Ga}(a_2, 1) \quad (l = 2, 3, \ldots). \tag{2.5}$$

Specific choices of (a_1, a_2) in (2.5) ensure that $E(\theta_h)$ decreases with h, thus allowing increasing shrinkage of the loadings as h grows. Instead, we keep $\sigma_j^2 \sim \mathrm{InvGa}(a_\sigma, b_\sigma)$ $(j = 1, \ldots, p)$ and define our cumulative shrinkage process prior by letting $(\lambda_{jh} \mid \theta_h) \sim N(0, \theta_h)$ $(j = 1, \ldots, p;\ h = 1, 2, \ldots)$ with

$$(\theta_h \mid \pi_h) \sim (1 - \pi_h)\mathrm{InvGa}(a_\theta, b_\theta) + \pi_h \delta_{\theta_\infty}, \quad \pi_h = \sum_{l=1}^{h} \omega_l, \quad \omega_l = \nu_l \prod_{m=1}^{l-1} (1 - \nu_m), \tag{2.6}$$

where $\nu_1, \nu_2, \ldots$ are independent $\mathrm{Beta}(1, \alpha)$. Integrating out θ_h, each loading λ_{jh} has the marginal prior $(1 - \pi_h)t_{2a_\theta}(0, b_\theta/a_\theta) + \pi_h N(0, \theta_\infty)$, where $t_{2a_\theta}(0, b_\theta/a_\theta)$ denotes the Student-t distribution with $2a_\theta$ degrees of freedom, location 0 and scale b_θ/a_θ. Hence, θ_∞ should be set close to zero to allow effective shrinkage of redundant factors, while (a_θ, b_θ) should be specified so as to induce a moderately diffuse prior with scale b_θ/a_θ for the active loadings. Although the choice $\theta_\infty = 0$ is possible, we follow Ishwaran & Rao (2005) by suggesting $\theta_\infty > 0$ to induce a continuous shrinkage prior on every λ_{jh} which improves mixing and identification of the inactive factors. Exploiting the marginals for λ_{jh}, it also follows that, if $b_\theta/a_\theta > \theta_\infty$ then $\mathrm{pr}(|\lambda_{j,h+1}| \leqslant \varepsilon) > \mathrm{pr}(|\lambda_{jh}| \leqslant \varepsilon)$ for each $j = 1, \ldots, p$, $h = 1, 2, \ldots$ and $\varepsilon > 0$. This allows cumulative shrinkage in distribution also for the loadings, and provides guidelines on (a_θ, b_θ) and θ_∞. Additional discussion on prior elicitation and empirical studies on sensitivity can be found in § 2.4.

To implement the analysis, we require a truncation H on the number of factors needed to characterize Ω, as discussed in § 2. Theorem 2.2 states that our shrinkage process truncated at H terms induces a well-defined prior for Ω with full-support, under the sufficient conditions that H is greater than the *true* H_0, and $E(\theta_h) < \infty$ for any h. These conditions are met when considering up to p active factors, with $a_\theta > 1$ and $\theta_\infty < \infty$.

Theorem 2.2. *Let Ω_0 be any $p \times p$ covariance matrix and define with Π the prior probability measure on $p \times p$ covariance matrices Ω induced by a Bayesian factor model having prior (2.6) on θ, truncated at H with $\nu_H = 1$. If $E(\theta_h) < \infty$ for any $h \geqslant 1$, then $\Pi\{\Omega \in \mathbb{R}^{p\times p} : \Omega$ has finite entries and is positive semi-definite$\} = 1$. In addition, if there exists a decomposition $\Omega_0 = \Lambda_0\Lambda_0^T + \Sigma_0$, such that $\Lambda_0 \in \mathbb{R}^{p\times H_0}$ and $H_0 < H$, then $\Pi\{B_\varepsilon^\infty(\Omega_0)\} > 0$ for any $\varepsilon > 0$, where $B_\varepsilon^\infty(\Omega_0)$ is an ε-neighborhood of Ω_0 under the sup-norm.*

Proof. We first prove that, for Gaussian factor models in § 2.3.1 with prior (2.6) truncated at H terms, $\Pi\{\Omega \in \mathbb{R}^{p\times p} : \Omega$ has finite entries and is positive semi-definite$\} = 1$. Since, by construction, Σ is diagonal with almost surely finite and non-negative entries, and $\Lambda\Lambda^T$ is trivially positive semi-definite, we only need to ensure that each entry $\lambda_{r\cdot}\lambda_{j\cdot}^T$ in $\Lambda\Lambda^T$ is almost surely finite. By the Cauchy-Schwartz inequality we obtain $|\lambda_{r\cdot}\lambda_{j\cdot}^T| \leqslant \|\lambda_{r\cdot}\|\|\lambda_{j\cdot}\| \leqslant \max_{1\leqslant j\leqslant p}\|\lambda_{j\cdot}\|^2$. Under the factor model in § 2.3.1 with prior (2.6) truncated at H terms, we have that

$$E(\|\lambda_{j\cdot}\|^2) = \sum_{h=1}^{H} E(\lambda_{jh}^2) = \sum_{h=1}^{H} E\{E(\lambda_{jh}^2 \mid \theta_h)\} = \sum_{h=1}^{H} E(\theta_h),$$

for every $j = 1, \ldots, p$, including the index of the maximum, thus ensuring that each entry in $\Lambda\Lambda^T$ is almost surely finite under the sufficient condition that $E(\theta_h) < \infty$ $(h = 1, \ldots, H)$. This holds when $a_\theta > 1$ and $\theta_\infty < \infty$.

Let us now prove the full support for Π. Since $H > H_0$, there always exists a $\Lambda \in \mathbb{R}^{p\times H}$ and a positive diagonal matrix Σ such that $\Lambda\Lambda^T + \Sigma = \Lambda_0\Lambda_0^T + \Sigma_0$. For instance, one can let $\Sigma = \Sigma_0$ and $\Lambda = [\Lambda_0, 0_{p\times(H-H_0)}]$. Hence, it suffices to prove full support for the priors induced on Λ and Σ by the truncated version of our cumulative shrinkage process. Such a property easily holds for Σ, whose diagonal elements σ_j^2 $(j = 1, \ldots, p)$ have independent inverse-gamma priors. Moreover, adapting the proof of Proposition 2 in Bhattacharya & Dunson (2011), full support can be proved also for the prior induced on Λ. Indeed, recalling § 2.3.1, we have that $\text{pr}\{\sum_{j=1}^{p}\sum_{h=1}^{H}(\lambda_{jh} - \lambda_{0jh})^2 < \epsilon_1^2\} \geqslant \text{pr}\{(\lambda_{jh} - \lambda_{0jh})^2 < \epsilon_1^2/(pH),$ for all $j = 1, \ldots, p;\ h = 1, \ldots, H\}$ with

$$\begin{aligned}&\text{pr}\{(\lambda_{jh} - \lambda_{0jh})^2 < \epsilon_1^2/(pH) \text{ for all } j = 1, \ldots, p;\ h = 1, \ldots, H\}\\ &= E\left[\prod_{j=1}^{p}\prod_{h=1}^{H}\text{pr}\{(\lambda_{jh} - \lambda_{0jh})^2 < \epsilon_1^2/(pH)|\theta\}\right] > 0.\end{aligned}$$

In fact, conditioned on $\theta = (\theta_1, \ldots, \theta_H)$, each λ_{jh} has independent $N(0, \theta_h)$ distribution. □

Recalling Theorem 2 in Bhattacharya & Dunson (2011), this result is also sufficient to ensure that the posterior of Ω is weakly consistent (Schwartz, 1965).

2.3.2 Posterior computation via Gibbs sampling

Posterior inference for the factor model in § 2.3.1 with cumulative shrinkage process (2.6) truncated at H terms for the loadings, proceeds via a Gibbs sampler cycling across the steps in Algorithm 1. This sampler relies on a data augmentation which exploits the fact that prior (2.6) can be obtained by marginalizing out the independent indicators z_h $(h = 1, \dots, H)$ with probabilities $\text{pr}(z_h = l \mid \omega_l) = \omega_l$ $(l = 1, \dots, H)$ in

$$(\theta_h \mid z_h) \sim \{1 - \mathbb{1}(z_h \leqslant h)\}\text{InvGa}(a_\theta, b_\theta) + \mathbb{1}(z_h \leqslant h)\delta_{\theta_\infty}, \tag{2.7}$$

where $\mathbb{1}(z_h \leqslant h) = 1$ if $z_h \leqslant h$ and 0 otherwise. As is clear from Algorithm 1, conditioned on $z_1, \dots, z_H$, it is possible to sample from conjugate full-conditionals, whereas the updating of the augmented data relies on the full-conditional distribution

$$\text{pr}(z_h = l \mid -) \propto \begin{cases} \omega_l N_p(\lambda_h; 0, \theta_\infty I_p), & \text{for } l = 1, \dots, h, \\ \omega_l t_{2a_\theta}\{\lambda_h; 0, (b_\theta/a_\theta) I_p\}, & \text{for } l = h+1, \dots, H, \end{cases} \tag{2.8}$$

where $N_p(\lambda_h; 0, \theta_\infty I_p)$ and $t_{2a_\theta}\{\lambda_h; 0, (b_\theta/a_\theta) I_p\}$ are the densities of p-variate Gaussian and Student-t distributions, respectively, evaluated at $\lambda_h = (\lambda_{1h}, \dots, \lambda_{ph})^T$. Equations (2.8) are obtained by marginalizing out θ_h, distributed as in (2.7), from the joint $N_p(\lambda_h; 0, \theta_h I_p)$. These calculations are straightforward in a variety of Bayesian models based on conditionally conjugate constructions, thus making (2.1) a general prior which can be easily incorporated, for instance, in Poisson factorizations (Gopalan et al., 2014).

2.3.3 Tuning the truncation index via adaptive Gibbs sampling

Recalling § 2.3.1, it is reasonable to perform Bayesian inference with at most p factors. Under our cumulative shrinkage process truncated at H terms this translates into $H \leqslant p + 1$, since there are at most $H - 1$ active factors, with the Hth one modelled with the spike by construction. However, this choice is too conservative, since we expect substantially fewer active factors than p, especially when p is very large. Hence, running Algorithm 1 with $H = p + 1$ would be computationally inefficient, since most of the columns in Λ would be modelled by the spike, thus providing a negligible contribution to the factorization of Ω.

Bhattacharya & Dunson (2011) address this issue via an adaptive Gibbs sampler tuning H as the sampler proceeds. To satisfy the diminishing adaptation condition in Roberts & Rosenthal (2007), they adapt H at iteration t with probability $p(t) = \exp(\alpha_0 + \alpha_1 t)$, where $\alpha_0 \leqslant 0$ and $\alpha_1 < 0$. The adaptation consists in dropping the inactive columns of Λ, if any, together with the corresponding parameters. If instead all columns are active, an extra factor is added, sampling the associated parameters from the prior.

Algorithm 1: One cycle of the Gibbs sampler for factor models with the cumulative shrinkage process

1 **for** *j from 1 to p* **do**
  sample the jth row of Λ from $N_H(V_j \eta^T \sigma_j^{-2} y_j, V_j)$, with $V_j = (D^{-1} + \sigma_j^{-2} \eta^T \eta)^{-1}$, $D = \text{diag}(\theta_1, \ldots, \theta_H)$, $\eta = (\eta_1, \ldots, \eta_n)^T$ and $y_j = (y_{1j}, \ldots, y_{nj})^T$;
2 **for** *j from 1 to p* **do**
  sample σ_j^2 from $\text{InvGa}\{a_\sigma + n/2, b_\sigma + \frac{1}{2}\sum_{i=1}^n (y_{ij} - \sum_{h=1}^H \lambda_{jh}\eta_{ih})^2\}$;
3 **for** *i from 1 to n* **do**
  sample η_i from $N_H\{(I_H + \Lambda^T\Sigma^{-1}\Lambda)^{-1}\Lambda^T\Sigma^{-1}y_i, (I_H + \Lambda^T\Sigma^{-1}\Lambda)^{-1}\}$;
4 **for** *h from 1 to H* **do**
  sample z_h from the categorical distribution with probabilities as in (2.8);
5 **for** *l from 1 to* $(H-1)$ **do**
  sample v_l from $\text{Beta}\{1 + \sum_{h=1}^H \mathbb{1}(z_h = l), \alpha + \sum_{h=1}^H \mathbb{1}(z_h > l)\}$;
  set $v_H = 1$ and update $\omega_1, \ldots, \omega_H$ from $v_1, \ldots, v_H$ through (2.6);
6 **for** *h from 1 to H* **do**
  if $z_h \leqslant h$ **then** $\theta_h = \theta_\infty$ **else** sample θ_h from $\text{InvGa}(a_\theta + p/2, b_\theta + \frac{1}{2}\sum_{j=1}^p \lambda_{jh}^2)$;
Output at the end of one cycle: one sample from the posterior of $\Omega = \Lambda\Lambda^T + \Sigma$.

This idea can be also implemented for the cumulative shrinkage process, as illustrated in Algorithm 2. Under our prior, the inactive Λ columns are naturally identified as those modelled by the spike and, hence, have index h such that $z_h \leqslant h$. Under the multiplicative gamma process, instead, a column is flagged as inactive if all its entries are within distance ϵ from zero. This ϵ plays a similar role as our spike location θ_∞. Indeed, lower values of ϵ and θ_∞ make it harder to discard inactive columns, thus affecting running time. Hence, although fixing θ_∞ close to zero is a key to enforce shrinkage, excessively low values should be avoided. Since under a truncated cumulative shrinkage process the number of active factors H^* is at most $H-1$, we increase H by one when $H^* = H-1$, and we decrease H to H^*+1 when $H^* < H-1$.

In our implementation no adaptation is allowed before a fixed number $\bar{t}$ of iterations to let the chain stabilize, while H and H^* are initialized to $p+1$ and p, which is the maximum possible rank for Ω. Further guidance for the choice of H can be obtained by monitoring how close $E(\pi_H)$ is to 1, via (2.2).

2.4 Simulation studies

We consider illustrative simulations to assess performance in learning the structure of the true covariance matrix $\Omega_0 = \Lambda_0\Lambda_0^T + \Sigma_0$ for the data y_i $(i = 1, \ldots, n)$ from a Gaussian factor model, with $\Sigma_0 = I_p$ and the entries in $\Lambda_0 \in \mathbb{R}^{p \times H_0}$ drawn from

Algorithm 2: One cycle of the adaptive version for the Gibbs sampler in Algorithm 1

Let t be the cycle number and $H^{(t)}$ the truncation index at cycle t.

perform a cycle of Algorithm 1;

if $t \geqslant \bar{t}$ **then**

  adapt with probability $p(t) = \exp(\alpha_0 + \alpha_1 t)$ as follows

  if $H^{*(t)} = \sum_h \mathbb{1}(z_h^{(t)} > h) < H^{(t-1)} - 1$ **then**

    set $H^{(t)} = H^{*(t)} + 1$, drop the inactive columns in Λ together with the associated parameters in η, θ, w, and add a final component to Λ, η, θ, w sampled from the prior;

  else

    set $H^{(t)} = H^{(t-1)} + 1$, add a final column sampled from the spike to Λ, and add the associated parameters to η, θ and w sampling from the corresponding priors;

Output: one sample from the posterior of $\Omega = \Lambda\Lambda^T + \Sigma$ and a value for H^*.

Table 2.1: Performance of CUSP and MGP in 25 simulations for different (p, H_0) scenarios

(p, H_0)	method	MSE		$E(H^* \mid y)$		averaged ESS	runtime (s)
		median	IQR	median	IQR	median	median
(20,5)	CUSP	0.75	0.29	5.00	0.00	655.04	310.76
	MGP	0.75	0.32	19.69	0.21	547.23	616.61
(50,10)	CUSP	2.25	0.33	10.00	0.00	273.55	716.23
	MGP	2.26	0.28	28.64	1.94	251.35	1845.88
(100,15)	CUSP	3.76	0.40	15.00	0.00	175.26	2284.87
	MGP	3.97	0.45	34.38	2.92	116.10	5002.33

CUSP, cumulative shrinkage process; MGP, multiplicative gamma process; MSE, mean square error; ESS, effective sample size; IQR, interquartile range.

independent $N(0,1)$. To study performance at varying dimensions, we consider three different combinations of (p, H_0): $(20,5)$, $(50,10)$ and $(100,15)$. For every pair (p, H_0) we sample 25 datasets of $n = 100$ observations from $N_p(0, \Omega_0)$ and, for each of the 25 replicates, we perform posterior inference on Ω via the Gaussian factor model in § 2.3.1 under both prior (2.5) and (2.6), exploiting the adaptive Gibbs sampler in Bhattacharya & Dunson (2011) and Algorithm 2, respectively.

For our cumulative shrinkage process, we set $\alpha = 5$, $a_\theta = b_\theta = 2$ and $\theta_\infty = 0.05$, whereas for the multiplicative gamma process, we follow Durante (2017) by considering $(a_1, a_2) = (1, 2)$, and set $\nu = 3$ as done by Bhattacharya & Dunson (2011) in their simulations. For both models, (a_σ, b_σ) are fixed at $(1, 0.3)$ as in Bhattacharya & Dunson (2011). The truncation H is initialized at p for the multiplicative gamma process and at $p + 1$ for the cumulative shrinkage process, both corresponding to at most p active factors. For the

Table 2.2: Sensitivity analysis for CUSP hyperparameters $(\alpha, a_\theta, b_\theta, \theta_\infty)$ in 25 simulations

(p, H_0)	$(\alpha, a_\theta, b_\theta, \theta_\infty)$	MSE		$E(H^* \mid y)$		averaged ESS	runtime (s)
		median	IQR	median	IQR	median	median
(20,5)	(2.5,2,2,0.05)	0.74	0.32	5.00	0.00	626.22	317.31
	(10,2,2,0.05)	0.74	0.33	5.00	0.00	636.61	314.82
	(5,2,1,0.05)	0.72	0.34	5.00	0.00	607.61	322.68
	(5,1,2,0.05)	0.79	0.30	5.00	0.00	602.28	309.39
	(5,2,2,0.025)	0.78	0.31	5.00	0.00	655.80	313.21
	(5,2,2,0.1)	0.74	0.30	5.00	0.04	604.88	315.51
(50,10)	(2.5,2,2,0.05)	2.25	0.40	10.00	0.00	280.39	719.11
	(10,2,2,0.05)	2.20	0.36	10.00	0.00	277.89	748.75
	(5,2,1,0.05)	2.16	0.42	10.00	0.00	266.82	722.67
	(5,1,2,0.05)	2.35	0.40	10.00	0.00	272.47	689.70
	(5,2,2,0.025)	2.22	0.35	10.00	0.00	280.60	717.19
	(5,2,2,0.1)	2.22	0.41	10.00	0.00	273.39	698.96
(100,15)	(2.5,2,2,0.05)	3.68	0.47	15.00	0.00	176.31	2247.44
	(10,2,2,0.05)	3.74	0.40	15.00	0.00	172.02	2205.78
	(5,2,1,0.05)	3.64	0.44	15.00	0.00	172.04	2287.32
	(5,1,2,0.05)	3.96	0.52	15.00	0.00	174.74	2178.47
	(5,2,2,0.025)	3.70	0.44	15.00	0.00	172.83	2200.20
	(5,2,2,0.1)	3.77	0.44	15.00	0.00	174.76	2284.80

CUSP, cumulative shrinkage process; MSE, mean square error; ESS, effective sample size; IQR, interquartile range.

two methods, adaptation is allowed only after 500 iterations and, following Bhattacharya & Dunson (2011), the parameters (α_0, α_1) are set to $(-1, -5 \times 10^{-4})$, while the adaptation threshold ϵ in the multiplicative gamma process is 10^{-4}. Both algorithms are run for 10000 iterations after a burn-in of 5000 and, by thinning every 5, we obtain a final sample of 2000 draws from the posterior of Ω. For each of the 25 simulations in every scenario, we compute a Monte Carlo estimate of $\sum_{j=1}^{p} \sum_{q=j}^{p} E\{(\Omega_{jq} - \Omega_{0jq})^2 \mid y\}/\{p(p+1)/2\}$ and $E(H^* \mid y)$. Since $E\{(\Omega_{jq} - \Omega_{0jq})^2 \mid y\} = \{E(\Omega_{jq} \mid y) - \Omega_{0jq}\}^2 + \text{var}(\Omega_{jq} \mid y)$, the posterior averaged mean square error accounts both for bias and variance in the posterior of Ω.

Table 2.1 shows, for each scenario and model, the median and the interquartile range of the above quantities computed from the 25 measures produced by the different simulations, together with the medians of the averaged effective sample sizes, out of 2000 samples, and of the running times. Such quantities rely on a R implementation run on an Intel Core i7-3632QM CPU laptop with 7.7 GB of RAM. The two methods have comparable mean square errors, but these measures and the performance gains of prior (2.6) over (2.5) increase with H_0. Our approach also provides some improvements in mixing and reduced running times. The latter is arguably due to the fact that the

multiplicative gamma process overestimates H^*, hence keeping more parameters to update than necessary. Instead, our cumulative shrinkage process recovers the true dimension H_0 in all settings, thus efficiently tuning the truncation level H. Such an improved learning of the true underlying dimension is confirmed by the 95% credible intervals highly concentrated around H_0 in all the scenarios considered. The multiplicative gamma process leads instead to wider credible intervals for H^*, with none of them including H_0. As shown in Table 2.2, results are robust to moderate and reasonable changes in the hyperparameters of the cumulative shrinkage process. We also tried to modify ϵ in Bhattacharya & Dunson (2011) so as to delete columns of Λ with values on the same scale of our spike. This setting provided lower estimates for H^* and, hence, a computational time more similar to our cumulative shrinkage process, but led to higher mean square errors and still some difficulties in learning H_0.

2.5 Application to personality data

We conclude with an application to a subset of the personality data available in the dataset `bfi` from the R package `psych`. Here, we focus on the association structure among $p = 25$ personality self-report items collected on a 6 point response scale for $n = 126$ individuals older than 50 years. These variables represent answers to questions organized into five personality traits known as agreeableness, conscientiousness, extraversion, neuroticism, and openness. Recalling common implementations of factor models, we center the 25 items, and then replace variables $1, 9, 10, 11, 12, 22$ and 25 with their negative version as suggested in the R documentation of the `bfi` dataset to have coherent answers within each personality trait. Posterior inference under priors (2.5)–(2.6) is performed with the same hyperparameters and Gibbs settings as in § 2.4.

Figure 2.1 shows posterior means and credible intervals for the absolute value of the entries in the correlation matrix $\overline{\Omega}$, under our model. Samples from $\overline{\Omega}$ are obtained computing $\overline{\Omega} = (\Omega \odot I_p)^{-\frac{1}{2}} \Omega (\Omega \odot I_p)^{-\frac{1}{2}}$ for every sample of $\Omega = \Lambda\Lambda^T + \Sigma$, with $\odot$ denoting the element-wise Hadamard product. Figure 2.1 highlights associations within each block of five answers measuring a main personality trait, while showing also interesting across-blocks correlations among agreeableness and extraversion as well as conscientiousness and neuroticism. Openness has less evident within-block and across-block associations. These results suggest three main factors as confirmed by the posterior mean and by the 95% credible intervals for H^* under the cumulative shrinkage process, which are 2.84 and $(2, 3)$, respectively. Such posterior summaries are 24.01 and $(18, 25)$ under the multiplicative gamma process, but the higher H^* does not lead to improved learning of $\overline{\Omega}$. In fact, when considering the Monte Carlo estimate

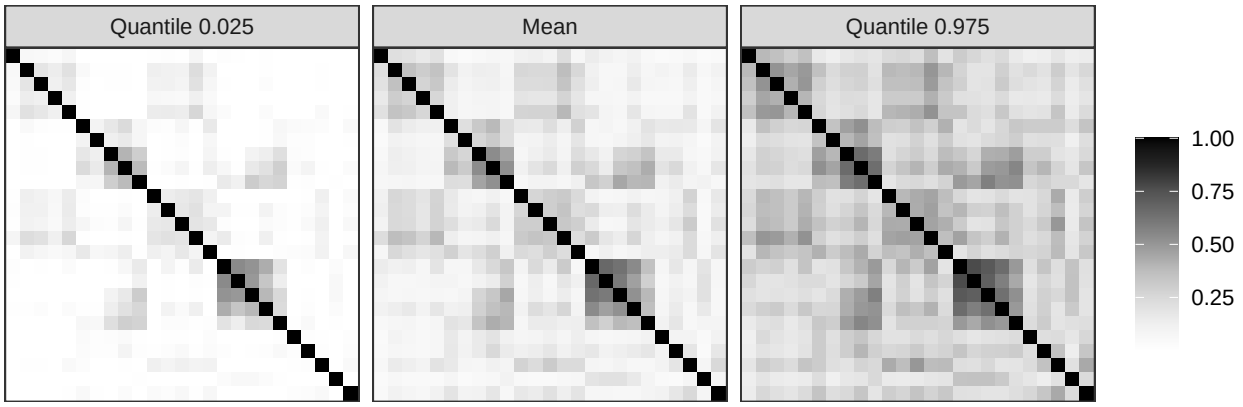

Figure 2.1: Posterior mean and credible intervals for each element of the absolute correlation matrix obtained through our adaptive Gibbs sampler (Algorithm 2)

of the mean squared deviations $\sum_{j=1}^{p}\sum_{q=j}^{p} E(\overline{\Omega}_{jq} - S_{jq})^2/\{p(p+1)/2\}$ from the sample correlation matrix S, we obtain 0.01 under both (2.5) and (2.6), thus suggesting that the multiplicative gamma process might overestimate H^* in this application. This leads to more redundant parameters to be updated in the adaptive Gibbs sampler, thus increasing the computational time from 400.69 to 1321.04 seconds. Our approach also increases the averaged effective sample size from 901.68 to 1070.83.

2.6 Variational methods

The adaptive Gibbs sampler proposed in § 2.3.3 for the cumulative shrinkage process has proved faster than the one for the multiplicative gamma process, on both simulated and real data; see § 2.4–2.5. However, larger datasets may require even faster algorithms. This need for scalability has pushed Bayesian statisticians towards approximate methods for posterior inference, including Laplace approximation, variational Bayes and expectation propagation (Bishop, 2006). Here we consider a variational approach, and in particular mean-field variational inference. This choice is suggested by the fact that Gaussian factor models endowed with a cumulative shrinkage process involve conditionally conjugate distributions within the exponential family, a setting well suited to mean-field variational inference (Blei et al., 2017). The derivation of the variational algorithm is further facilitated by a prior specification which slightly differs from the one in § 2.3.1, and which is described in § 2.6.1.

2.6.1 Simplified prior specification

We focus on the same Gaussian factor model considered in § 2.3.1 and, as we did there, we let $\sigma_j^2 \sim \text{InvGa}(a_\sigma, b_\sigma)$ for $j = 1, \ldots, p$. However, in contrast to § 2.3.1, we place a

cumulative shrinkage process prior directly on the loadings

$$(\lambda_{jh} \mid \pi_h) \sim (1-\pi_h)\mathrm{N}(0,\theta_0) + \pi_h \mathrm{N}(0,\theta_\infty), \qquad (j=1,\dots,p;\ h=1,\dots,H), \tag{2.9}$$

with θ_0 and θ_∞ both set to positive values, and π_h as in Def. 2.1, truncated at H with $v_H = 1$. This prior specification facilitates the derivation of the variational algorithm, since the marginals for the loadings under the slab and the spike are both Gaussian. Nevertheless, as long as θ_0 and θ_∞ are set sensibly, this specification preserves the increasing shrinkage property. In fact, setting $\theta_0 > \theta_\infty$, the loadings are increasingly shrunk towards zero in probability, i.e. $\mathrm{pr}\{|\lambda_{j,h+1}| < \epsilon\} \geqslant \mathrm{pr}\{|\lambda_{jh}| < \epsilon\}$ for any $\epsilon > 0$, thus encoding the prior assumption that additional factors provide a decreasing contribution to the model. In principle, setting both the spike and the slab to Gaussians as in (2.9) might be suboptimal compared to the specification in § 2.3.1, where the Student-t slab is more differentiated from the Gaussian spike. However, we will see in § 2.6.3 that this simplified prior specification and the corresponding variational algorithm provide similar accuracy, while being more than five times faster than the adaptive Gibbs sampler for the original specification in § 2.3.

Another small difference from § 2.3 is the data augmentation scheme. In fact, our variational algorithm employs the augmented data $z_h = (z_{h1},\dots,z_{hH}) \sim \mathrm{Mult}\{1,(\omega_1,\dots,\omega_H)\}$, exploiting the fact that (2.9) can be obtained by marginalizing out z_h from

$$(\lambda_{jh} \mid z_h) \sim \{1-\textstyle\sum_{l=1}^{h} z_{hl}\}\mathrm{N}(0,\theta_0) + \sum_{l=1}^{h} z_{hl}\mathrm{N}(0,\theta_\infty), \quad (j=1,\dots,p;\ h=1,\dots,H).$$

This multinomial data augmentation is slightly different from the categorical one in § 2.3.2, but totally equivalent. This minor modification is aimed at simplifying the derivation of the variational algorithm described in the next section.

2.6.2 Variational inference

Variational Bayes approximates the posterior with the density q^* that is closest to it, in Kullback-Leibler (KL) divergence, within a family $\mathcal{Q}$ of tractable densities; see Blei et al. (2017) for a review. The ideal variational family $\mathcal{Q}$ should combine flexibility, that allows for a good approximation, and tractability. Motivated by the remarks at the beginning of § 2.6, we consider the mean-field variational family, whose elements factorize as follows:

$$q(\lambda,\eta,\sigma,z,v) = q(\lambda)q(\eta)q(\sigma)q(z)q(v). \tag{2.10}$$

The KL divergence between such a q and the intractable posterior cannot be computed or minimized directly. Equivalently, we maximize the evidence lower bound

$$\begin{aligned} \mathrm{ELBO}(q) &= \log p(y) - \mathrm{KL}(q(\lambda, \eta, \sigma, z, \nu) \| p(\lambda, \eta, \sigma, z, \nu \mid y)) \\ &= \mathrm{E}_q[\log p(y, \lambda, \eta, \sigma, z, \nu)] - \mathrm{E}_q[\log q(\lambda, \eta, \sigma, z, \nu)]. \end{aligned} \tag{2.11}$$

The first line of (2.11) highlights that, since the KL divergence is always non-negative, the ELBO lower-bounds the log-evidence, thus justifying its name. It also shows that, since $\log p(y)$ does not depend on q, maximizing the ELBO is equivalent to minimizing the KL divergence with respect to q. However, since this expression involves the intractable posterior, the equivalent formula in the second line of (2.11) is used to actually compute the ELBO. Optimization is solved through coordinate ascent, iteratively maximizing the ELBO with respect to each factor on the right-hand side of (2.10). Following Bishop (2006, Ch. 10), each factor update is derived as follows (we report only the loadings term for illustrative purposes):

$$\log q^*(\lambda) = \mathrm{E}_{\neq\lambda}[\log p(y, \lambda, \eta, \sigma, z, \nu)] + \mathrm{const},$$

where $\mathrm{E}_{\neq\lambda}$ denotes the expectation under q with respect to all variables other than λ. With no parametric assumption on the factors in (2.10), we obtain:

$$\begin{aligned} q^*(\lambda, \dots, \nu) \;=\; & \prod_{j=1}^{p} \mathrm{N}_H(\lambda_{j\cdot}; \mu_j^{(\lambda)}, V_j^{(\lambda)}) \prod_{i=1}^{n} \mathrm{N}_H(\eta_{i\cdot}; \mu_i^{(\eta)}, V^{(\eta)}) \prod_{j=1}^{p} \mathrm{InvGa}(\sigma_j^2; A^{(\sigma)}, B_j^{(\sigma)}) \cdot \\ & \cdot \prod_{h=1}^{H} \mathrm{Mult}(z_h; 1, \kappa_h) \prod_{h=1}^{H-1} \mathrm{Beta}(\nu_h; A_h^{(\nu)}, B_h^{(\nu)}). \end{aligned}$$

Each factor further factorizes into exponential-family distributions, thus facilitating computations. The update equations for the parameters are coupled, meaning that each factor update involves expectations with respect to other factors. We then proceed iteratively, cycling over the steps of Algorithm 3. This routine converges to a local maximum, hence should be run from several initializations (Blei et al., 2017). Convergence of each run can be assessed by monitoring the monotone growth of the ELBO. From the optimal variational parameters we can also compute the variational expectation of the number H^* of factors that are active, in the sense that they are modeled by the slab:

$$\mathrm{E}_{q^*}[H^*] = \sum_{h=1}^{H} \sum_{l=h+1}^{H} \kappa_{hl}.$$

Algorithm 3: One cycle of the variational algorithm for Gaussian factor models

1 **for** *j from 1 to p* **do**
  set $V_j^{(\lambda)} = \{diag(\theta_1^*, \dots, \theta_H^*) + (A^{(\sigma)}/B_j^{(\sigma)})(\mu^{(\eta)^T}\mu^{(\eta)} + nV^{(\eta)})\}^{-1}$,
  where $\theta_h^* = (1 - \sum_{l=1}^{h} \kappa_{hl})\theta_0^{-1} + (\sum_{l=1}^{h} \kappa_{hl})\theta_\infty^{-1}$, and
  $\mu_j^{(\lambda)} = (A^{(\sigma)}/B_j^{(\sigma)})V_j^{(\lambda)}\mu^{(\eta)^T}y_{\cdot j}$;

2 Set $A^{(\sigma)} = a_\sigma + n/2$ and **for** *j from 1 to p* **do**
  $B_j^{(\sigma)} = b_\sigma + \frac{1}{2}\sum_{i=1}^{n}\{y_{ij}^2 - 2y_{ij}\mu_i^{(\eta)^T}\mu_j^{(\lambda)} + \sum_{h=1}^{H}\sum_{k=1}^{H}(\mu_{ih}^{(\eta)}\mu_{ik}^{(\eta)} + V_{hk}^{(\eta)})(\mu_{jh}^{(\lambda)}\mu_{jk}^{(\lambda)} + V_{j;hk}^{(\lambda)})\}$;

3 Set $V^{(\eta)} = (I_H + \mu^{(\lambda)^T}diag(A^{(\sigma)}/B^{(\sigma)})\mu^{(\lambda)} + \sum_{j=1}^{p}(A^{(\sigma)}/B_j^{(\sigma)})V_j^{(\lambda)})^{-1}$;
  for *i from 1 to n* **do**
  set $\mu_i^{(\eta)} = V^{(\eta)}\mu^{(\lambda)^T}diag(A^{(\sigma)}/B^{(\sigma)})y_{i\cdot}$;

4 **for** *h from 1 to H* **do**
  for *l from 1 to h* **do**
  set $\kappa_{hl} \propto \exp\{E(\log\omega_l) - 0.5 \cdot p\log\theta_\infty - 0.5\cdot\theta_\infty^{-1}E[\lambda_{\cdot h}^T\lambda_{\cdot h}]\}$
  for *l from h+1 to H* **do**
  set $\kappa_{hl} \propto \exp\{E(\log\omega_l) - 0.5 \cdot p\log\theta_0 - 0.5\cdot\theta_0^{-1}E[\lambda_{\cdot h}^T\lambda_{\cdot h}]\}$

  where $E[\lambda_{\cdot h}^T\lambda_{\cdot h}] = \sum_{j=1}^{p}(\mu_{jh}^{(\lambda)^2} + V_{j;hh}^{(\lambda)})$ and, with Ψ being the digamma function,
  $E(\log\omega_l) = \mathbb{1}\{l < H\}\{\Psi(A_l^{(\nu)}) - \Psi(A_l^{(\nu)} + B_l^{(\nu)})\} + \mathbb{1}\{l > 1\}\sum_{m=1}^{l-1}\{\Psi(B_m^{(\nu)}) - \Psi(A_m^{(\nu)} + B_m^{(\nu)})\}$;

5 **for** *h from 1 to* $(H-1)$ **do**
  set $A_h^{(\nu)} = 1 + \sum_{l=1}^{H}\kappa_{lh}$ and $B_h^{(\nu)} = \alpha + \sum_{l=1}^{H}\sum_{m=h+1}^{H}\kappa_{lm}$.

2.6.3 Application to personality data

We compare the proposed variational algorithm (Algorithm 3) and the adaptive Gibbs sampler described in § 2.3.3 (Algorithm 2) on the same dataset considered in § 2.5. For Algorithm 2 we adopt the same hyperparameters and Gibbs settings as in § 2.5, while for Algorithm 3 we set $\alpha = 5$, $\theta_0 = 1$, $\theta_\infty = 10^{-6}$ and we let $H = p + 1$, which coincides with the initial value of H in the adaptive Gibbs sampler and corresponds to at most p latent factors. We run the variational algorithm from 20 random initializations, stopping each run when the ELBO grows less than 0.05. We then pick the run reaching the highest ELBO. Using the optimal variational parameters of this run, we get a sample of size 2000 for the covariance matrix Ω, from which we derive a sample for the correlation matrix $\overline{\Omega} = (\Omega \odot I_p)^{-1/2}\Omega(\Omega \odot I_p)^{-1/2}$, with $\odot$ denoting the element-wise Hadamard product. From this sample we compute a Monte Carlo estimate of the mean squared deviations $\sum_{j=1}^{p}\sum_{q=j}^{p}E(\overline{\Omega}_{jq} - S_{jq})^2/\{p(p+1)/2\}$ between $\overline{\Omega}$ and the sample correlation matrix S. The same quantity is computed from a posterior sample of equal size obtained running Algorithm 2 for 10000 iterations after a burn-in of 5000 and then thinning every five.

Table 2.3: Performance of adaptive Gibbs sampler and variational algorithm on the `bfi` dataset

Method	MSE	E[H*]	Running time (s)
Adaptive Gibbs sampler	0.01	2.7	340
Variational algorithm	0.01	3.0	63

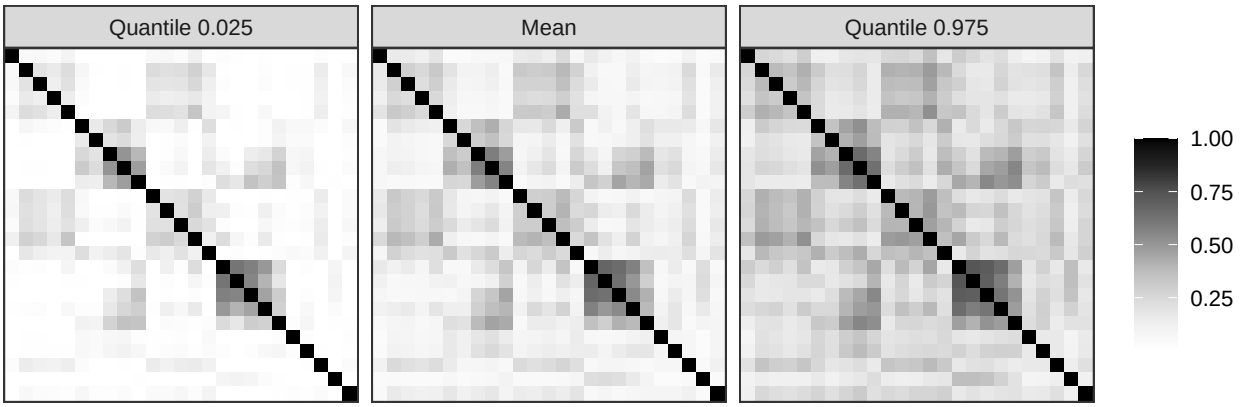

Figure 2.2: Posterior mean and credible intervals for each element of the absolute correlation matrix obtained through our mean-field variational algorithm (Algorithm 3)

The two quantities are reported as MSE (Mean Square Error) in Table 2.3, together with the posterior expectation of the number of active factors and the total running time for each of the two methods. The proposed variational algorithm is more than five times faster than the adaptive Gibbs sampler, while providing a similar expected number of active factors and the same MSE (rounded off to the second decimal digit).

Mean-field variational approximations are typically good at point estimation, while they may underestimate the posterior variability, since they suppress some posterior correlations (Bishop, 2006, Ch. 10). To get a first graphical assessment of whether this is the case with our variational approximation, in Figure 2.2 we plot the posterior means and credible intervals for the absolute value of the entries in the correlation matrix $\overline{\Omega}$, sampled from the optimal variational distribution. A visual comparison among Figure 2.2 and the corresponding plot for the adaptive Gibbs sampler (Figure 2.1), suggests that most of the posterior variability of $\overline{\Omega}$ has been preserved in our variational approximation, though this surely deserves deeper investigation in future work.

Chapter 3

Extended stochastic block models

3.1 Introduction

Network data are ubiquitous in science and there is recurring interest in community structure. Interacting units – such as brain regions (Crossley et al., 2013), social actors (Zhao et al., 2011) and transportation nodes (Guimera et al., 2005) – can often be grouped into clusters which share similar connectivity patterns in the corresponding network. The relevance of such a property and the interdisciplinary nature of network science have motivated a collective effort by various disciplines towards the development of methods for community detection, ranging from algorithmic strategies (Girvan & Newman, 2002; Newman & Girvan, 2004; Newman, 2006; Blondel et al., 2008; Priebe et al., 2019) to model-based solutions (Holland et al., 1983; Nowicki & Snijders, 2001; Kemp et al., 2006; Airoldi et al., 2008; Athreya et al., 2017; Geng et al., 2019); see also Fortunato (2010) and Lee & Wilkinson (2019) for a comprehensive overview. Despite being widely used in practice, most algorithmic approaches lack uncertainty quantification and can only detect communities characterized by dense within-block connectivity and sparser connections between different blocks (Fortunato & Hric, 2016). These issues have motivated a growing interest in model-based solutions which rely on generative statistical models. This choice allows coherent uncertainty quantification, model selection and hypothesis testing, while accounting for more general connectivity patterns, where nodes in the same community are not necessarily more densely connected, but simply share the same connectivity behavior (Fortunato & Hric, 2016), which may even characterize core-periphery, disassortative or weak community patterns (Fortunato & Hric, 2016, Figure 8). These alternative structures are also found in the motivating 2013–2018 Italian bill co-sponsorship network (Briatte, 2016) displayed in Figure 3.1, thus supporting our focus on model-based solutions.

Among the generative models for learning communities in network data, the stochastic block model (SBM) (Holland et al., 1983; Nowicki & Snijders, 2001) is arguably the most

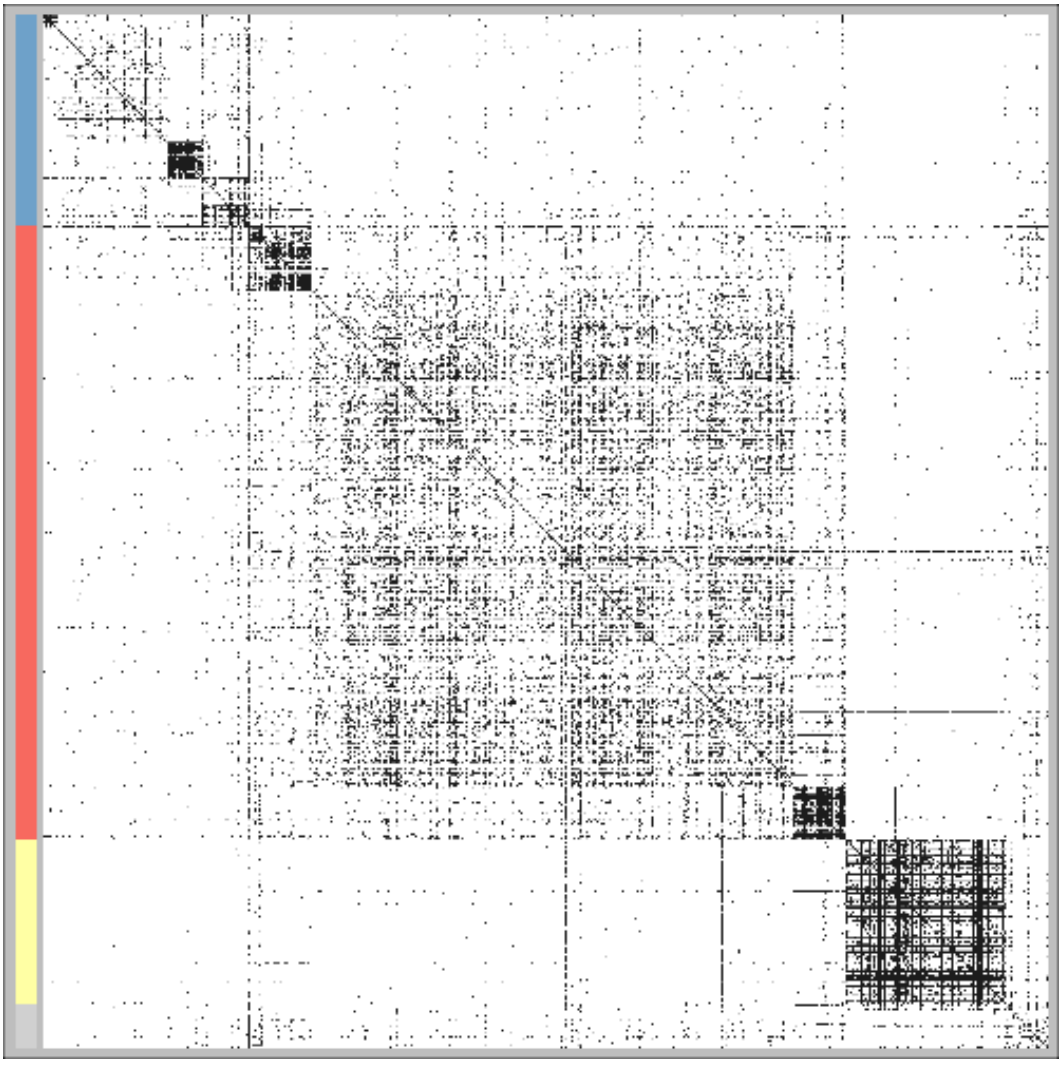

Figure 3.1: Adjacency matrix of the 2013–2018 bill co-sponsorship network in the Italian parliament. Edges and non-edges are depicted as black and white pixels, respectively. Colors on the left denote the right wing (blue), left wing (red), *Movimento 5 Stelle* (yellow) and mixed group (grey).

widely-implemented and well-established one, owing also to its unique balance between simplicity and flexibility (Lee & Wilkinson, 2019). In SBMs, the probability of an edge between two nodes only depends on their cluster memberships, thus allowing efficient inference on communities and on block probabilities, which can characterize assortative, disassortative, core-periphery or weak community patterns, and combinations thereof (Fortunato & Hric, 2016). These desirable properties have motivated extensive theoretical studies (Zhao et al., 2012; Olhede & Wolfe, 2014; van der Pas & van der Vaart, 2018; Ghosh et al., 2019) and various generalizations (Tallberg, 2004; Kemp et al., 2006; Handcock et al., 2007; Airoldi et al., 2008; Karrer & Newman, 2011; Schmidt & Morup, 2013; Newman & Clauset, 2016; Geng et al., 2019; Stanley et al., 2019) of the original SBM.

Most of the above extensions address two fundamental open problems with classical SBMs. First, in real-world applications the number of underlying communities is typically unknown and has to be learned from the data. Therefore, classical SBM formulations based on a fixed and pre-specified number of communities (Holland et al., 1983; Nowicki & Snijders, 2001) are not suitable to address this goal. Second, it is common to observe nodal attributes that may effectively inform the community assignment mechanism. Hence, SBMs require extensions to include such information in the process regulating the node partitions. A successful answer to the first open issue has been provided by Bayesian nonparametric solutions replacing the original Dirichlet-multinomial process for node partitioning (Nowicki & Snijders, 2001) with alternative priors that allow the number of communities to grow adaptively with the size of the network via the Chinese restaurant process (CRP) (Kemp et al., 2006; Schmidt & Morup, 2013) or be finite and random under a mixture-of-finite-mixtures representation (Geng et al., 2019). Inclusion of nodal attributes within the community assignment is instead obtained via multinomial probit (Tallberg, 2004) or mixture models (Newman & Clauset, 2016; Stanley et al., 2019). Unfortunately, all these different extensions have been developed separately and SBMs still lack a unifying framework, which would be useful to clarify common properties, develop broad computational and inferential strategies, and identify novel solutions.

Motivated by the above discussion, we unify the aforementioned formulations within a general extended stochastic block model (ESBM) framework based on Gibbs-type priors (Gnedin & Pitman, 2005; De Blasi et al., 2013a), which also allows the inclusion of node attributes in a principled manner via product partition models (PPMs) (Hartigan, 1990). Within this class, we focus on the Gnedin process (Gnedin, 2010) as an example of a prior which has not been yet employed in the context of SBMs, but exhibits analytical tractability, desirable properties, theoretical guarantees and promising empirical performance when combined with such models. Our framework allows posterior computation via an easy-to-implement collapsed Gibbs sampler, and motivates general methods for

uncertainty quantification and model assessment, thus exploiting the advantages of a model-based approach over algorithmic strategies. The performance of key priors within the ESBM class is evaluated in simulations. In light of these results, we opt for the Gnedin process to analyze the political network in Figure 3.1.

3.2 Model formulation

Consider a binary undirected network with V nodes and let $\mathbf{Y}$ denote its $V \times V$ symmetric adjacency matrix, with elements $y_{vu} = y_{uv} = 1$ if nodes v and u are connected, and $y_{vu} = y_{uv} = 0$ otherwise. Since the focus is on community detection, self-loops are not relevant and, hence, are not included in the generative model. We first review SBMs and then introduce our general ESBM class along with associated properties and extensions to incorporate node attributes. For simplicity, we focus on binary undirected networks and categorical attributes, but our approach can be naturally extended to other types of networks and covariates, as discussed in Chapter 5.

3.2.1 Stochastic block models

SBMs (Holland et al., 1983; Nowicki & Snijders, 2001) partition the nodes into $\overline{H}$ mutually exclusive and exhaustive communities, with nodes in the same community sharing common connectivity patterns. More specifically, SBMs assume that the sub-diagonal entries y_{vu} ($v = 2, \ldots, V$; $u = 1, \ldots, v-1$) of the symmetric adjacency matrix $\mathbf{Y}$ are conditionally independent Bernoulli random variables with probabilities depending only on the community memberships of the involved nodes v and u. Denoting with $\bar{\mathbf{z}} = (\bar{z}_1, \ldots, \bar{z}_V)^\intercal \in \{1, \ldots, \overline{H}\}^V$ the vector of community assignments of the V nodes, and with $\boldsymbol{\Theta}$ the $\overline{H} \times \overline{H}$ symmetric matrix whose generic element θ_{hk} is the probability of a connection between a node in community h and a node in community k, the likelihood for the adjacency matrix $\mathbf{Y}$ is

$$\begin{aligned} p(\mathbf{Y} \mid \bar{\mathbf{z}}, \boldsymbol{\Theta}) &= \prod_{v=2}^{V} \prod_{u=1}^{v-1} \theta_{\bar{z}_v \bar{z}_u}^{y_{vu}} (1 - \theta_{\bar{z}_v \bar{z}_u})^{1-y_{vu}} \\ &= \prod_{h=1}^{\overline{H}} \prod_{k=1}^{h} \theta_{hk}^{m_{hk}} (1 - \theta_{hk})^{\overline{m}_{hk}}, \end{aligned} \tag{3.1}$$

where m_{hk} and $\overline{m}_{hk}$ denote the number of edges and non-edges between communities h and k, respectively. Classical SBMs (Holland et al., 1983; Nowicki & Snijders, 2001) assume independent Beta(a, b) priors for the block probabilities θ_{hk}. Thus the joint density for the diagonal and sub-diagonal elements of the symmetric matrix $\boldsymbol{\Theta}$ is

$$p(\boldsymbol{\Theta}) = \prod_{h=1}^{\overline{H}} \prod_{k=1}^{h} \frac{\theta_{hk}^{a-1} (1 - \theta_{hk})^{b-1}}{B(a, b)}, \tag{3.2}$$

where $B(\cdot,\cdot)$ is the Beta function. Since the overarching goal in SBMs is to provide inference on communities, $\boldsymbol{\Theta}$ is usually treated as a nuisance parameter and marginalized out in (3.1) via beta-binomial conjugacy, obtaining

$$p(\mathbf{Y} \mid \bar{\mathbf{z}}) = \prod_{h=1}^{\overline{H}} \prod_{k=1}^{h} \frac{B(a + m_{hk}, b + \overline{m}_{hk})}{B(a, b)}. \tag{3.3}$$

As we will clarify in the following sections, such a collapsed representation is also useful for computation and inference. The likelihood in equation (3.3) is common to several extensions of SBMs, which instead differ in the choice of the probabilistic mechanism underlying the assignments $\bar{\mathbf{z}}$. A natural choice is a Dirichlet-multinomial distribution on $\bar{\mathbf{z}}$, obtained by marginalizing the vector of community assignment probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{\overline{H}}) \sim \text{Dirichlet}(\boldsymbol{\beta})$ out of the likelihood for $\bar{\mathbf{z}}$, assuming $\text{pr}(\bar{z}_v = h \mid \boldsymbol{\pi}) = \pi_h$, $v = 1, \ldots, V$. If $\overline{H}$ is fixed and finite, this leads to the original SBM of Nowicki & Snijders (2001). However, as already discussed, the number of communities is usually unknown and has to be inferred from the data. A possible solution consists in placing a prior on $\overline{H}$, which leads to the mixture-of-finite-mixtures (MFM) version of the SBM by Geng et al. (2019). Another option is a Dirichlet process partition mechanism, corresponding to the infinite relational model (Kemp et al., 2006). Such an infinite mixture model differs from the MFM in that $\overline{H} = \infty$, meaning that infinitely many nodes would give rise to infinitely many communities. The total number of possible clusters $\overline{H}$ should not be confused with the number of occupied clusters H, defined as the number of distinct labels in $\bar{\mathbf{z}}$. H is upper bounded by $\min\{V, \overline{H}\}$, hence it cannot exceed V, even when $\overline{H} = \infty$.

So far we have introduced *labeled* clusters, identified by $\bar{\mathbf{z}}$. This means that a vector $\bar{\mathbf{z}}$ and its relabelings are regarded as distinct objects, even though they identify the same partition. Throughout the rest of the chapter we will rely on a generic $\mathbf{z}$ to denote all relabelings of $\bar{\mathbf{z}}$ that lead to the same partition. For convenience, one may assume that $z_v \in \{1, \ldots, H\}$, which corresponds to avoiding empty communities. Note that (3.3) is invariant under relabeling and, hence, $p(\mathbf{Y} \mid \mathbf{z}) = p(\mathbf{Y} \mid \bar{\mathbf{z}})$.

3.2.2 Extended stochastic block model

As illustrated in the previous section, several priors for community memberships have been considered in the context of SBMs, including the Dirichlet-multinomial (Nowicki & Snijders, 2001), the Dirichlet process (Kemp et al., 2006), and mixtures of finite Dirichlet mixtures (Geng et al., 2019). These are all Gibbs-type priors, which were introduced by Gnedin & Pitman (2005) and stand out for analytical and computational tractability (De Blasi et al., 2013a). In this section we propose the ESBM as a unifying framework characterized by the choice of a Gibbs-type prior for the assignments. This formulation

includes the previously-mentioned SBMs as special cases and offers new alternatives by exploring the whole Gibbs-type class and its connections with PPMs (Hartigan, 1990).

Gibbs-type priors are defined over the space of the unlabeled community indicators $\mathbf{z}$. For $a > 0$, denote the ascending factorial with $(a)_n = a(a+1)\cdots(a+n-1)$ for any $n \geqslant 1$, and set $(a)_0 = 1$. A probability mass function $p(\mathbf{z})$ is of Gibbs-type if and only if

$$p(\mathbf{z}) = \mathcal{W}_{V,H} \prod_{h=1}^{H} (1-\sigma)_{n_h-1}, \tag{3.4}$$

where n_h denotes the number of nodes in cluster h, $\sigma < 1$ is the so-called *discount parameter* and $\{\mathcal{W}_{V,H} : 1 \leqslant H \leqslant V\}$ is a collection of non-negative weights satisfying the recursion $\mathcal{W}_{V,H} = (V - H\sigma)\mathcal{W}_{V+1,H} + \mathcal{W}_{V+1,H+1}$, with $\mathcal{W}_{1,1} = 1$. Gibbs-type priors are a special case of PPMs (Hartigan, 1990; Quintana & Iglesias, 2003), which are probability models for random partitions $\mathbf{z}$ of the form $p(\mathbf{z}) \propto c(Z_1)\cdots c(Z_H)$, where $\{Z_1, \ldots, Z_H\}$ is the partition associated to $\mathbf{z}$, so that $v \in Z_h$ if and only if $z_v = h$, whereas $c(\cdot)$ is a non-negative *cohesion function* measuring the homogeneity within each cluster. Such a connection will be useful to incorporate node-specific attributes effects in ESBMs. Interestingly, Gibbs-type priors represent the largest class of PPMs which are also species sampling models (Pitman, 1996), meaning that the membership indicators $\mathbf{z}$ can be obtained in a sequential and interpretable manner. Specifically, a Gibbs-type random partition $\mathbf{z}$ can be sequentially generated according to

$$\text{pr}(z_{V+1} = h \mid \mathbf{z}) \propto \begin{cases} \mathcal{W}_{V+1,H}(n_h - \sigma) & \text{for } h = 1, \ldots, H, \\ \mathcal{W}_{V+1,H+1} & \text{for } h = H+1. \end{cases} \tag{3.5}$$

Hence, the community assignment process can be easily interpreted as a seating mechanism in which a new node is assigned to an existing community h with probability proportional to the current size n_h of that community discounted by a global factor σ and further rescaled by a weight $\mathcal{W}_{V+1,H}$, which may depend both on the size of the network and on the current number of non-empty communities. Alternatively, the incoming node is assigned to a new community with probability proportional to $\mathcal{W}_{V+1,H+1}$. According to (3.5), when $\sigma > 0$ the mass assigned to existing communities is less than proportional to their cardinality, particularly affecting small clusters, and the remaining mass is added to the probability of creating a new community. This gives an intuition for why the number of occupied clusters grows with V as $\mathcal{O}(V^\sigma)$ when $\sigma > 0$. When $\sigma = 0$ the growth is slower, namely $\mathcal{O}(\log V)$, while $\sigma < 0$ yields a finite $\overline{H}$ even for infinitely many nodes. This is due to the fact that the reinforcement mechanism is reversed and each new community decreases the probability of creating future ones (De Blasi et al., 2013a).

In the examples below we show how commonly used partition processes in SBMs and unexplored alternatives can be obtained as special cases of (3.5).

Example 3.1 (Dirichlet-multinomial – DM). Let $\sigma < 0$ and define $\mathcal{W}_{V,H} = \beta^{H-1}/(\beta\overline{H}+1)_{V-1}\prod_{h=1}^{H-1}(\overline{H}-h)\mathbb{1}(H \leqslant \overline{H})$ for some $\beta = -\sigma$ and $\overline{H} \in \{1, 2, \dots\}$. Then (3.5) coincides with the Dirichlet-multinomial urn-scheme

$$\text{pr}(z_{V+1} = h \mid \mathbf{z}) \propto \begin{cases} n_h + \beta & \text{for } h = 1, \dots, H, \\ \beta(\overline{H} - H)\mathbb{1}(H \leqslant \overline{H}) & \text{for } h = H+1. \end{cases}$$

Example 3.2 (Dirichlet process – DP). Let $\sigma = 0$ and $\mathcal{W}_{V,H} = \alpha^H/(\alpha)_V$ for some $\alpha > 0$. Then (3.5) leads to a CRP scheme

$$\text{pr}(z_{V+1} = h \mid \mathbf{z}) \propto \begin{cases} n_h & \text{for } h = 1, \dots, H, \\ \alpha & \text{for } h = H+1. \end{cases}$$

The CRP can also be obtained as a limiting case of a DM with $\beta = \alpha/\overline{H}$, as $\overline{H} \to \infty$.

Example 3.3 (Pitman-Yor process – PY). Let $\sigma \in [0, 1)$ and set $\mathcal{W}_{V,H} = \prod_{h=1}^{H-1}(\alpha + h\sigma)/(\alpha+1)_{V-1}$ for some $\alpha > -\sigma$. Then (3.5) characterizes the Pitman-Yor process

$$\text{pr}(z_{V+1} = h \mid \mathbf{z}) \propto \begin{cases} n_h - \sigma & \text{for } h = 1, \dots, H, \\ \alpha + H\sigma & \text{for } h = H+1. \end{cases}$$

This scheme clearly reduces to the CRP when $\sigma = 0$.

Example 3.4 (Gnedin process – GN). Let $\sigma = -1$ and $\mathcal{W}_{V,H} = (\gamma)_{V-H}\prod_{h=1}^{H-1}(h^2 - \gamma h)/\prod_{v=1}^{V-1}(v^2 + \gamma v)$ for some $\gamma \in (0, 1)$. Then (3.5) identifies the Gnedin process

$$\text{pr}(z_{V+1} = h \mid \mathbf{z}) \propto \begin{cases} (n_h + 1)(V - H + \gamma) & \text{for } h = 1, \dots, H, \\ H^2 - H\gamma & \text{for } h = H+1. \end{cases}$$

Other known and popular examples of tractable Gibbs-type priors can be found in Lijoi et al. (2007), De Blasi et al. (2013a,b) and Miller & Harrison (2018).

3.2.3 Learning the number of communities

A key focus in community detection is inferring the number of occupied clusters H. As the number of nodes V grows, H converges to $\overline{H}$, which can be assumed, depending on the application, to be finite (scenario I), random but almost surely finite (scenario II), or infinite (scenario III). Classical SBMs (Nowicki & Snijders, 2001) fall into scenario I, the MFM approach of Geng et al. (2019) into scenario II, and the infinite relational model of

	$\overline{H}$	σ	H (growth)	Example
I	Fixed	$\sigma < 0$	–	Dirichlet-multinomial (DM)
II	Random	$\sigma < 0$	–	Gnedin process (GN)
III.a	Infinite	$\sigma = 0$	$\mathcal{O}(\log V)$	Dirichlet process (DP)
III.b	Infinite	$\sigma \in (0,1)$	$\mathcal{O}(V^{\sigma})$	Pitman-Yor process (PY)

Table 3.1: A classification of Gibbs-type priors.

Kemp et al. (2006) into scenario III. As shown in Table 3.1, Gibbs-type priors cover all these three scenarios, allowing analysts to choose the most suitable for a given study.

The only Gibbs-type prior in scenario I is the Dirichlet-multinomial, which serves as building block for Gibbs-type priors in scenario II. In fact, the latter can be derived from the Dirichlet-multinomial by placing a prior on $\overline{H}$, thus making it random. For instance, the distribution of $\mathbf{z}$ under the Gnedin process in Example 3.4 can be expressed as

$$p_G(\mathbf{z};\gamma) = \sum\nolimits_{h=1}^{\infty} \mathrm{pr}_G(\overline{H}=h)\, p_{DM}(\mathbf{z};1,h),$$

where $p_{DM}(\mathbf{z};\beta,\overline{H})$ denotes the Dirichlet-multinomial distribution in Example 3.1, and $\mathrm{pr}_G(\overline{H}=h) = \gamma(1-\gamma)_{h-1}/h!$ can be interpreted as a prior distribution on $\overline{H}$. Although different prior choices for $\overline{H}$ might be considered (Miller & Harrison, 2018), the Gnedin process has considerable advantages. First, the sequential mechanism in Example 3.4 has a simple analytical expression. Moreover, the distribution $\mathrm{pr}_G(\overline{H}=h) = \gamma(1-\gamma)_{h-1}/h!$ has the mode at one, heavy tail and infinite expectation (Gnedin, 2010). Hence, the associated MFM favors simpler models with fewer communities while being also a robust specification for $\overline{H}$, because considerable mass is allocated on the tail of the distribution.

Priors on $\overline{H}$ quantify uncertainty in the total number of communities that one would expect if $V \to \infty$. In practice, the number of non-empty communities H occupied by the observed V nodes is of more direct interest. Under Gibbs-type priors such a quantity has a closed form probability mass function (Gnedin & Pitman, 2005) that coincides with

$$\mathrm{pr}(H=h) = (\mathcal{W}_{V,h}/\sigma^h)\, \mathscr{C}(V,h;\sigma), \qquad h = 1,\dots,V, \tag{3.6}$$

where $\mathscr{C}(V,h;\sigma) = 1/h! \sum_{j=0}^{h}(-1)^j h!\{j!(h-j)!\}^{-1}(-j\sigma)_V$ is the generalized factorial coefficient. The CRP is recovered when $\sigma \to 0$.

In addition to its practical relevance, (3.6) clarifies the asymptotic behavior of H. Indeed, the distribution of H converges to a point mass in scenario I, to a proper distribution in scenario II and to a point mass at infinity in scenario III. For instance, for

the Gnedin process in Example 3.4, we have that (3.6) reduces to

$$\mathrm{pr}_G(H = h) = \binom{V}{h} \frac{(1-\gamma)_{h-1}(\gamma)_{V-h}}{(1+\gamma)_{V-1}}, \qquad h = 1, \dots, V,$$

and hence the expected value can be easily computed as

$$\mathbb{E}_G(H) = \sum_{h=1}^{V} h \cdot \binom{V}{h} \frac{(1-\gamma)_{h-1}(\gamma)_{V-h}}{(1+\gamma)_{V-1}}.$$

Note that $\lim_{V\to\infty} \mathrm{pr}_G(H = h) = \mathrm{pr}_G(\overline{H} = h) = \gamma(1-\gamma)_{h-1}/h!$.

Dirichlet and Pitman-Yor processes may lead to inconsistent estimates for the number of communities if the data are generated from a model with $\overline{H}_0 < \infty$ (Miller & Harrison, 2014). Intuitively, priors in scenario III fail in estimating a finite $\overline{H}_0$ because, by assumption, $\overline{H} = \infty$. Hence, we suggest Gibbs-type priors with $\sigma \geqslant 0$ only if the analyst believes that $\overline{H}_0 = \infty$, that is, when the true number of communities is assumed to grow without bound with the number of nodes V. If the analyst believes that $\overline{H}_0 < \infty$, then Gibbs-type priors of scenario II may be more suitable. In the context of SBMs, Geng et al. (2019) proved a consistency result for a MFM, that actually applies to any Dirichlet-multinomial with a prior on $\overline{H}$ supported on all positive integers. For instance, consistency holds for the Gnedin process in Example 3.4.

3.2.4 Inclusion of node attributes

When node-specific attributes $\mathbf{x}_v = (x_{v1}, \dots, x_{vd})^\intercal$ are available for $v = 1, \dots, V$, such information may support inference on community structures, both in term of point estimation and in reduction of posterior uncertainty. An option to include attributes within ESBMs in a principled manner is to rely on the PPM structure of Gibbs-type priors. Adapting results in Park & Dunson (2010) and Müller et al. (2011) to our network setting, this solution is based on the idea of replacing (3.4) with

$$p(\mathbf{z} \mid \mathbf{x}) \propto \mathcal{W}_{V,H} \prod_{h=1}^{H} p(\mathbf{x}_h)(1-\sigma)_{n_h-1}, \tag{3.7}$$

where $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_V)^\intercal$, whereas $\mathbf{x}_h = (\mathbf{x}_v : z_v = h)$ are the attributes for the nodes in cluster h. In (3.7), $p(\mathbf{x}_h)$ controls the contribution of the attributes to the cluster cohesion and, as we will clarify later, it favors communities that are homogeneous with respect to attribute values. Even if attributes are not considered random, in this context Müller et al. (2011) suggest choosing $p(\mathbf{x}_h)$ as the probability distribution induced by an auxiliary model $p(\mathbf{x}_h \mid \xi_h)$, with ξ_h denoting community-specific parameters, thus obtaining

$p(\mathbf{x}_h) = \int p(\mathbf{x}_h \mid \xi_h) dp(\xi_h)$. We refer to Müller et al. (2011) for further discussion about the choice of $p(\cdot)$.

Here we consider the case in which each node attribute $\mathbf{x}_v = x_v$ is a single categorical variable taking values in $\{1, \ldots, C\}$. This is a common setting in applications, where node attributes often come in the form of exogenous partitions. For example, in the Italian bill co-sponsorship network in Figure 3.1, possible attributes are party or coalition memberships, that we expect to influence voting behaviors. Following Müller et al. (2011), we consider a Dirichlet-multinomial auxiliary model for such attributes, which leads to the following cohesion function

$$p(\mathbf{x}_h) \propto \frac{1}{\Gamma(n_h + \alpha_0)} \prod_{c=1}^{C} \Gamma(n_{hc} + \alpha_c), \tag{3.8}$$

where n_{hc} is the number of nodes in cluster h with attribute value c, and $\alpha_0 = \sum_{c=1}^{C} \alpha_c$, with $\alpha_c > 0$ for $c = 1, \ldots, C$.

3.3 Posterior computation and inference

We derive a collapsed Gibbs sampler that holds for any model within the ESBM class, and allows inclusion of node attributes. Then, we provide extensive tools not only for point estimation of the community structure, but also for uncertainty quantification and model selection. Despite their importance, these two aspects have been largely neglected in the SBM literature.

3.3.1 Collapsed Gibbs sampler

The availability of the urn scheme in (3.5) for the whole class of Gibbs-type priors allows us to derive a collapsed Gibbs sampler that holds for any ESBM (see Algorithm 4). At each iteration, we sample the community assignment of each node v from its full-conditional distribution given the adjacency matrix $\mathbf{Y}$ and the vector $\mathbf{z}_{-v}$ of the cluster assignments of all the other nodes. By simple application of the Bayes rule, these full conditional probabilities are equal to

$$\text{pr}(z_v = h \mid \mathbf{Y}, \mathbf{z}_{-v}) = \text{pr}(z_v = h \mid \mathbf{z}_{-v}) \frac{p(\mathbf{Y} \mid z_v = h, \mathbf{z}_{-v})}{p(\mathbf{Y} \mid \mathbf{z}_{-v})}. \tag{3.9}$$

Recalling (Schmidt & Morup, 2013), the last term in (3.9) simplifies to

$$\prod_{k=1}^{H} \frac{B(a + m_{hk}^{-} + r_{vk}, b + \overline{m}_{hk}^{-} + n_{k}^{-} - r_{vk})}{B(a + m_{hk}^{-}, b + \overline{m}_{hk}^{-})}, \tag{3.10}$$

where m^-_{hk} and $\overline{m}^-_{hk}$ are the number of edges and non-edges between clusters h and k, excluding node v, and r_{vk} is the number of edges between node v and the nodes in cluster k. The prior term $\text{pr}(z_v = h \mid \mathbf{z}_{-v})$ in (3.9) is derived from (3.5) and coincides with

$$\text{pr}(z_v = h \mid \mathbf{z}_{-v}) \propto \begin{cases} \mathcal{W}_{V,H^-}(n^-_h - \sigma) & \text{for } h \leqslant H^-, \\ \mathcal{W}_{V,H^-+1} & \text{for } h = H^- + 1, \end{cases} \tag{3.11}$$

where n^-_h and H^- are the cardinality of cluster h and the number of occupied communities, respectively, after removing node v from $\mathbf{Y}$. Under the priors in Table 3.1, (3.11) admits the simple closed-form expressions reported in Examples 3.1–3.4.

When available, nodal attributes can be incorporated via (3.7), leading to an attribute-dependent collapsed Gibbs sampler. In this case, the full conditionals in (3.9) become

$$\text{pr}(z_v = h \mid \mathbf{Y}, \mathbf{x}, \mathbf{z}_{-v}) \propto \text{pr}(z_v = h \mid \mathbf{Y}, \mathbf{z}_{-v}) \frac{p(\mathbf{x}_h)}{p(\mathbf{x}_{h,-v})}, \tag{3.12}$$

where $\mathbf{x}_h$ and $\mathbf{x}_{h,-v}$ are the attributes for the nodes in the hth community, including and excluding node v, respectively. In the case of categorical attributes with $p(\mathbf{x}_h)$ as in (3.8), the last term in (3.12) can be written as

$$\frac{p(\mathbf{x}_h)}{p(\mathbf{x}_{h,-v})} = \frac{n^-_{hx_v} + \alpha_{x_v}}{n^-_h + \alpha_0}, \tag{3.13}$$

where n^-_{hc} is the number of nodes in cluster h with covariate value c and n^-_h is the total number of nodes in cluster h, both without counting node v. The introduction of this additional factor favors the attribution of node v to the cluster(s) containing a higher fraction of nodes with its same covariate value x_v. In fact, (3.13) tends to the fraction of nodes in cluster h that have the same attribute value as node v. Instead, for $h = H^- + 1$ the additional term is equal to α_{x_v}/α_0.

Finally, although Algorithm 4 leverages the marginal likelihood in (3.3) with block probabilities integrated out, estimates for each θ_{hk} can be easily obtained. In particular, since $(\theta_{hk} \mid \mathbf{Y}, \mathbf{z}) \sim \text{Beta}(a + m_{hk}, b + \overline{m}_{hk})$, we estimate θ_{hk} by

$$\widehat{\theta}_{hk} = \mathbb{E}[\theta_{hk} \mid \mathbf{Y}, \mathbf{z} = \widehat{\mathbf{z}}] = \frac{a + \widehat{m}_{hk}}{a + \widehat{m}_{hk} + b + \widehat{\overline{m}}_{hk}}, \tag{3.14}$$

where $\widehat{m}_{hk}$ and $\widehat{\overline{m}}_{hk}$ denote the number of edges and non-edges between nodes in communities h and k computed from the estimated cluster assignment $\widehat{\mathbf{z}}$. In the next subsection, we describe methods for estimation of $\mathbf{z}$, uncertainty quantification in community detection, and model selection.

Algorithm 4: Gibbs sampler for ESBM

At each iteration, update the cluster assignments as follows:
For $v = 1, \dots, V$ **do**:
1. Remove node v from the network;
2. If the cluster which contained node v contains no other node, discard it (so that clusters $1, \dots, H^-$ are non-empty);
3. Sample z_v from a categorical distribution with probabilities

$$\text{pr}(z_v = h \mid \mathbf{Y}, \mathbf{z}_{-v}) = \text{pr}(z_v = h \mid \mathbf{z}_{-v}) \cdot \frac{p(\mathbf{Y} \mid z_v = h, \mathbf{z}_{-v})}{p(\mathbf{Y} \mid \mathbf{z}_{-v})},$$

for $h = 1, \dots, H^- + 1$, with $\text{pr}(z_v = h \mid \mathbf{z}_{-v})$ as in (3.11) and $p(\mathbf{Y} \mid z_v = h, \mathbf{z}_{-v})/p(\mathbf{Y} \mid \mathbf{z}_{-v})$ as in (3.10). If categorical node attributes are available and have to be incorporated via (3.7)–(3.8), rescale the above expression by (3.13).

3.3.2 Estimation, uncertainty quantification, and model selection

While algorithmic methods return a single estimated partition, ESBMs provide the whole posterior distribution over the space of partitions. To fully exploit such a posterior, we adapt the decision-theoretic approach of Wade & Ghahramani (2018) to the community detection setting. In this way, we summarize posterior distributions on partitions leveraging the *variation of information* (VI) metric (Meilă, 2007), which quantifies distances between two clusterings by comparing their individual and joint entropies, and ranges from zero to $\log_2 V$. Intuitively, VI measures the amount of information in two clusterings relative to the information shared between them, thus providing a metric that decreases to zero as the overlap between two partitions grows; see Wade & Ghahramani (2018) for a discussion of the key properties of VI that facilitate uncertainty quantification on partitions. Under this framework, a formal Bayesian point estimate for $\mathbf{z}$ is the partition with the lowest posterior averaged VI distance from the other clusterings, thus obtaining

$$\hat{\mathbf{z}} = \arg\min_{\mathbf{z}'} \mathbb{E}_{\mathbf{z}}[\text{VI}(\mathbf{z}, \mathbf{z}') \mid \mathbf{Y}], \tag{3.15}$$

where the expectation is taken with respect to $\mathbf{z}$. Due to the huge cardinality of the partition space, even for moderate V, the optimization in (3.15) is typically carried out through a greedy algorithm (Wade & Ghahramani, 2018) as in the R package `mcclust.ext`.

The VI distance also provides natural strategies to construct credible sets around point estimates. In particular, one can define a $1 - \alpha$ credible ball around $\hat{\mathbf{z}}$ by ordering the partitions according to their VI distance from $\hat{\mathbf{z}}$, and defining the ball as containing all the partitions having less than a threshold distance from $\hat{\mathbf{z}}$, with the threshold chosen to minimize the size of the ball while ensuring it contains at least $1 - \alpha$ posterior probability.

Summarizing this ball is non-trivial given the high-dimensional discrete nature of the space of partitions. In practice, as we illustrate in our examples below, one can report the partition at the edge of the ball, which we call credible bound. This form of uncertainty quantification complements the posterior *co-clustering matrix* which presents, for every pair of nodes, the relative frequency of MCMC samples in which such nodes are assigned to the same community.

Another advantage of a Bayesian approach over algorithmic techniques is the possibility of model comparison. In particular, we can compare two models $\mathcal{M}$ and $\mathcal{M}'$ by studying the Bayes factor (Kass & Raftery, 1995)

$$\mathcal{B}_{\mathcal{M},\mathcal{M}'} = \frac{p(\mathbf{Y} \mid \mathcal{M})}{p(\mathbf{Y} \mid \mathcal{M}')} = \frac{\sum_{\mathbf{z}} p(\mathbf{Y} \mid \mathbf{z}) p(\mathbf{z} \mid \mathcal{M})}{\sum_{\mathbf{z}} p(\mathbf{Y} \mid \mathbf{z}) p(\mathbf{z} \mid \mathcal{M}')}. \tag{3.16}$$

Due to the unified structure of ESBMs, this approach is highly general and allows comparisons between any two models in the ESBM class, covering – for example – representations relying on different priors for $\mathbf{z}$ and including or not node attributes. While for degenerate models, with $p(\mathbf{z} \mid \mathcal{M}) = \delta_{\mathbf{z}'}$, computing $p(\mathbf{Y} \mid \mathcal{M})$ reduces to evaluating (3.3) at a specific $\mathbf{z}'$ (see Chapter 4), for non-degenerate models we must rely on posterior samples $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(T)}$ from $p(\mathbf{z} \mid \mathcal{M})$ to obtain an estimate of $p(\mathbf{Y} \mid \mathcal{M})$, for example through the harmonic mean (Newton & Raftery, 1994; Raftery et al., 2007)

$$\widehat{p}(\mathbf{Y} \mid \mathcal{M}) = [T^{-1} \sum_{t=1}^{T} p(\mathbf{Y} \mid \mathbf{z}^{(t)})^{-1}]^{-1}, \tag{3.17}$$

where $p(\mathbf{Y} \mid \mathbf{z}^{(t)})$ is the marginal likelihood in (3.3) at $\mathbf{z}^{(t)}$. We shall note that (3.17) may face instabilities and slow convergence to $p(\mathbf{Y} \mid \mathcal{M})$, thus motivating other estimators (Gelman & Meng, 1998). Such issues did not occur in our empirical studies and the results were always coherent with other model assessment measures. Hence, we maintain (3.17) for its simplicity. As a global measure of goodness-of-fit we also study the misclassification error when predicting each y_{vu} with $\hat{\theta}_{\hat{z}_v \hat{z}_u}$.

3.4 Simulation studies

To assess ESBM performance and highlight benefits over algorithmic strategies (Blondel et al., 2008), we consider two simulated networks of $V = 100$ nodes with various types of block structures, both sampled from a SBM with $\overline{H}_0 = 5$ communities and block probabilities either 0.7 or 0.3. As illustrated in Figure 3.2, the first network has equally-sized groups of 20 nodes each, displaying either community or core-periphery patterns, whereas the second has a cluster of size 40, one of size 30 and the remaining three of size 10, all characterized by classical community structures. State-of-the-art

algorithmic strategies (Blondel et al., 2008) applied to these two networks failed in recovering the true underlying blocks and showed a tendency to over-collapse different communities, due to their inability to incorporate unbalanced noisy partitions and behaviors beyond community patterns.

This motivates implementation of ESBMs, both without and with attributes coinciding with the true partition $\mathbf{z}_0$. Within the Gibbs-type class, we test the four representative priors (DM, DP, PY and GN) for $\mathbf{z}$ presented in Table 3.1. Their hyperparameters are set so that the prior mean for the number H of non-empty clusters is close to $10 > \overline{H}_0$ under all priors. In this way we can check whether our results are robust to hyperparameters settings. Specifically, we set $\alpha = 2.55$ for the DP, $\sigma = 0.575$ and $\alpha = -0.325$ for the PY, $\overline{H} = 50$ and $\beta = 3/50$ for the DM and $\gamma = 0.475$ for the GN. In implementing these models we consider the default setting $a = b = 1$ for the prior on the block probabilities and let $\alpha_1 = \cdots = \alpha_C = 1$ when including node attributes. From Algorithm 4 we obtain 15000 samples for $\mathbf{z}$, after a conservative burn-in of 5000. In our experiments, inference was robust with respect to the initialization of $\mathbf{z}$, but starting with one community for each node provided the best mixing when monitored on the chain of the likelihood in (3.3) evaluated at the MCMC samples of $\mathbf{z}$. The traceplots for such a chain suggested rapid convergence under all models, and Algorithm 4 provided 120 samples of $\mathbf{z}$ per second when implemented on an iMac with 1 Intel Core i5 3.4 GHz processor and 8 GB RAM, thus showing good efficiency. Table 3.2 summarizes the performance of the four priors, both with and without node attributes, in each of the two scenarios.

Among the four Gibbs-type priors considered for $\mathbf{z}$, the Gnedin process always performed slightly better in terms of marginal likelihood and posterior mean of the VI distance from the true partition $\mathbf{z}_0$. More notably, it typically offered more accurate learning of the number of communities, with tighter interquartile ranges that always included the true $\overline{H}_0 = 5$, and tighter credible balls around the VI-optimal posterior point estimate $\hat{\mathbf{z}}$. The posterior bias in terms of VI distance between $\hat{\mathbf{z}}$ and true $\mathbf{z}_0$ is comparable under all four considered priors and much smaller than the maximum achievable VI among two partitions of 100 nodes, which is $\log_2 100 \approx 6.644$. In our trials, the Gnedin process also tested the most robust to hyperparameters.

As expected, including informative attributes improved performance, lowering by one order of magnitude the $\mathbb{E}[VI(\mathbf{z}, \mathbf{z}_0) \mid \mathbf{Y}]$, bringing $VI(\hat{\mathbf{z}}, \mathbf{z}_0)$ to zero and shrinking the credible balls. In a sense, this is the best scenario, since we used the true memberships $\mathbf{z}_0$ as attributes. We also tried supervising with a random permutation of $\mathbf{z}_0$. This resulted in a slight performance deterioration relative to the model without attributes, which is doubly reassuring. In fact, on one hand it shows that the unsupervised model would be preferred to one with non-informative attributes under the proposed model-selection

criteria. On the other, the fact that performance deterioration is not dramatic suggests robustness in learning. According to Figure 3.2, unbalanced partitions are harder to infer, especially without attributes. However this gap vanishes when including informative attributes that can successfully support inference.

As a further fitting measure, we computed the error rate when predicting $\hat{y}_{vu} = 1$ if the estimated block probability $\hat{\theta}_{\hat{z}_v \hat{z}_u}$ is larger than 0.5 and $\hat{y}_{vu} = 0$ otherwise. Under all the considered ESBM specifications, such a rate is 0.29, close to the one expected when predicting based on the true $\theta_{z_v z_u}$. In fact, since in our simulations the true $\theta_{z_v z_u}$ is either 0.7 or 0.3, we have:

$$\begin{aligned} \text{pr}\{y_{vu} = 1 \mid \hat{y}_{vu} = 0\} &= \text{pr}\{y_{vu} = 1 \mid \theta_{vu} = 0.3\} = 0.3; \\ \text{pr}\{y_{vu} = 0 \mid \hat{y}_{vu} = 1\} &= \text{pr}\{y_{vu} = 0 \mid \theta_{vu} = 0.7\} = 0.3. \end{aligned}$$

This suggests accurate calibration and a tendency to avoid overfitting in the considered ESBM specifications.

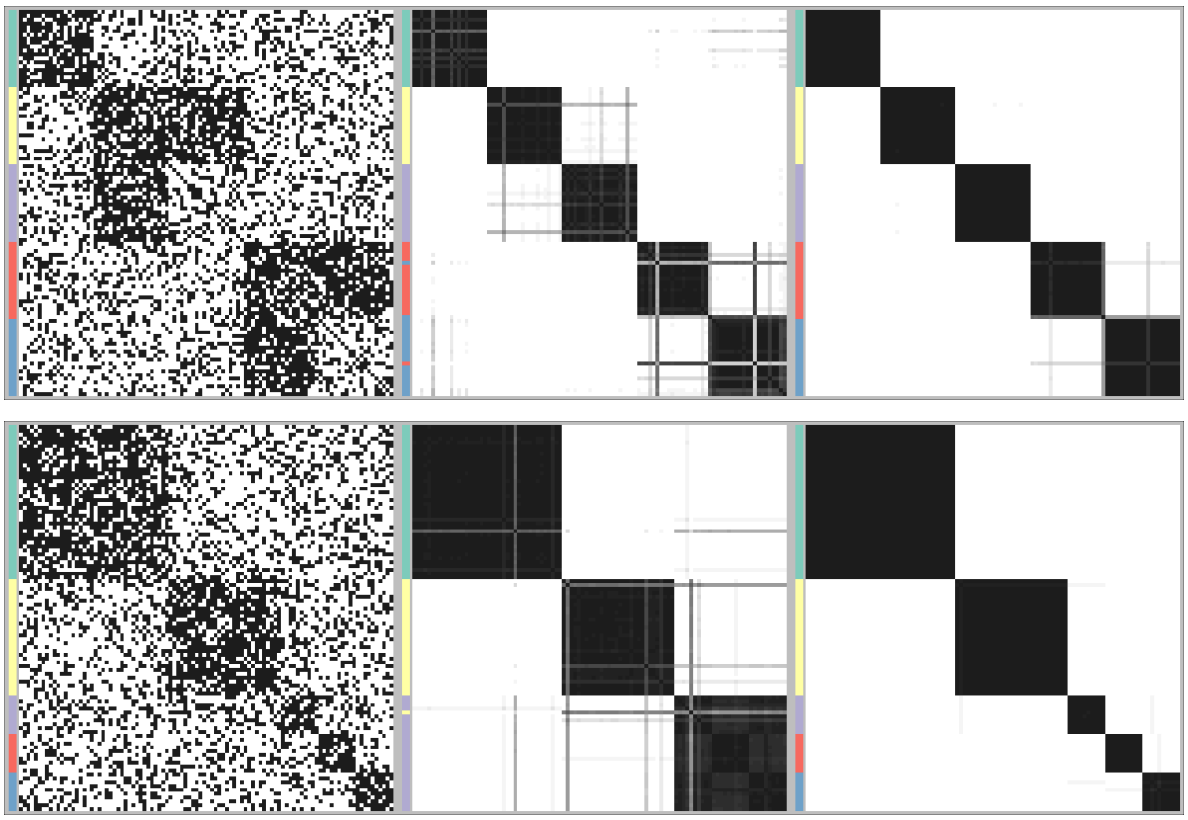

Figure 3.2: **Top:** first simulated network. **Bottom:** second simulated network. **Left:** observed adjacency matrix, with colors on the side corresponding to the true communities. **Center and right:** posterior co-clustering matrix under the Gnedin process from the ESBM without and with node attributes, respectively. Colors on the side correspond to the estimated partition.

	$\log p(\mathbf{Y})$	$\mathbb{E}[\text{VI}(\mathbf{z}, \mathbf{z}_0) \mid \mathbf{Y}]$	H	$\text{VI}(\hat{\mathbf{z}}, \mathbf{z}_0)$	$\text{VI}(\hat{\mathbf{z}}, \mathbf{z}_b)$
			NETWORK 1 [WITHOUT ATTRIBUTES]		
DM	−3101.7.0	0.648	7 [6,8]	0.303	0.887
DP	−3101.6	0.631	7 [6,8]	0.303	0.860
PY	−3123.6	0.554	6 [6,7]	0.303	0.780
GN	**−3097.5**	**0.519**	**5 [5,6]**	0.303	**0.724**
			NETWORK 1 [WITH ATTRIBUTES]		
DM	−3084.5	0.108	5 [5,6]	0.000	0.285
DP	−3083.7	0.105	5 [5,6]	0.000	0.265
PY	−3084.7	0.105	5 [5,6]	0.000	0.250
GN	**−3083.3**	**0.085**	**5 [5,5]**	0.000	**0.230**
			NETWORK 2 [WITHOUT ATTRIBUTES]		
DM	−3148.4	0.837	6 [5,7]	0.570	1.009
DP	−3146.6	0.819	6 [5,7]	0.570	0.979
PY	−3145.3	0.762	4 [3,5]	0.570	0.776
GN	**−3142.9**	**0.725**	4 [3,5]	0.570	**0.649**
			NETWORK 2 [WITH ATTRIBUTES]		
DM	−3123.7	0.052	5 [5,6]	0.000	0.189
DP	−3123.2	0.063	5 [5,6]	0.000	0.238
PY	−3124.0	0.081	5 [5,5]	0.000	0.285
GN	**−3119.9**	**0.031**	5 [5,5]	0.000	**0.116**

Table 3.2: Results of ESBMs in the two scenarios with $\overline{H}_0 = 5$ clusters. Performance is measured by marginal likelihood $\log p(\mathbf{Y})$, posterior mean of the variation of information distance $\mathbb{E}[\text{VI}(\mathbf{z}, \mathbf{z}_0) \mid \mathbf{Y}]$ from the true partition $\mathbf{z}_0$, posterior median number of non-empty clusters H (with first and third quartiles in brackets), distance $\text{VI}(\hat{\mathbf{z}}, \mathbf{z}_0)$ among the estimated and true partitions, and distance $\text{VI}(\hat{\mathbf{z}}, \mathbf{z}_b)$ among the estimated partition and the 95% credible bound.

3.5 Application to bill co-sponsorship networks

Motivated by the growing interest in the analysis of political networks (Fowler, 2006; Cranmer & Desmarais, 2011; Briatte, 2016; Signorelli & Wit, 2018), we apply our proposed ESBM class to the bill co-sponsorship network among the $V = 655$ members of the Italian parliament during the 2013–2018 mandate, whose composition is reported in Table 3.3. This dataset is extracted from the more extensive openly-available data in Briatte (2016), which contains legislative networks from 20 countries across years. Here, we consider the last available mandate of the Italian parliament and transform the original directed and weighted network into an undirected and binary one. Namely, we study the symmetric adjacency matrix $\mathbf{Y}$ with elements $y_{vu} = y_{uv} = 1$ if v co-sponsored a bill authored by u or viceversa, and $y_{vu} = y_{uv} = 0$ otherwise. The original edges were directed

PARTY	SEATS	LEFT-RIGHT	WINGS
Sinistra Ecologia Libertà	34	1.3	LEFT
Movimento 5 Stelle	104	2.6	M5S
Partito Democratico	314	2.6	LEFT
Per l'Italia: Centro Democratico	11	6.0	LEFT
Scelta Civica per Monti	30	6.0	LEFT
Forza Italia (Il Popolo della Libertà)	72	7.1	RIGHT
Lega Nord	22	7.8	RIGHT
Alleanza Nazionale	9	8.1	RIGHT
Area Popolare	31	–	RIGHT
Mixed or minor group	28	–	MIXED

Table 3.3: Composition of the Italian parliament during the 2013–2018 mandate. For each party, we report the number of seats, the left-right score (Briatte, 2016, with zero corresponding to extreme-left and 10 to extreme-right) and the political wing, that we use as node attribute

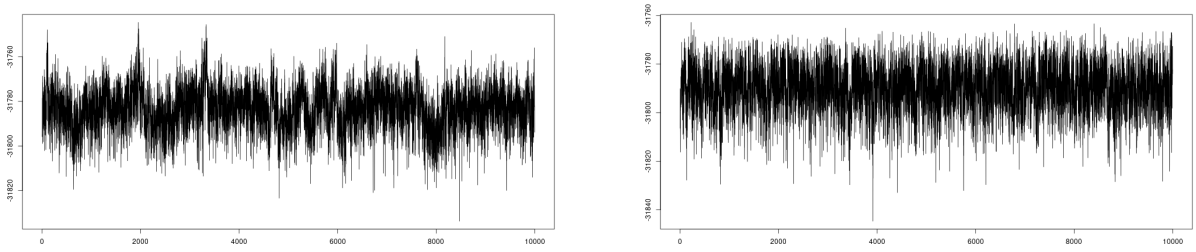

Figure 3.3: Traceplot of the marginal log-likelihood $\log p(\mathbf{Y})$ with wings as covariates (left) and without covariates (right), both after burn-in

towards the first author of the bill and weighted in inverse proportion to the number of co-sponsors and total bills. However, for our purposes, such direction and weight seem less informative than the presence or absence of a co-sponsorship.

Given its appealing theoretical properties and in the light of the results in simulations, we adopt a Gnedin process as a prior on $\mathbf{z}$ within our ESBM formulation. The hyperparameter γ is set to 0.5, corresponding to 20 expected clusters a priori, twice as many as the parties in the legislature. We run Algorithm 4, both with and without node attributes, for 10000 iterations after a burn-in of 10000. These settings provided adequate mixing, as shown in Figure 3.3. Due to the larger size of the network, running times were longer than in simulations, taking about one minute to produce 120 samples of $\mathbf{z}$. However, inference under Algorithm 4 remains feasible even for large networks such as the one in this study.

As node attributes, we employed the political wings reported in Table 3.3, which we found to be more informative than single parties, based on the marginal log-likelihood. In fact, the marginal log-likelihood with parties as covariates is -31833, a value sitting in

	$\log p(\mathbf{Y})$	H	$\text{VI}(\hat{\mathbf{z}}, \mathbf{z}_b)$
WITHOUT ATTRIBUTES	-31835	32 [32,33]	0.55
WITH ATTRIBUTES [WINGS]	$\mathbf{-31824}$	31 [31,31]	0.60

Table 3.4: Marginal log-likelihood $\log p(\mathbf{Y})$ and posterior summaries for the bill co-sponsorship network under the Gnedin process: posterior median number of occupied communities H (with first and third quartiles in brackets) and distance $\text{VI}(\hat{\mathbf{z}}, \mathbf{z}_b)$ among the estimated partition $\hat{\mathbf{z}}$ and the 95% VI credible bound.

between the ones obtained with wings as covariates and without covariates; see Table 3.4. These results suggest that parties are less effective in assisting inference compared to wings, which instead seem to carry information about the block structures in the network, thus making the wing-assisted model preferable to the unassisted one. This is confirmed by the adjacency matrix in the left panel of Figure 3.4, in which, by reordering the nodes according to the inferred partition, we can observe a recurrent core-periphery pattern within each wing. Such a structure was not visible in Figure 3.1 and suggests that only a subset of politicians are active in proposing new bills, whereas the others just tend to support the bills proposed by members of the same wing. This core-periphery pattern could not have been captured by algorithmic approaches, as shown in the right panel of Figure 3.4, which represents the co-clustering matrix obtained from the algorithm of Blondel et al. (2008), implemented in the function `cluster_louvain` from the R package `igraph`. The middle panel of Figure 3.4, instead, represents the posterior co-clustering matrix obtained from our Algorithm 4. This matrix is quite sharp, suggesting limited posterior uncertainty, as confirmed by Figure 3.5 and by the posterior summaries in Table 3.4. In fact, the radius of the credible ball is far below 1 under both models, while the maximum achievable VI distance is $\log_2 655 \approx 9.355$. The misclassification error of 0.05 confirms the satisfactory fit of the models.

The co-clustering matrix in the middle panel of Figure 3.4 also shows alliances among parties of the same wing and splits within larger parties, mostly due to core-periphery structures. This can be observed also in Figure 3.6, where clusters are visualized as nodes of a new weighted network, with weights given by the block probabilities estimated via (3.14). Party memberships within each cluster are represented via pie-charts, highlighting different fragmentation and aggregation levels for the different parties. For example, all members of *Lega Nord* belong to the same community, while *Movimento 5 Stelle* and *Partito Democratico* are split over several blocks. Right-wing politicians, instead, mostly belong to two main communities with different party proportions. The "geography" of such communities, induced by the block probabilities, reflects the left-right placement of Italian parties in Table 3.3, and highlights a polarization around three main forces, covering left parties, right parties and *Movimento 5 Stelle*, that are almost equidistant.

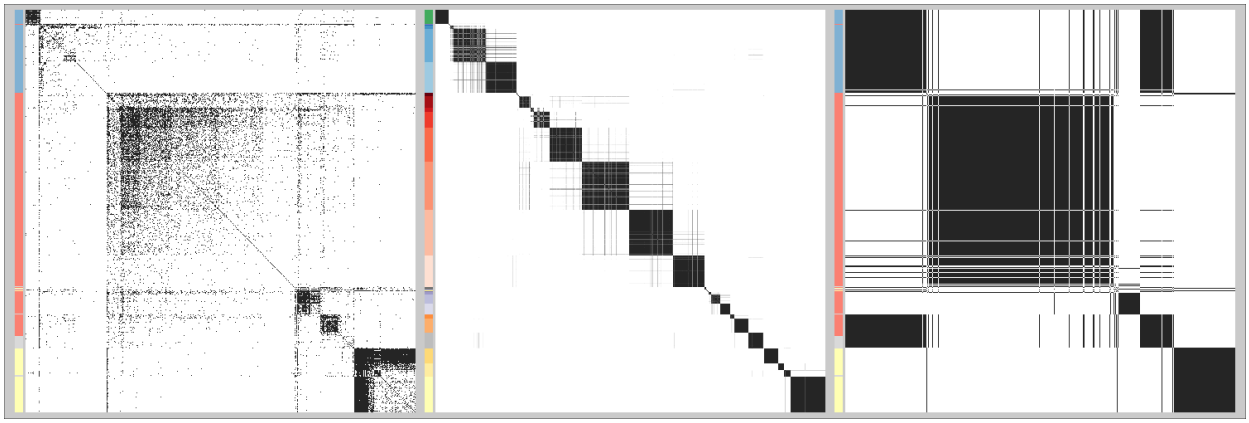

Figure 3.4: **Left:** observed bill co-sponsorship adjacency matrix, reordered according to the clusters estimated from Algorithm 4; colors on the side correspond to political wings, used as node attributes: blue for the right wing, red for the left wing, yellow for *Movimento 5 Stelle* and grey for the mixed group. **Middle:** posterior co-clustering matrix obtained from Algorithm 4 with wings as covariates; colors on the side denote estimated clusters, with shades proportional to the prevailing party: green for *Lega Nord*, blue for the rest of the right wing, red for *Partito Democratico* and *Per l'Italia: Centro Democratico*, orange for *Scelta Civica per Monti*, purple for *Sinistra Ecologia Libertà*, yellow for *Movimento 5 stelle*, grey for the mixed group. **Right:** co-clustering matrix obtained from the algorithm of Blondel et al. (2008), colors on the side correspond to political wings, as in the left panel

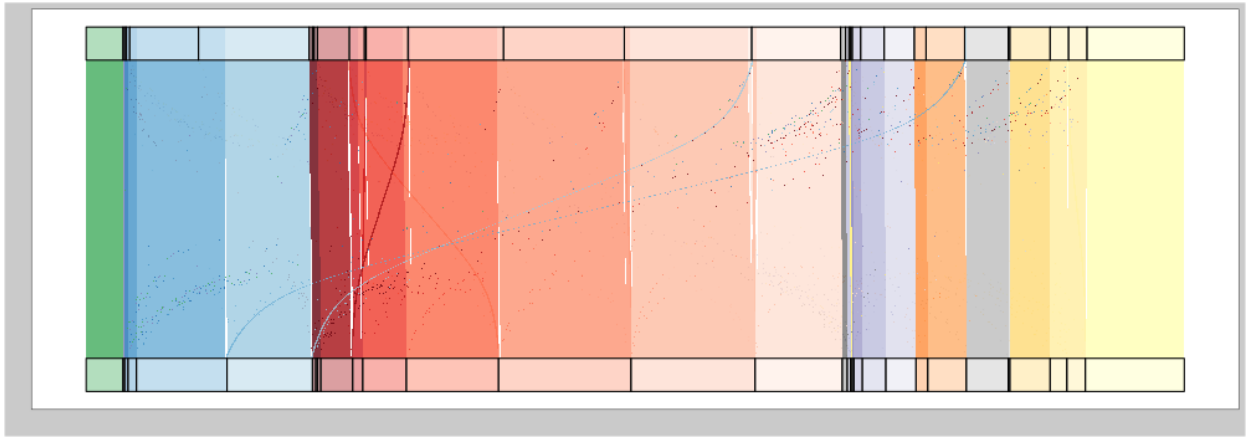

Figure 3.5: Riverplot highlighting which nodes change community when comparing the estimated partition $\hat{\mathbf{z}}$ with the bound $\mathbf{z}_b$ of the 95% credible ball around $\hat{\mathbf{z}}$. Party colors are the same as in Figure 3.4.

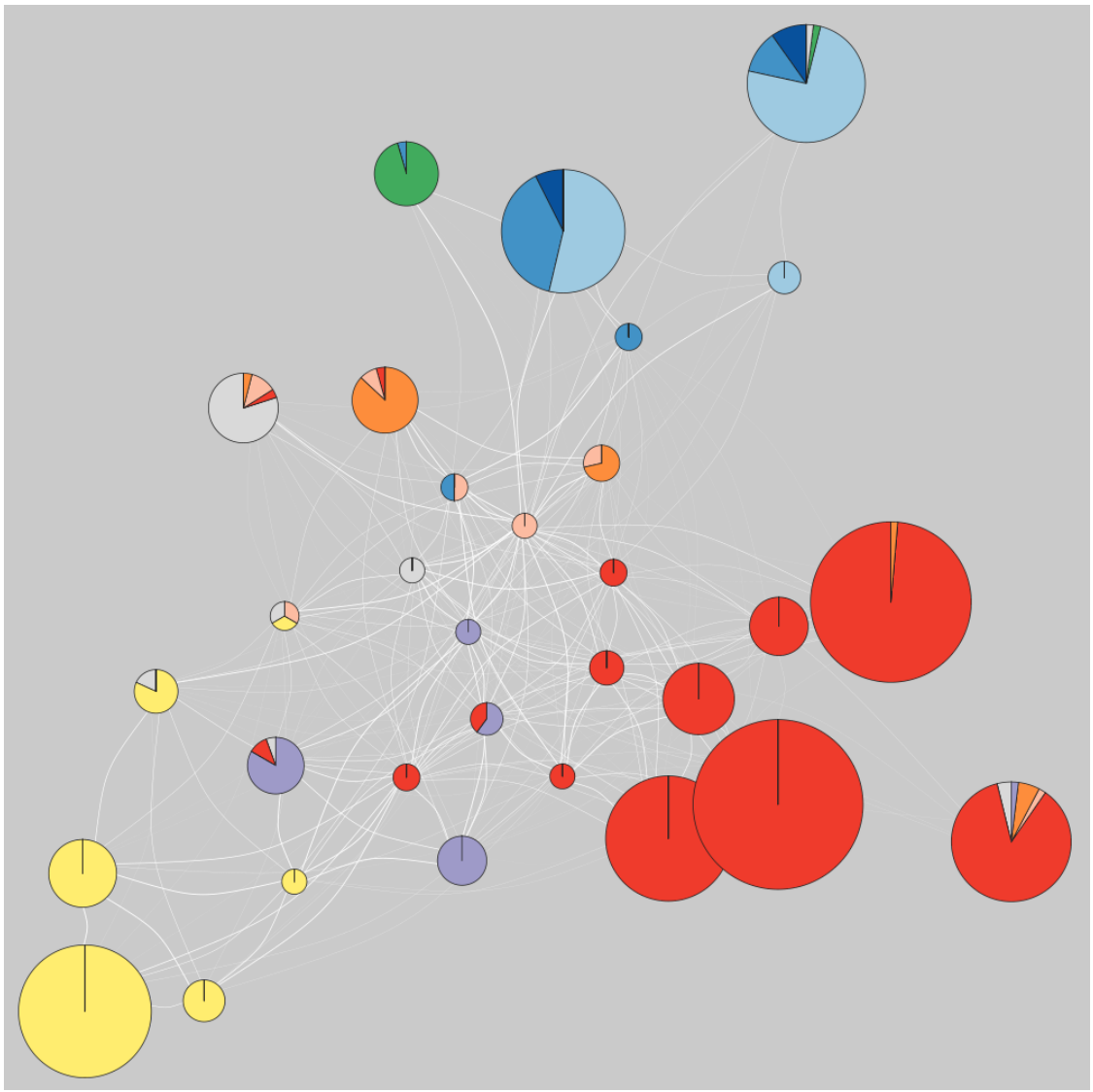

Figure 3.6: Network representation of inferred clusters. Each node represents a cluster, and edges are weighted by estimated block probabilities. Node sizes are proportional to cluster cardinalities, while pie-charts represent compositions with respect to party affiliations (party colors are the same as in Figure 3.4). Node placement reflects the strength of connections (higher block probabilities yield to closer nodes).

Chapter 4

Bayesian testing for partition structures in stochastic block models

4.1 Introduction

As shown in Chapter 3, stochastic block models (SBMs) can effectively capture community structures in networks, and typically benefit from incorporating node-specific attributes in the model. Such attributes often come in the form of exogenous partitions; for example, in the bill co-sponsorship application of § 3.5, politicians are pre-clustered in parties. Common proposals incorporate nodal attributes within the generative model for community assignments (e.g. Tallberg, 2004; White & Murphy, 2016; Newman & Clauset, 2016; Stanley et al., 2019) or define edge probabilities as a function of block-specific parameters, as in classical SBMs, and of pairwise similarity measures among node attributes (e.g. Mariadassou et al., 2010; Choi et al., 2012; Sweet, 2015; Roy et al., 2019).

However, less attention has been paid to the development of Bayesian testing methods to quantify whether or not these external partitions and the endogenous network block structures are related. For example, in structural brain network studies it is often of interest whether exogenous anatomical partitions of brain regions depart from endogenous block structures of brain networks (e.g. Sporns, 2013; Faskowitz et al., 2018). This goal could be partially addressed by the previously-mentioned models via inference on the posterior distribution for the parameters regulating the effect of the node-specific attributes, but these formulations are prone to identifiability and computational issues.

Motivated by the above discussion, in this chapter we propose a formal, yet simple, Bayesian testing procedure quantifying to what extent an exogenous node partition effectively characterizes block structures induced by the endogenous stochastic equivalence relations within the network. In this regard, our solution is more similar to those of Bianconi et al. (2009) and Peel et al. (2017). However such contributions compute, under a frequentist perspective, the entropy of a SBM whose communities coincide

with the external node partition and compare it with the distribution of the entropies derived under the same network with communities given by a random permutation of the exogenous node labels. Besides taking a Bayesian approach, our procedure quantifies proximities to endogenous block structures rather than studying departures from a random partition. This allows, as a byproduct, inference on node communities supported by the data, either exogenous or endogenous.

As described in § 4.2, this goal is accomplished by the calculation of the Bayes factor (e.g. Kass & Raftery, 1995) among a SBM with known community structure given by the exogenous node partition and an infinite relational model (Kemp et al., 2006) that quantifies uncertainty in the endogenous community assignments via a CRP prior (Aldous, 1985) allowing the total number of non-empty communities to be inferred. In § 4.3, we derive a collapsed Gibbs sampler to sample from the posterior of the endogenous partition, thus allowing Monte Carlo estimation of the marginal likelihood (Newton & Raftery, 1994; Raftery et al., 2007) required for calculating the Bayes factor. As illustrated in a simulation in § 4.4 and in an application to Alzheimer's brain networks in § 4.5, this Gibbs sampler is also useful to perform posterior inference on endogenous partitions, in case the exogenous ones are not sufficient to characterize the network topology.

4.2 Model formulation and hypothesis testing

Consider an undirected binary network without self-loops and let $\mathbf{Y}$ be the associated $V \times V$ symmetric adjacency matrix having elements $y_{vu} = y_{uv} = 1$ if nodes $v = 2, \dots, V$ and $u = 1, \dots, v-1$ are connected, and $y_{vu} = y_{uv} = 0$ otherwise. The absence of self-loops implies that all the diagonal entries of $\mathbf{Y}$ are not considered for inference. We employ the same SBM formulation as described in § 3.1, resulting in the following marginal likelihood for $\mathbf{Y}$ given the vector of community membership indicators for the V nodes, $\mathbf{z} = (z_1, \dots, z_V)^\intercal \in \mathcal{Z} = \{1, \dots, H\}^V$:

$$p(\mathbf{Y} \mid \mathbf{z}) = \prod_{h=1}^{H} \prod_{k=1}^{h} \frac{B(a + m_{hk}, b + \overline{m}_{hk})}{B(a, b)}, \tag{4.1}$$

where m_{hk} and $\overline{m}_{hk}$ denote the number of edges and non-edges among nodes in communities h and k, respectively, while $B(\cdot, \cdot)$ is the beta function. As clarified later, (4.1) is fundamental to compute Bayes factors and develop a collapsed Gibbs sampler.

Recalling § 4.1, our goal is to develop a Bayesian testing procedure to assess whether assuming $\mathbf{z}$ as known and equal to an exogenous assignment vector $\mathbf{z}^*$ produces an effective characterization of the block structures in $\mathbf{Y}$, relative to what would be obtained by letting $\mathbf{z}$ unknown and endogenously determined by the stochastic equivalence

relations in $\mathbf{Y}$. Letting $\mathcal{M}^*$ and $\mathcal{M}$ be the first and the second hypothesized models, respectively, and assuming that these two competing formulations are equally likely a priori, i.e. $p(\mathcal{M}) = p(\mathcal{M}^*)$, we address this goal by studying the Bayes factor

$$\mathcal{B}_{\mathcal{M},\mathcal{M}^*} = \frac{p(\mathbf{Y} \mid \mathcal{M})}{p(\mathbf{Y} \mid \mathcal{M}^*)} = \frac{\sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{Y} \mid \mathbf{z})p(\mathbf{z})}{p(\mathbf{Y} \mid \mathbf{z}^*)}, \tag{4.2}$$

where $\sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{Y} \mid \mathbf{z})p(\mathbf{z})$ and $p(\mathbf{Y} \mid \mathbf{z}^*)$ are the marginal likelihoods of $\mathcal{M}$ and $\mathcal{M}^*$. Under $\mathcal{M}^*$, the quantity $p(\mathbf{Y} \mid \mathbf{z}^*)$ can be computed by evaluating (4.1) at $\mathbf{z} = \mathbf{z}^*$. Instead, model $\mathcal{M}$ also requires a prior $p(\mathbf{z})$ on the assignments, which are unknown and have to be inferred from the equivalence relations in the observed network. Recalling, e.g., Kass & Raftery (1995), equation (4.2) defines a formal Bayesian procedure to assess evidence against $\mathcal{M}^*$ relative to $\mathcal{M}$, with high values suggesting that the exogenous assignments in $\mathbf{z}^*$ are not as effective in characterizing the endogenous block structures in $\mathbf{Y}$ as the posterior for $\mathbf{z}$ under $\mathcal{M}$. Note that, since $p(\mathcal{M}) = p(\mathcal{M}^*)$, the Bayes factor in (4.2) coincides with the posterior odds $p(\mathcal{M} \mid \mathbf{Y})/p(\mathcal{M}^* \mid \mathbf{Y})$. When $p(\mathcal{M}) \neq p(\mathcal{M}^*)$, it suffices to multiply $\mathcal{B}_{\mathcal{M},\mathcal{M}^*}$ by $p(\mathcal{M})/p(\mathcal{M}^*)$ to assess posterior evidence against $\mathcal{M}^*$ relative to $\mathcal{M}$.

To complete our Bayesian specification, we specify a prior distribution for the indicators $\mathbf{z} = (z_1, \ldots, z_V)^\intercal$ under model $\mathcal{M}$. Among the SBM variants illustrated in Chapter 3, here we employ the widely used infinite relational model (Kemp et al., 2006; Schmidt & Morup, 2013). This formulation is based on the CRP (Aldous, 1985), so that each community attracts new nodes in proportion to its size, and the formation of new communities depends only on the size of the network and on a tuning parameter $\alpha > 0$. More specifically, we assume the following prior over community indicators for node v, conditioned on the memberships $\mathbf{z}_{-v} = (z_1, \ldots, z_{v-1}, z_{v+1}, \ldots, z_V)^\intercal$ of the other nodes:

$$\text{pr}(z_v = h \mid \mathbf{z}_{-v}) = \begin{cases} n_{h,-v}/(V-1+\alpha), & \text{if } h = 1, \ldots, H_{-v}, \\ \alpha/(V-1+\alpha), & \text{if } h = H_{-v}+1, \end{cases} \tag{4.3}$$

where $n_{h,-v}$ is the number of nodes associated with community h, excluding node v, whereas $\alpha > 0$ is a tuning parameter controlling the expected number of non-empty communities; see also Gershman & Blei (2012) for an introductory overview of the CRP and related priors.

The above urn representation of the CRP is induced by the joint probability mass function $p(\mathbf{z}) = \alpha^H[\prod_{h=1}^H (n_h - 1)!][\prod_{v=1}^V (v-1+\alpha)]^{-1}$ for the entire vector $\mathbf{z}$, which can be used to compute the summation in (4.2). Although theoretically feasible, this approach is clearly computationally impractical due to the huge cardinality of the set $\mathcal{Z}$, thus requiring alternative strategies relying on Monte Carlo estimation of $p(\mathbf{Y} \mid \mathcal{M}) = \sum_{\mathbf{z}\in\mathcal{Z}} p(\mathbf{Y} \mid \mathbf{z})p(\mathbf{z})$ via importance sampling methods. Here, we rely on the

harmonic mean approach (Newton & Raftery, 1994; Raftery et al., 2007), thus obtaining

$$\hat{p}(\mathbf{Y} \mid \mathcal{M}) = \left[\frac{1}{R} \sum_{r=1}^{R} \frac{1}{p(\mathbf{Y} \mid \mathbf{z}^{(r)})} \right]^{-1}, \tag{4.4}$$

where $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(R)}$ are samples from the posterior distribution of $\mathbf{z}$ and $p(\mathbf{Y} \mid \mathbf{z}^{(r)})$ is the marginal likelihood in (4.1) evaluated at $\mathbf{z} = \mathbf{z}^{(r)}$, for every $r = 1, \ldots, R$. The harmonic mean approach provides a consistent strategy to evaluate marginal likelihoods and, due to its simplicity, is widely implemented. Although recent refinements have been proposed to address some shortcomings of the harmonic estimate (Lenk, 2009; Pajor, 2017), here we consider the original formula which is computationally more tractable and has proved stable in our simulations and applications; see Figures 4.2 and 4.4.

Leveraging equations (4.1) and (4.4), our estimate of the Bayes factor in (4.2) is

$$\hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}^*} = \frac{\hat{p}(\mathbf{Y} \mid \mathcal{M})}{p(\mathbf{Y} \mid \mathcal{M}^*)} = \frac{\left[\frac{1}{R} \sum_{r=1}^{R} \prod_{h=1}^{H^{(r)}} \prod_{k=1}^{h} \frac{B(a,b)}{B(a+m_{hk}^{(r)}, b+\overline{m}_{hk}^{(r)})} \right]^{-1}}{\prod_{h=1}^{H^*} \prod_{k=1}^{h} \frac{B(a+m_{hk}^*, b+\overline{m}_{hk}^*)}{B(a,b)}}, \tag{4.5}$$

where $m_{hk}^{(r)}$ and $\overline{m}_{hk}^{(r)}$ refer to the counts of edges and non-edges among nodes in communities h and k induced by the rth MCMC sample of $\mathbf{z}$, whereas m_{hk}^* and $\overline{m}_{hk}^*$ denote the number of edges and non-edges among the nodes in communities h and k identified by the exogenous assignments $\mathbf{z}^*$. Finally, $H^{(r)}$ and H^* are the total numbers of unique labels in $\mathbf{z}^{(r)}$ and $\mathbf{z}^*$. In § 4.3 we describe the collapsed Gibbs algorithm to sample the assignment vectors $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(R)}$ from the posterior $p(\mathbf{z} \mid \mathbf{Y})$ under model $\mathcal{M}$. These samples are required to compute (4.5) and also allow inference on the posterior distribution of the endogenous partitions in $\mathbf{Y}$.

4.3 Posterior computation via collapsed Gibbs sampling

Posterior inference for the marginal model (4.1) with CRP prior (4.3) proceeds via a simple collapsed Gibbs sampler which updates the community assignment of each node v conditioned on those of the others by sampling from the full-conditional distribution $p(z_v \mid \mathbf{Y}, \mathbf{z}_{-v})$. Algorithm 5 provides the detailed steps of one cycle of the Gibbs sampler. Note that since (4.1) is the joint probability for a large set of binary edges, manipulating this quantity within Algorithm 5 and in computing the Bayes factor in (4.5) may lead to practical difficulties due to dealing with quantities very close to zero. In such settings, we suggest to work with the logarithm whenever possible, and to leverage the identity $\log[\sum_i \exp(v_i)] = d + \log[\sum_i \exp(v_i - d)]$, where d usually coincides with $\max_i v_i$.

Algorithm 5: One step of the Gibbs sampler for $\mathbf{z}$ under $\mathcal{M}$

for $v = 1, \dots, V$ **do**

1. Remove node v from the node set.
2. If no other node belongs to the community of v, such a community is removed.
3. Reorder the community indices so that $1, \dots, H_{-v}$ are non-empty and sample z_v from the categorical variable with full-conditional probabilities

$$\text{pr}(z_v = h \mid \mathbf{Y}, \mathbf{z}_{-v}) \propto \begin{cases} \frac{n_{h,-v}}{V-1+\alpha} p(\mathbf{Y} \mid z_v = h, \mathbf{z}_{-v}), & \text{if } h = 1, \dots, H_{-v}, \\ \frac{\alpha}{V-1+\alpha} p(\mathbf{Y} \mid z_v = h, \mathbf{z}_{-v}), & \text{if } h = H_{-v} + 1, \end{cases}$$

where $p(\mathbf{Y} \mid z_v = h, \mathbf{z}_{-v})$ is computed as in (4.1) conditioned on $\mathbf{z}_{-v}$ and $z_v = h$.

return $\mathbf{z} = (z_1, \dots, z_V)^\intercal$

As for point estimation of $\mathbf{z}$ and uncertainty quantification about it, in the subsequent quantitative studies we adapt to the network setting the approach of Wade & Ghahramani (2018), which is based on *variation of information* metrics and offers desirable theoretical properties alongside accurate empirical performance.

4.4 Simulation studies

We consider an illustrative simulation to assess the performance of the testing procedure we proposed in § 4.2, and to evaluate the ability of model $\mathcal{M}$ to recover underlying endogenous partition structures. Consistent with this goal, we simulate a symmetric binary adjacency matrix $\mathbf{Y}$ from a stochastic block model for $V = 60$ nodes partitioned into $H_0 = 3$ communities of equal size. More specifically, we let $\mathbf{z}_0 = (z_{1,0} = 1, \dots, z_{20,0} = 1, z_{21,0} = 2, \dots, z_{40,0} = 2, z_{41,0} = 3, \dots, z_{60,0} = 3)^\intercal$, and simulate each $y_{vu} = y_{uv}$ for $v = 2, \dots, V$, $u = 1, \dots, v-1$ from a Bernoulli with probability 0.8 if nodes v and u are in the same community, and 0.2 otherwise.

In performing posterior inference on the endogenous community structure under model $\mathcal{M}$, we set $a = b = 1$ to induce a uniform prior on the block probabilities. This choice is theoretically supported (e.g. Ghosh et al., 2019) and has been widely employed in routine implementations of SBMs (e.g. Nowicki & Snijders, 2001; Kemp et al., 2006; Geng et al., 2019). As for α in prior (4.3), we set it equal to 1 following default specifications of CRP, thus circumventing the need to include a hyper-prior which could affect mixing and convergence of Algorithm 5. Such a default specification has proved effective both in simulations and applications, and we found the results robust to moderate changes in α.

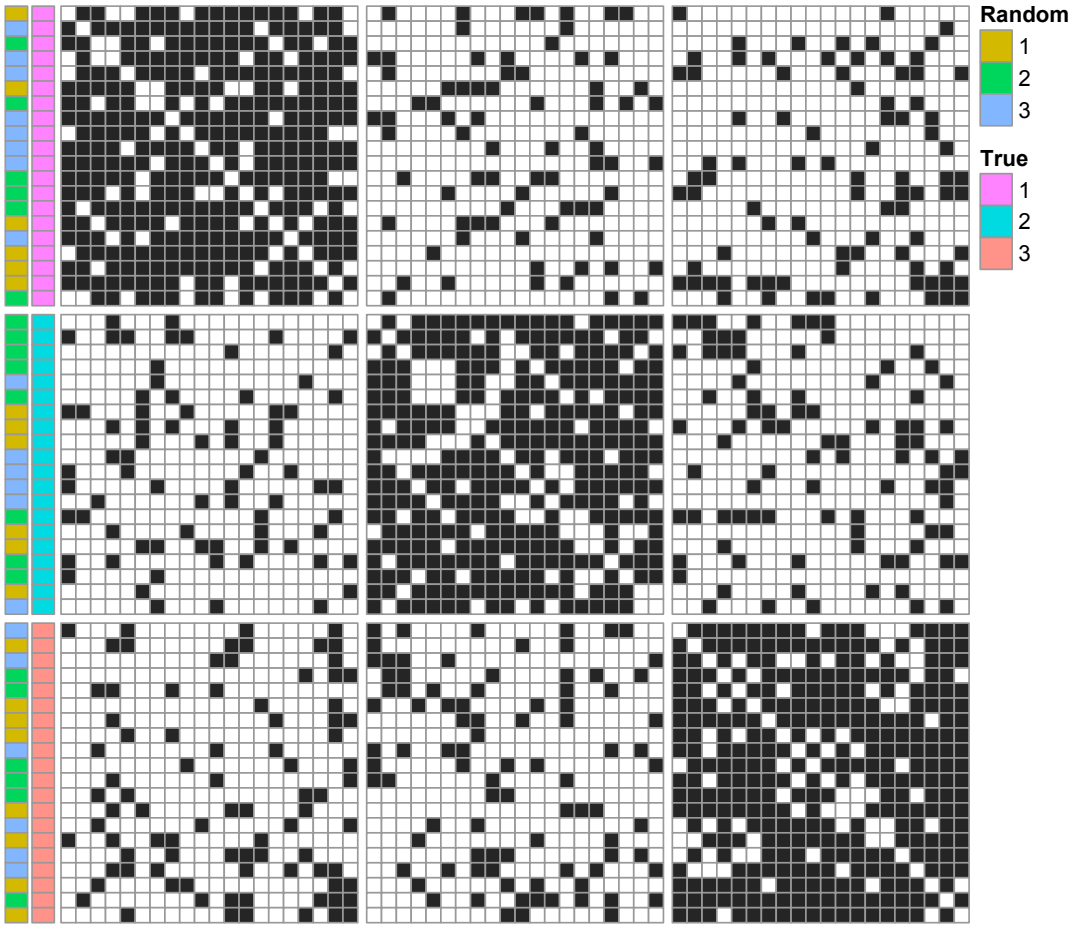

Figure 4.1: Simulated adjacency matrix $\mathbf{Y}$ divided in blocks according to the estimated endogenous assignments $\hat{\mathbf{z}}$. Black and white cells denote edges and non-edges, respectively, whereas the colors on the side represent the true assignments and a random permutation of the same.

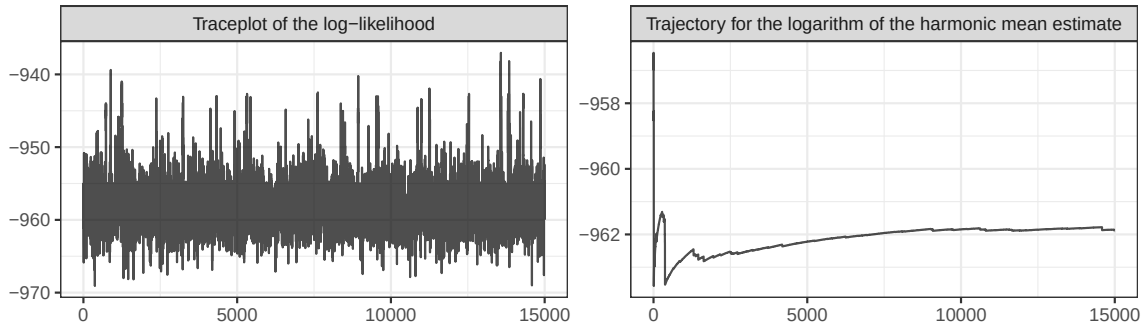

Figure 4.2: MCMC diagnostics for the simulation study. Left: traceplot for the logarithm of the likelihood in equation (4.1) computed at the MCMC samples of $\mathbf{z}$ after burn-in. Right: trajectory of the logarithm of the harmonic mean estimate in equation (4.4) for growing R.

Figure 4.1 shows the simulated adjacency matrix $\mathbf{Y}$ and highlights its endogenous block partition according to the estimated $\hat{\mathbf{z}}$ under model $\mathcal{M}$. Such an estimate relies on 15000 MCMC samples produced by Algorithm 5, after a burn-in of 2000. As shown in Figure 4.2, such settings are sufficient for good convergence and mixing according to the MCMC diagnostics of key measures for posterior inference, covering the traceplot of the log-likelihood in (4.1) under model $\mathcal{M}$ and the trajectory of the logarithm of the harmonic mean estimate for the associated marginal likelihood in (4.4). As is clear from the block partition of $\mathbf{Y}$ in Figure 4.1, the posterior for $\mathbf{z}$ under model $\mathcal{M}$ is able to concentrate around the true underlying endogenous partition and allows learning of the correct number of non-empty groups. These results support the use of $\mathcal{M}$ as a benchmark model to test for differences between endogenous and exogenous partitions under the methods presented in § 4.2.

To assess the quality of such strategies, we consider the two external assignment vectors $\mathbf{z}_1^*$ and $\mathbf{z}_2^*$ displayed in Figure 4.1. The first coincides with the true underlying community structure $\mathbf{z}_0$, whereas the second is obtained by a random permutation of the indices in $\mathbf{z}_0$. Due to this, we expect $\hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}_1^*}$ and $\hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}_2^*}$ to provide evidence in favor of $\mathcal{M}_1^*$ and against $\mathcal{M}_2^*$ relative to $\mathcal{M}$, respectively. Indeed, we obtain $2\log\hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}_1^*} = -5.25$ and $2\log\hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}_2^*} = 518.93$, which confirm our expectations when compared with the thresholds in Kass & Raftery (1995). Note that, although $\hat{\mathbf{z}} = \mathbf{z}_1^* = \mathbf{z}_0$, we obtain a negative $\log\hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}_1^*}$, which leads to strong preference for $\mathcal{M}_1^*$ relative to $\mathcal{M}$. Indeed, even if the point estimate of $\mathbf{z}$ under $\mathcal{M}$ exactly recovers $\mathbf{z}_0$, there is still some amount of posterior uncertainty induced by the CRP prior on $\mathbf{z}$. On the contrary, $\mathcal{M}_1^*$ is defined by conditioning on the true underlying partition with no uncertainty, thus providing a formulation much closer to the true data-generative mechanism relative to $\mathcal{M}$.

4.5 Application to brain networks of Alzheimer's individuals

There is an intensive research effort aimed at finding the sources of the Alzheimer's disease in human brain networks. Such an increasing interest has been motivated by recent developments in brain imaging technologies and by the constant growth of elderly population in the age range mostly affected by Alzheimer's, thus making such a disease a major concern for developed countries both in terms of disability and mortality (Ashford et al., 2011a,b; Stam, 2014). Here, we focus on the analysis of structural brain networks encoding presence or absence of white matter fibers among anatomical regions in the human brain. These connectivity data have been a source of interest in several recent studies mostly focused on topological summary measures of Alzheimer's brains and on how these measures change as the disease progresses (e.g. Daianu et al., 2013; Sulaimany

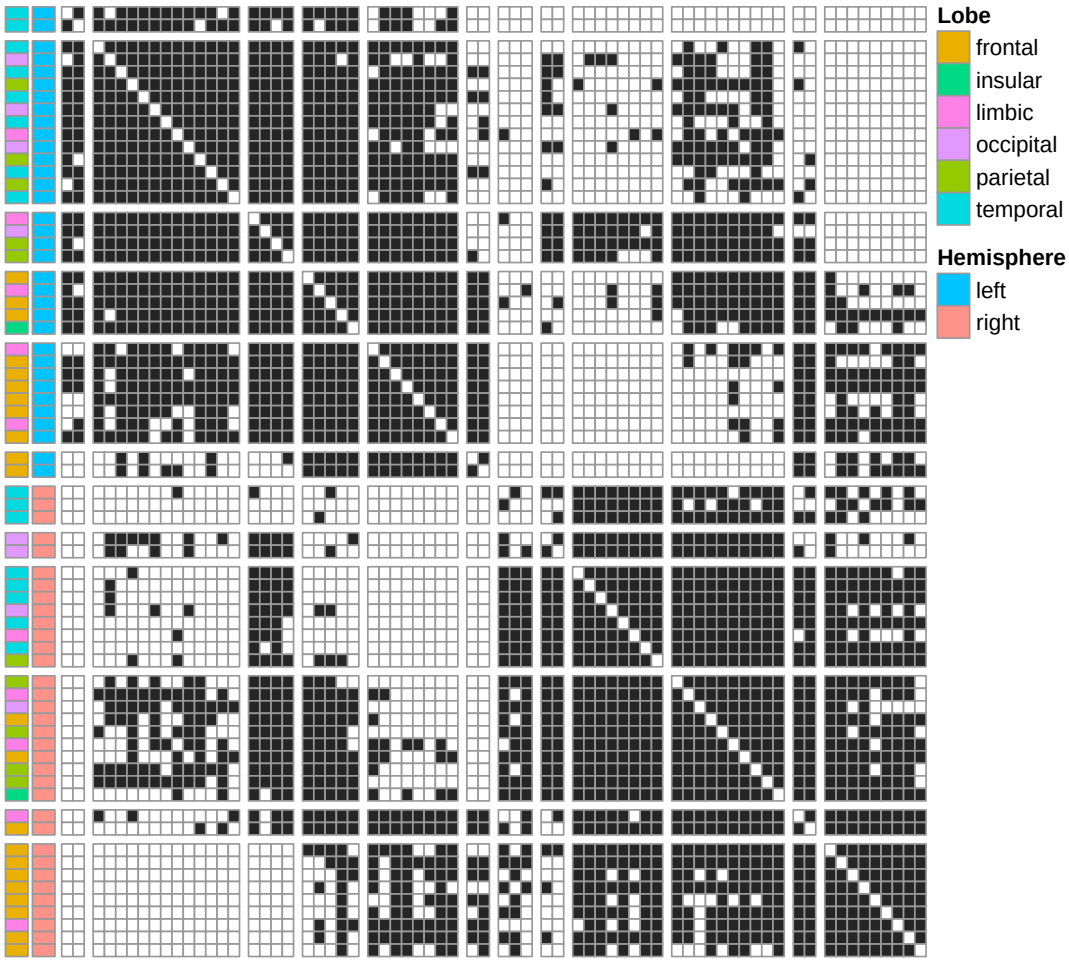

Figure 4.3: Adjacency matrix $\mathbf{Y}$ of a representative brain network for Alzheimer's individuals. Brain regions are re-ordered and divided in blocks according to the estimated endogenous assignments $\hat{\mathbf{z}}$. Black and white cells denote edges and non-edges, respectively, whereas the colors on the side represent the two exogenous anatomical brain partitions corresponding to lobes and hemispheres, respectively.

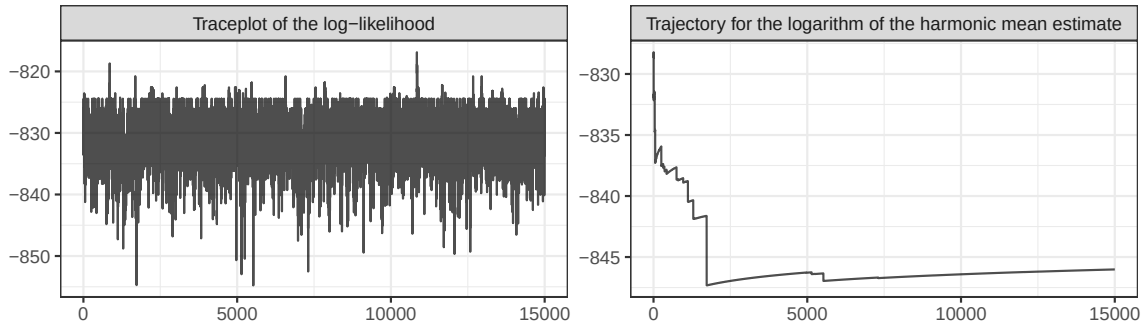

Figure 4.4: MCMC diagnostics for the application. Left: traceplot for the logarithm of the likelihood in equation (4.1) computed at the MCMC samples of $\mathbf{z}$ after burn-in. Right: trajectory of the logarithm of the harmonic mean estimate in equation (4.4) for growing R.

Table 4.1: Results of our proposed procedure for testing to what extent exogenous brain partitions are as effective as model $\mathcal{M}$ in characterizing the endogenous block structure for a representative brain network of Alzheimer's individuals. Here, we focus on anatomical partitions and on community structures identified in representative brains of individuals characterized by three ordered early stages of the disease progress.

	Anatomical		Cognitive Decline Progression		
	Hemisphere	Lobes	Normal Aging	Early Decline	Late Decline
$2 \log \hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}^*}$	712.33	1290.50	155.01	100.21	39.88

et al., 2017; John et al., 2017; Mårtensson et al., 2018). Instead, we consider a different perspective by studying the endogenous block structures in a representative Alzheimer's brain network, while assessing whether exogenous partitions of interest can effectively characterize these endogenous blocks.

Consistent with the above goal, we apply methods in § 4.2–4.3 to the 68×68 binary adjacency matrix $\mathbf{Y}$ encoding presence or absence of white matter fibers among anatomical regions in a representative Alzheimer's brain network provided by Sulaimany et al. (2017). In this study, brain regions are defined by the Desikan atlas (Desikan et al., 2006), which provides additional information on hemisphere and lobe memberships (Kang et al., 2012); see Sulaimany et al. (2017) for additional details on the construction of $\mathbf{Y}$. Figure 4.3 provides a graphical representation of $\mathbf{Y}$ with brain regions suitably reordered and partitioned in blocks according to the estimated endogenous assignments $\hat{\mathbf{z}}$. The latter are obtained by considering the same MCMC settings and hyperparameters of the simulation study, which proved effective and robust also in this application; see Figure 4.4. As shown in Figure 4.3, we learn $\hat{\mathrm{H}} = 12$ endogenous communities equally divided between the two hemispheres and showing an overall coherence of the partition structure across left and right regions. As expected, there is an evident block connectivity within hemispheres, although some communities also display a tendency to connect across hemispheres. For example, brain regions in the frontal lobe tend to create two highly interconnected communities, one in each hemisphere, with these two blocks showing also a preference to create bridges among the two hemispheres. Despite these anatomical homophily structures, as highlighted in Figure 4.3 and in Table 4.1, hemisphere and lobe partitions are not sufficient to fully characterize the endogenous block structures in Alzheimer's brains. There are, in fact, various sub-blocks within each hemisphere and these communities typically comprise regions in different lobes.

We conclude by studying if exogenous partitions inferred from representative brains of individuals in three ordered early stages of the disease progress can effectively explain also the endogenous block structures in Alzheimer's brains. To accomplish this goal, we

first apply Algorithm 5 to the representative adjacency matrices of individuals characterized by normal aging, early and late cognitive decline (Sulaimany et al., 2017), and then quantify, via the Bayes factors in Table 4.1, whether these exogenously-determined partitions are also effective in modeling the block structures within the Alzheimer's brain. Although $2 \log \hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}^*}$ is above the threshold in Kass & Raftery (1995) suggesting strong evidence against this hypothesis for all three stages, it is interesting to notice how $2 \log \hat{\mathcal{B}}_{\mathcal{M},\mathcal{M}^*}$ decreases as cognitive decline progresses towards Alzheimers' disease. This means that the inferred partitions could be used, with caution, as a diagnostic strategy to identify the progress of the disease.

Chapter 5

Discussion

The ideas presented in this thesis open up several directions for future research. For instance, the cumulative shrinkage process proposed in Chapter 2 can be employed within several models other than Gaussian factor models, which served here as an illustrative example. Another setting in which the cumulative shrinkage process can be applied is Poisson factorization (Dunson & Herring, 2005), a probabilistic model for count data that has been successfully used for recommendation systems, both in a finite-dimensional version (Gopalan et al., 2015) and in a non-parametric one (Gopalan et al., 2014). In Poisson factorization, the number y_{ij} of purchases of item j by user i is modelled with a Poisson random variable of rate λ_{ij} $(i = 1, \dots, n;\ j = 1, \dots, p)$, independently for each pair (i,j), with $\lambda_{ij} = \gamma_i^T \psi_j \in \mathbb{R}_+$ being the quadratic combination of user preferences $\gamma_i = (\gamma_{i1}, \dots, \gamma_{iH})^T \in \mathbb{R}_+^H$ and item attributes $\psi_j = (\psi_{j1}, \dots, \psi_{jH})^T \in \mathbb{R}_+^H$. This corresponds to factorizing the matrix of Poisson rates $\Lambda = [\lambda_{ij}]$ as

$$\Lambda = \Gamma \Psi^T, \tag{5.1}$$

with $\Gamma = [\gamma_{ih}]$ $(i = 1, \dots, n;\ h = 1, \dots, H)$ and $\Psi = [\psi_{jh}]$ $(j = 1, \dots, p;\ h = 1, \dots, H)$. The similarity between (5.1) and (1.1) highlights the connection with factor models and suggests that increasing shrinkage, namely across the columns of Γ and Ψ, could be appropriate in Poisson factorization as well. A characteristic feature of Λ in recommendation systems is that even non-negligible columns are sparse. This requires to carefully design an instance of the cumulative shrinkage process that is able to capture such a sparsity.

Another possible extension of the cumulative shrinkage process involves the parameter space. Namely, in Chapter 2 we have defined the cumulative shrinkage process for sequences of real parameters. However, Definition 2.1 can be extended to sequences in $\mathbb{R}^p$ or even more general metric spaces without affecting the properties in § 2.2. This allows to apply the proposed approach to other data types.

In Chapter 3 we have proposed the extended stochastic block model (ESBM), which is already very general in that it encompasses the main existing SBM formulations, while facilitating the proposal of new ones by exploiting the whole class of Gibbs–type priors (Gnedin & Pitman, 2005; De Blasi et al., 2013a). However, here we focused on undirected binary networks, but our approach can be extended to directed, bipartite and weighted networks. In this sense, the ESBM is as extensible as the infinite relational model (see Schmidt & Morup, 2013, §II.D). These extensions only involve the structure of the adjacency matrix $\mathbf{Y}$ and the block–probabilities matrix $\boldsymbol{\Theta}$, hence impacting their likelihoods (3.1) and (3.2), but the rest of the model remains unchanged. Namely, directed networks allow asymmetric $\mathbf{Y}$ and $\boldsymbol{\Theta}$, implying that (3.1) and (3.2) factorize over all their entries rather than just on sub-diagonal ones. Instead, bipartite networks (i.e. networks with edges among two different sets of nodes, possibly with different cardinalities) have in general non-square adjacency matrices and admit separate clusterings for the two sets of nodes, which can be achieved through two independent Gibbs–type priors. Finally, weighted networks require to substitute the Bernoulli distributions in (3.1) with suitable ones, for example Poisson distributions for positive integer weights and Gaussians for continuous ones. To preserve the availability of a marginal likelihood for $\mathbf{Y}$ as in (3.3), the priors on the entries of $\boldsymbol{\Theta}$ must be chosen in order to exploit conjugacy. For example, the Beta priors in (3.2) may be replaced by Gamma distributions in case of Poisson weights, or by Normal-inverse-gamma distributions in case of continuous Gaussian weights. Moreover, here we considered categorical node attributes, but it is easy to imagine settings with a different type of attributes. In this sense, the approach of Müller et al. (2011) that we adopted is particularly suitable since it applies to general attributes, just requiring to choose the appropriate function $p(\cdot)$ in (3.7), for which the authors provide default choices also for the cases of continuous, ordinal and count–type attributes.

Finally, in Chapter 4 we outlined a Bayesian testing procedure to assess whether exogenous node partitions can effectively capture endogenous block structures in a network. To accomplish this goal we employed the infinite relational model (Kemp et al., 2006) to learn endogenous communities. However, our testing procedure applies not only to the infinite relational model, but to the whole ESBM class. Moreover, instead of testing an exogenous partition against an endogenous one, inferred via a SBM, our approach also allows to compare two exogenous partitions or two endogenous ones. The first task is even simpler than the one analyzed in Chapter 4, since the marginal likelihood (4.1) can be computed in closed–form for both the external partitions under comparison, thus avoiding the need of MCMC methods. The second task is computationally similar to the one considered in Chapter 4, and allows to compare different instances of the ESBM class (see also § 3.3.2) in order to identify the most appropriate for the given context.

Bibliography

Airoldi, E. M., Blei, D. M., Fienberg, S. E. & Xing, E. P. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* **9**, 1981–2014.

Aldous, D. (1985). Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII–1983* **1117**, 1–198.

Ashford, J. W., Rosen, A., Adamson, M., Bayley, P., Sabri, O., Furst, A. & Black, S. E. (2011a). *Handbook of Imaging the Alzheimer Brain*. IOS Press.

Ashford, J. W., Salehi, A., Furst, A., Bayley, P., Frisoni, G. B., Jack Jr, C. R., Sabri, O., Adamson, M. M., Coburn, K. L. & Olichney, J. (2011b). Imaging the Alzheimer brain. *Journal of Alzheimer's Disease* **26**, 1–27.

Assent, I. (2012). Clustering high dimensional data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **2**, 340–350.

Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lyzinski, V. & Qin, Y. (2017). Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research* **18**, 8393–8484.

Bellman, R. E. (1961). *Adaptive control processes: a guided tour*. Princeton University Press.

Bertoletti, M., Friel, N. & Rastelli, R. (2015). Choosing the number of clusters in a finite mixture model using an exact integrated completed likelihood criterion. *Metron* **73**, 177–199.

Bhattacharya, A. & Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98**, 291–306.

Bianconi, G., Pin, P. & Marsili, M. (2009). Assessing the relevance of node features for network structure. *Proceedings of the National Academy of Sciences* **106**, 11433–11438.

Bickel, P. J. & Levina, E. (2008). Regularized estimation of large covariance matrices. *Annals of Statistics* **36**, 199–227.

Biernacki, C., Celeux, G. & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 719–725.

Bishop, C. M. (1999). Bayesian PCA. In *Advances in Neural Information Processing Systems*.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bissiri, P. G. & Ongaro, A. (2014). On the topological support of species sampling priors. *Electronic Journal of Statistics* **8**, 861–882.

Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *Journal of the American Statistical Association* **112**, 859–877.

Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics* **1**, 1–12.

Bouveyron, C., Girard, S. & Schmid, C. (2007). High-dimensional data clustering. *Computational Statistics & Data Analysis* **52**, 502–519.

Briatte, F. (2016). Network patterns of legislative collaboration in twenty parliaments. *Network Science* **4**, 266–271.

Canale, A., Durante, D. & Dunson, D. B. (2018). Convex mixture regression for quantitative risk assessment. *Biometrics* **74**, 1331–1340.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q. & West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association* **103**, 1438–1456.

Carvalho, C. M., Polson, N. G. & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480.

Choi, D. S., Wolfe, P. J. & Airoldi, E. M. (2012). Stochastic blockmodels with a growing number of classes. *Biometrika* **99**, 273–284.

Cranmer, S. J. & Desmarais, B. A. (2011). Inferential network analysis with exponential random graph models. *Political analysis* **19**, 66–86.

Crossley, N. A., Mechelli, A., Vértes, P. E., Winton-Brown, T. T., Patel, A. X., Ginestet, C. E., McGuire, P. & Bullmore, E. T. (2013). Cognitive relevance of the community structure of the human brain functional coactivation network. *Proceedings of the National Academy of Sciences* **110**, 11583–11588.

Daianu, M., Jahanshad, N., Nir, T. M., Toga, A. W., Jack Jr, C. R., Weiner, M. W. & Thompson, P. M. (2013). Breakdown of brain connectivity between normal aging and Alzheimer's disease: a structural k–core network analysis. *Brain Connectivity* **3**, 407–422.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I. & Ruggiero, M. (2013a). Are Gibbs–type priors the most natural generalization of the Dirichlet process? *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**, 212–229.

De Blasi, P., Lijoi, A. & Prünster, I. (2013b). An asymptotic analysis of a class of discrete nonparametric priors. *Statistica Sinica* **23**, 1299–1321.

De Blasi, P., Martínez, A. F., Mena, R. H. & Prünster, I. (2020). On the inferential implications of decreasing weight structures in mixture models. *Computational Statistics & Data Analysis* **147**, 106940.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P. & Hyman, B. T. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968–980.

Dunson, D. B. & Herring, A. H. (2005). Bayesian latent variable models for mixed discrete outcomes. *Biostatistics* **6**, 11–25.

Durante, D. (2017). A note on the multiplicative gamma process. *Statistics & Probability Letters* **122**, 198–204.

Durante, D. & Dunson, D. B. (2014). Nonparametric Bayes dynamic modelling of relational data. *Biometrika* **101**, 883–898.

Durante, D. & Dunson, D. B. (2018). Bayesian inference and testing of group differences in brain networks. *Bayesian Analysis* **13**, 29–58.

Durante, D., Dunson, D. B. & Vogelstein, J. T. (2017a). Nonparametric Bayes modeling of populations of networks. *Journal of the American Statistical Association* **112**, 1516–1530.

Durante, D., Mukherjee, N. & Steorts, R. C. (2017b). Bayesian learning of dynamic multilayer networks. *Journal of Machine Learning Research* **18**, 1414–1442.

Faskowitz, J., Yan, X., Zuo, X.-N. & Sporns, O. (2018). Weighted stochastic block models of the human connectome across the life span. *Scientific Reports* **8**, 1–16.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports* **486**, 75–174.

Fortunato, S. & Hric, D. (2016). Community detection in networks: a user guide. *Physics Reports* **659**, 1–44.

Fowler, J. H. (2006). Connecting the congress: a study of cosponsorship networks. *Political Analysis* **14**, 456–487.

Fraley, C. & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal* **41**, 578–588.

Gelman, A. & Meng, X.-L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science* **13**, 163–185.

Geng, J., Bhattacharya, A. & Pati, D. (2019). Probabilistic community detection with unknown number of communities. *Journal of the American Statistical Association* **114**, 893–905.

Gershman, S. J. & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology* **56**, 1–12.

Ghosh, J. & Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics* **18**, 306–320.

Ghosh, P., Pati, D. & Bhattacharya, A. (2019). Posterior contraction rates for stochastic block models. *Sankhya A* , 1–29.

Girvan, M. & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* **99**, 7821–7826.

Gnedin, A. (2010). Species sampling model with finitely many types. *Electronic Communications in Probability* **15**, 79–88.

Gnedin, A. & Pitman, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zapiski Nauchnykh Seminarov, POMI* **325**, 83–102.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61**, 215–231.

Gopalan, P., Hofman, J. M. & Blei, D. M. (2015). Scalable recommendation with hierarchical Poisson factorization. In *UAI*.

Gopalan, P., Ruiz, F. J., Ranganath, R. & Blei, D. (2014). Bayesian nonparametric Poisson factorization for recommendation systems. *Journal of Machine Learning Research, Workshop and Conference Proceedings* **33**, 275–283.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Guimera, R., Mossa, S., Turtschi, A. & Amaral, L. N. (2005). The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Sciences* **102**, 7794–7799.

Handcock, M. S., Raftery, A. E. & Tantrum, J. M. (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A* **170**, 301–354.

Hartigan, J. (1990). Partition models. *Communications in Statistics - Theory and Methods* **19**, 2745–2756.

Hoff, P. (2008). Modeling homophily and stochastic equivalence in symmetric relational data. In *Advances in Neural Information Processing Systems*.

Hoff, P. D., Raftery, A. E. & Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association* **97**, 1090–1098.

Holland, P. W., Laskey, K. B. & Leinhardt, S. (1983). Stochastic blockmodels: first steps. *Social Networks* **5**, 109–137.

Ishwaran, H. & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

Ishwaran, H. & Rao, J. S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics* **33**, 730–773.

John, M., Ikuta, T. & Ferbinteanu, J. (2017). Graph analysis of structural brain networks in Alzheimer's disease: beyond small world properties. *Brain Structure and Function* **222**, 923–942.

Kang, X., Herron, T. J., Cate, A. D., Yund, E. W. & Woods, D. L. (2012). Hemispherically–unified surface maps of human cerebral cortex: reliability and hemispheric asymmetries. *PLoS One* **7**, 1–15.

Karrer, B. & Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E* **83**, 016107.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T. & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*.

Knowles, D. & Ghahramani, Z. (2011). Nonparametric Bayesian sparse factor models with application to gene expression modeling. *Annals of Applied Statistics* **5**, 1534–1552.

Lazarsfeld, P. F. & Henry, N. W. (1968). *Latent structure analysis*. Houghton Mifflin Co.

Lee, C. & Wilkinson, D. J. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science* **4**, 1–50.

Lee, J. A. & Verleysen, M. (2007). *Nonlinear dimensionality reduction*. Springer Science & Business Media.

Lenk, P. (2009). Simulation pseudo–bias correction to the harmonic mean estimator of integrated likelihoods. *Journal of Computational and Graphical Statistics* **18**, 941–960.

Lijoi, A., Mena, R. H. & Prünster, I. (2007). Controlling the reinforcement in Bayesian non-parametric mixture models. *Journal of the Royal Statistical Society: Series B* **69**, 715–740.

Lopes, H. F. & West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41–68.

Mariadassou, M., Robin, S. & Vacher, C. (2010). Uncovering latent structure in valued graphs: a variational approach. *Annals of Applied Statistics* **4**, 715–742.

Mårtensson, G., Pereira, J. B., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Lovestone, S., Simmons, A. & Volpe, G. (2018). Stability of graph theoretical measures in structural brain networks in Alzheimer's disease. *Scientific Reports* **8**, 1–15.

Meilă, M. (2007). Comparing clusterings — an information based distance. *Journal of Multivariate Analysis* **98**, 873–895.

Miller, J. W. & Harrison, M. T. (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *Journal of Machine Learning Research* **15**, 3333–3370.

Miller, J. W. & Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association* **113**, 340–356.

Müller, P., Quintana, F. & Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics* **20**, 260–278.

Murphy, K., Viroli, C. & Gormley, I. C. (2020). Infinite mixtures of infinite factor analysers. *Bayesian Analysis* **15**, 937–963.

Neal, R. M. (1992). Bayesian mixture modeling. In *Maximum Entropy and Bayesian Methods*. Springer, pp. 197–211.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103**, 8577–8582.

Newman, M. E. & Clauset, A. (2016). Structure and inference in annotated networks. *Nature Communications* **7**, 1–11.

Newman, M. E. J. & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E* **69**, 1–15.

Newton, M. A. & Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B* **56**, 3–26.

Nowicki, K. & Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association* **96**, 1077–1087.

Olhede, S. C. & Wolfe, P. J. (2014). Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences* **111**, 14722–14727.

Pajor, A. (2017). Estimating the marginal likelihood using the arithmetic mean identity. *Bayesian Analysis* **12**, 261–287.

Park, A. J.-h. & Dunson, D. B. (2010). Bayesian generalized product partition model. *Statistica Sinica* **20**, 1203–1226.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559–572.

Peel, L., Larremore, D. B. & Clauset, A. (2017). The ground truth about metadata and community detection in networks. *Science Advances* **3**, 1–8.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. *Statistics, Probability and Game Theory* **30**, 245–267.

Priebe, C. E., Park, Y., Vogelstein, J. T., Conroy, J. M., Lyzinski, V., Tang, M., Athreya, A., Cape, J. & Bridgeford, E. (2019). On a two-truths phenomenon in spectral graph clustering. *Proceedings of the National Academy of Sciences* **116**, 5995–6000.

Quintana, F. A. & Iglesias, P. L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B* **65**, 557—-574.

Raftery, A. E., Newton, M. A., Satagopan, J. M. & Krivitsky, P. N. (2007). Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics* **8**, 1–45.

Rastelli, R., Latouche, P. & Friel, N. (2018). Choosing the number of groups in a latent stochastic blockmodel for dynamic networks. *Network Science* **6**, 469–493.

Roberts, G. O. & Rosenthal, J. S. (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability* **44**, 458–475.

Rousseau, J. & Mengersen, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B* **73**, 689–710.

Roweis, S. T. & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–2326.

Roy, S., Atchadé, Y. & Michailidis, G. (2019). Likelihood inference for large scale stochastic blockmodels with covariates based on a divide-and-conquer parallelizable algorithm with communication. *Journal of Computational and Graphical Statistics* **28**, 609–619.

Ryan, C., Wyse, J. & Friel, N. (2017). Bayesian model selection for the latent position cluster model for social networks. *Network Science* **5**, 70–91.

Schmidt, M. N. & Morup, M. (2013). Nonparametric Bayesian modeling of complex networks: an introduction. *IEEE Signal Processing Magazine* **30**, 110–128.

Schölkopf, B., Smola, A. & Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation* **10**, 1299–1319.

Schwartz, L. (1965). On Bayes procedures. *Probability Theory and Related Fields* **4**, 10–26.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Signorelli, M. & Wit, E. C. (2018). A penalized inference approach to stochastic block modelling of community structure in the Italian Parliament. *Journal of the Royal Statistical Society: Series C* **67**, 355–369.

Spearman, C. (1904). "General intelligence", objectively determined and measured. *American Journal of Psychology* **15**, 201–292.

Sporns, O. (2013). Structure and function of complex brain networks. *Dialogues in Clinical Neuroscience* **15**, 247–262.

Stam, C. J. (2014). Modern network science of neurological disorders. *Nature Reviews Neuroscience* **15**, 683–695.

Stanley, N., Bonacci, T., Kwitt, R., Niethammer, M. & Mucha, P. J. (2019). Stochastic block models with multiple continuous attributes. *Applied Network Science* **4**, 1–22.

Sulaimany, S., Khansari, M., Zarrineh, P., Daianu, M., Jahanshad, N., Thompson, P. M. & Masoudi-Nejad, A. (2017). Predicting brain network changes in Alzheimer's disease with link prediction algorithms. *Molecular BioSystems* **13**, 725–735.

Sweet, T. M. (2015). Incorporating covariates into stochastic blockmodels. *Journal of Educational and Behavioral Statistics* **40**, 635–664.

Tadesse, M. G., Sha, N. & Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association* **100**, 602–617.

Tallberg, C. (2004). A Bayesian approach to modeling stochastic blockstructures with covariates. *Journal of Mathematical Sociology* **29**, 1–23.

Tenenbaum, J. B., De Silva, V. & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* **58**, 267–288.

Van Der Maaten, L., Postma, E. & Van den Herik, J. (2009). Dimensionality reduction: a comparative. *Journal of Machine Learning Research* **10**, 13.

van der Pas, S. & van der Vaart (2018). Bayesian community detection. *Bayesian Analysis* **13**, 767–796.

Wade, S. & Ghahramani, Z. (2018). Bayesian cluster analysis: point estimation and credible balls. *Bayesian Analysis* **13**, 559–626.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian Statistics* **7**, 733–742.

White, A. & Murphy, T. B. (2016). Mixed–membership of experts stochastic blockmodel. *Network Science* **4**, 48–80.

Young, S. J. & Scheinerman, E. R. (2007). Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*.

Zhao, Y., Levina, E. & Zhu, J. (2011). Community extraction for social networks. *Proceedings of the National Academy of Sciences* **108**, 7321–7326.

Zhao, Y., Levina, E. & Zhu, J. (2012). Consistency of community detection in networks under degree–corrected stochastic block models. *Annals of Statistics* **40**, 2266–2292.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* **67**, 301–320.