Submitted to Statistical Science

Exchangeability, prediction and predictive modeling in Bayesian statistics

Sandra Fortini and Sonia Petrone

Abstract. There is currently a renewed interest in the Bayesian *predictive* approach to statistics. This paper offers a review on foundational concepts and focuses on 'predictive modeling', which by directly reasoning on prediction, bypasses inferential models or may characterize them. We detail predictive characterizations in exchangeable and partially exchangeable settings, for a large variety of data structures, and hint at new directions. The underlying concept is that Bayesian predictive rules are probabilistic *learning* rules, formalizing through conditional probability how we learn on future events given the available information. This concept has implications in any statistical problem and in inference, from classic contexts to less explored challenges, such as providing Bayesian uncertainty quantification to predictive algorithms in data science, as we show in the last part of the paper. The paper gives a historical overview, but also includes a few new results, presents some recent developments and poses some open questions.

Key words and phrases: Bayesian foundations, Predictive characterizations, Bayesian nonparametrics, Predictive sufficiency, Partial exchangeability, Recursive algorithms.

1. INTRODUCTION

There is currently a renewed interest in the *Bayesian predictive approach* to statistics. The approach is just Bayesian, but the additional adjective 'predictive' underlines conceptual emphasis on predictive tasks; while the more common 'inferential approach' is centered on inference on parameters, here one focuses on observable quantities and prediction, evaluates models and priors based on their implications on prediction, and even deduce models and parameters from the predictive rule (the long list of references includes [34], [62], [63], [24]). With the major focus on prediction in data science and machine learning ([18], [116]), this approach appears natural and is adopted in novel research directions ([66], [56], [50], [129] [12], [90]). In fact, the predictive approach has a long tradition in Bayesian statistics and is rooted in its same foundations (de Finetti [29], [30], [23], Savage [114], [45], and Diaconis [35], [36], Regazzini [109], [52], Dawid [27] and more; see the book by Bernardo and Smith [10]).

Sandra Fortini is Associate Professor of Statistics at the Department of Decision Sciences, Bocconi University (e-mail: sandra.fortini@unibocconi.it). Sonia Petrone is Professor of Statistics, Department of Decision Sciences and Bocconi Institute of Data Science and Analytics, Bocconi University (e-mail: sonia.petrone@unibocconi.it).

The first aim of this paper is to offer a review, starting from foundations and going through methods for predictive constructions in a variety of contexts, with focus on exchangeable structures, which play a basic role. Thus we also review, from a predictive perspective, the use of exchangeability and of forms of partial exchangeability in Bayesian statistics.

A second aim of the paper is to show how a Bayesian predictive approach can be usefully adopted in less explored situations, beyond exchangeability; in particular, (a) to obtain computationally tractable approximations of (exchangeable) Bayesian inferences and (b) to provide Bayesian uncertainty quantification of some classes of algorithms (a novel example we provide is online gradient descent), without the need of an explicit likelihood and prior law. This is developed in the last part of the paper and relies on the foundational principles that we discuss in the first part.

Along our review, we include a few novel results and open problems. We hope that the paper may be of some interest, especially to young researchers, as both a reminder of the foundations and of some remarkable results, and as an inspiration for new work.

1.1 Basic concepts and paper overview

In Bayesian statistics, prediction is expressed through the *predictive distribution* of future observations given

the available information. In the simplest setting (and with an abuse of notation, in this introduction identifying distributions through their arguments) one has a sample from a sequence of random variables (r.v.'s) $(X_n)_{n\geq 1}$, has specified a conditional model $(X_1,\dots,X_n) \mid \tilde{\theta} \sim p(x_1,\dots,x_n \mid \tilde{\theta})$, $n \geq 1$ and a prior distribution π on $\tilde{\theta}$, and computes the predictive density of X_{n+1} given $x_{1:n} \equiv (X_1 = x_1,\dots,X_n = x_n)$ as

$$(1.1)\quad p(x_{n+1} \mid x_{1:n}) = \int_\Theta p(x_{n+1} \mid x_{1:n},\theta) d\pi(\theta \mid x_{1:n}),$$

where $\pi(\cdot \mid x_{1:n})$ is the posterior distribution of $\tilde{\theta}$ (we use the notation $\tilde{\theta}$ to underline that it is a r.v.). Summaries of the predictive distribution naturally include point prediction and predictive credible intervals. Thus, while standard frequentist prediction would move from a model $(X_1,\dots,X_n) \sim p_\theta(x_1,\dots,x_n)$ and deal with parameters' uncertainty by plugging their estimates into $p_\theta(x_{n+1} \mid x_{1:n})$, in the Bayesian approach uncertainty is taken into account by 'averaging' the possible models $p(x_{n+1} \mid x_{1:n},\theta)$ with respect to the posterior distribution of $\tilde{\theta}$.

We already see distinctive features of Bayesian prediction; but this all may sound as 'the usual Bayesian story'. Actually, Bayesian statistics is often described as consisting of assigning a prior on $\tilde{\theta}$ and using Bayes rule to compute the posterior distribution. Obtaining the predictive distribution as in (1.1) is then just a matter of computations. Of course, Bayesian statistics is deeper than that; and a first basic concept we should recall for this paper is the interpretation of the Bayesian predictive distribution.

Bayesian statistics is about acting under uncertainty, or incomplete *information*. This can be information from the data, from domain knowledge, etc; the point is to formalize that information, and probability is the prescribed formal language for this. If probability describes (incomplete) information, then the evolution of information, or *learning*, is expressed through *conditional probabilities*. In particular, learning on the next observation based on the observed $x_{1:n}$ is expressed through the conditional distribution $p(x_{n+1} \mid x_{1:n})$. This leads us to the interpretation of the Bayesian predictive distribution: it is a *learning rule* that formalizes, through conditional probability, how we learn about future events given the available information ([54]. Thus, it is not meant as the 'physical mechanism' generating X_{n+1} given the past – in the classic setting, that might be the interpretation of $p_{\theta_0}(x_{n+1} \mid x_{1:n})$ for a true θ_0).

This principle is the basis of our discussion in the paper, and we return on it in a rather novel way in Section 5. Here, to see a first implication, let us consider the basic case, random sampling. In the Bayesian approach, one does not assume independence, as it would give $p(x_{n+1} \mid x_{1:n}) = p(x_{n+1})$, expressing no learning. One would rather elicit a joint probability $p(x_1,\dots,x_n)$ that expresses dependence: not because the X_i are 'physically' dependent, but because each X_i brings information about the others. The X_i are dependent in our probability assessment formalizing the learning process. In random sampling, the natural assessment is that the order of the observations does not bring any information: the X_i are exchangeable. Then they are only *conditionally* independent. We devote substantial space in the paper to exchangeability; simply because it is the natural predictive requirement in random sampling, and random sampling is the basic setting. The fundamental concepts are treated in Section 2.

In practice, we usually specify the joint distribution $p(x_1,\dots,x_n)$, for any n, with the help of models and parameters

$$(1.2)\qquad p(x_1,\dots,x_n) = \int p(x_1,\dots,x_n \mid \theta) d\pi(\theta);$$

and compute the predictive distribution as in (1.1). But, especially if interest is in prediction, we could in principle bypass the inferential model and directly specify the predictive distributions - typically, the one-step-ahead predictions, which give, for any n,

(1.3)

$$p(x_1,\dots,x_n) = p(x_1)p(x_2 \mid x_1)\cdots p(x_n \mid x_{1:n-1}).$$

In this *predictive approach*, that we refer to as "predictive modeling", one reasons on the observable quantities, for example on symmetry properties as in the case of exchangeability, and on what information from the sample is relevant for prediction. This is well rooted in Bayesian foundations and is particularly attractive in complex settings where models and parameters tend to lose interpretability. Still, predictive modeling may seem quite impracticable; it has in fact a long tradition, however the available literature is rather fragmented. Thus in Section 3 our effort is to trace concepts and methods that may provide a methodological basis to predictive constructions. We mostly refer to exchangeable settings, but a predictive approach can be taken for any kind of data structures (see e.g. [12]).

Prediction and inference. Predictive modeling is also intriguing as a form of "Bayesian learning without the prior". In fact, an inferential model and a prior law may be implicitly subintended, and unveiling them is important both practically and conceptually. This is typically obtained through representation theorems; roughly speaking, one can move from the predictive specification (1.3) of the joint distribution $p(x_1,\dots,x_n)$, for any $n \geq 1$, and might *represent* it in a form as (1.2); see Section 2.1. Although an inferential model is not needed in a purely predictive approach, representation theorems significantly provide the link from prediction to inference. The celebrated de Finetti's representation theorem has a central

role in Bayesian statistics as it leads from foundations, where probability is expressed on *observable* events (see Section 2), to inference. From an assumption on the observable X_i (exchangeability), the representation theorem gives the theoretical justification of the basic Bayesian inferential scheme where the parameter $\tilde{\theta}$ is random and the X_i are conditionally i.i.d. given $\tilde{\theta}$, as an implication of exchangeability. Moreover, it shows how the inferential model is related to frequencies. In Section 2.2, we will underline how *prediction* is related to frequencies, thus to the inferential model, and in particular we give a result (Section 2.4) showing how the uncertainty expressed in the posterior distribution is determined by the way the predictive distribution learns from the data.

Representation theorems have been extended in numerous directions (Sect 2.3 of [23] includes extensive references) and predictive constructions are applied well beyond simple random sampling. In Section 4 we consider more structured data for which it is natural to express a predictive judgment of *partial* exchangeability; we provide predictive characterizations of some forms of partial exchangeability (Theorems 2.3, 4.4 and 4.7), and review de Finetti-like representation theorems, which give the predictive-theoretical basis in many problems including stochastic design regression (as reducible to random sampling), fixed design regression and multiple experiments (Section 4.1), Markov chains (potentially, models for temporal data based on Markov chains) (Section 4.3) and arrays and networks data (Section 4.4). There are authoritative and comprehensive references on the theory of exchangeability, see Kingman [80], Aldous [2], Kallenberg [78], to which we refer interested readers. The more specific aim of our - necessarily brief - review is to point out some main aspects that we believe are relevant in Bayesian statistics from a predictive perspective.

Open directions. Although the above discussion shows that the predictive approach is theoretically sound and that predictive modeling can be applied in many contexts, we acknowledge that proceeding solely through predictive constructions may not be easy, especially if one wants to satisfy exchangeability constraints. On the other hand – and this is a further point we want to make in this paper – there are many predictive algorithms in data science that lack clean uncertainty quantification, or there are, in fields such as economics, subjective predictions implicitly guided by the agent's explanation of the phenomena, that would be interesting to reveal (see e.g. [5]). A Bayesian predictive approach can be usefully employed. In particular, we show that some classes of recursive predictive algorithms can in fact be read as Bayesian predictive learning rules, that assume exchangeability only asymptotically. The relevance of this approach is not merely theoretical, but allows to understand the underlying modeling assumptions and to provide formal uncertainty quantification, and can lead to principled extensions. This is treated in Section 5. Brief final remarks conclude the paper. All the proofs are collected in the Supplement [58].

1.2 Preliminaries and notation

In this paper, all the random variables take values in a Polish space $\mathbb{X}$, endowed with its Borel sigma-algebra $\mathcal{X}$. The topology on spaces of probability measures is implicitly assumed as the topology of weak convergence. Hence, for any P_n and P, $P_n \to P$ means weak convergence.

The underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ for a random sequence $(X_n)_{n\geq 1}$ is implicitly assumed to be the canonical space $(\mathbb{X}^\infty, \mathcal{X}^\infty, \mathbb{P})$, where $\mathbb{P}$ is the probability law of the sequence, denoted as $(X_n)_{n\geq 1} \sim \mathbb{P}$. We write $\mathbb{P}$-a.s. for "with $\mathbb{P}$-probability one". We use the short notation $x_{1:n}$ for $(X_1 = x_1, \ldots, X_n = x_n)$. All conditional distributions must be understood as regular versions. For random variables taking values in Euclidean spaces, we denote with the same symbol a probability measure and the corresponding distribution function. Sequences are denoted as $(Z_n)_{n\geq 1}$ and arrays as $[Z_{i,j}]_{i\in I, j\in J}$.

2. EXCHANGEABILITY AND PREDICTION

Let us begin by recalling in some more detail the foundational role of prediction in Bayesian statistics and the notion of exchangeability as a basic predictive judgment.

Bayesian statistics has decision-theoretic roots in the work of the 1920s in mathematical logic aimed at founding a normative theory of rational decisions under risk (Ramsey [108], and later, Savage [114], [45]; two book references are [10] and [98]). In this perspective, probability arises as the prescribed rational (*coherent*; see [23]) formalization of the agent's information on uncertain events, as advocated in the foundations of modern Bayesian statistics by Bruno de Finetti; see e.g. [30] and [33]. de Finetti emphasises that probability is expressed on *observable* events (we do not discuss, here, issues on the notions of observability or of imprecise probability; see e.g. [128]). In this perspective, unobservable parameters are not assigned a probability *per se*, but simply as an intermediate step for ultimately expressing the probability of observable events. They are just a tool in the learning process that goes from past observable events to prediction of future events. Of course, parameters may be interpretable and inference is a core problem, but it is prediction that has a foundational role.

The focus on probability of observable events is well demonstrated in de Finetti's notion and use of exchangeability. As mentioned in the Introduction, in the context of homogeneous replicates of an experiment (random sampling) the researcher would judge that the labels of the X_i "do not matter". This is formalized through a joint probability law that is invariant under permutations of the labels:

$$(X_1, \ldots, X_n) \stackrel{d}{=} (X_{\sigma(1)}, \ldots, X_{\sigma(n)})$$

for each permutation σ of $(1,\ldots,n)$, where $\stackrel{d}{=}$ means equal in distribution. An infinite sequence $(X_n)_{n\geq 1}$ is exchangeable if it is invariant to each finite permutation of $\{1,2,\ldots\}$, i.e. each permutation that only switches a finite set of indexes. Exchangeability is an elegant probabilistic structure and exchangeable processes arise in many fields. In de Finetti's work on Bayesian foundations, however, exchangeability is not meant as a physical property of the sequence $(X_n)_{n\geq 1}$, but as an expression of the agent's information.

EXAMPLE 2.1. Consider random sampling from a two-color urn, and let $X_i = 1$ if the color of the ball picked on the i-th draw is white, and zero otherwise. The agent judges that the order of the draws is not informative and the sequence $(X_n)_{n\geq 1}$ is exchangeable. By the representation theorem (Section 2.1), $(X_n)_{n\geq 1}$ has the same probability law of a sequence arising from an experiment where the urn composition is picked from a 'prior' distribution and balls are then sampled at random with replacement. The physical experiment is not as such: the urn composition is not sampled, it is given although unknown. Here, exchangeability is not referring to the mechanism generating the data, but to the way we use information. □

We should keep in mind this use of exchangeability in what follows. See also [57], and the discussion in [123] for the more general setting of stationary sequences.

Although exchangeability is a predictive requirement, it has an immediate inferential implication, established by the celebrated de Finetti's representation theorem.

THEOREM 2.2 (Law of large numbers and representation theorem for infinite exchangeable sequences). *Let $(X_n)_{n\geq 1}$ be an infinite exchangeable sequence and denote by $\mathbb{P}$ its probability law. Then:*

i) *With $\mathbb{P}$-probability one, the sequence of the empirical distributions $\hat{F}_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ converges weakly as $n\to\infty$ to a random distribution $\tilde{F}$,*

$$\hat{F}_n \to \tilde{F};$$

ii) *For all $n\geq 1$ and measurable sets $A_1,\ldots,A_n$,*

$$\mathbb{P}(X_1\in A_1,\ldots,X_n\in A_n) = \int \prod_{i=1}^{n} F(A_i)\, d\pi(F), \tag{2.1}$$

where π is the probability law of $\tilde{F}$.

See Aldous [2], who refers to $\tilde{F}$ as the *directing random measure* of the exchangeable sequence $(X_n)_{n\geq 1}$. The representation 2.1 is often phrased as "$X_i \mid \tilde{F} = F \stackrel{i.i.d.}{\sim} F$, with $\tilde{F}\sim\pi$"; a subtle difference is that this latter formulation may (in principle, misleadingly) suggest the existence of a true F. In Bayesian inference, $\tilde{F}$ plays the role of the statistical model, and its probability law is the prior. The prior law is unique, and is a probability measure on the class of all the possible distributions on the sample space. The representation theorem is a high-level result: the probability law $\mathbb{P}$ characterizes the random $\tilde{F}$; in other words, it shapes it (the model) through its implied distribution (the prior). In applications, one has to choose a specific law $\mathbb{P}$. In particular, further information may restrict the support of the prior to a parametric class, so that $X_i \mid \tilde{\theta} \stackrel{i.i.d.}{\sim} p(\cdot \mid \tilde{\theta})$ (see Section 3.2). In this paper we will mostly keep the general framework (2.1).

Remark. Note that $\tilde{F}$ in Theorem 2.2 is random; as the limit of the empirical distributions, it depends on $(X_1, X_2, \ldots)$. Given a sample path $\omega = (x_1, x_2, \ldots)$, we have a realization of the random $\tilde{F}$, that we denote by $\tilde{F}(\cdot)(\omega)$. For i.i.d. observations from a distribution F, the limit of the empirical distribution is F; the fact that the limit is random for exchangeable sequences may sound surprising. Formally, this is because exchangeable sequences are *mixtures of i.i.d. sequences*; let us give some intuition. By the representation theorem, an exchangeable sequence $(X_n)_{n\geq 1}$ can be obtained by first picking a distribution F from the prior law, then sampling the X_i at random from F. If we pick F and restrict ourselves to the set of the sample paths $\omega = (x_1, x_2, ...)$ that may be obtained by sampling at random from it, we have the usual properties of the i.i.d. case; in particular, for almost all these ω, the empirical distribution converges to F, which is not random. However, when we observe a finite sample $(x_1,\ldots,x_n)$, we do not know what F was chosen, hence the limit of the empirical distribution may still be any distribution we could have picked from the prior. We would know which one if we could observe the entire $\omega = (x_1, x_2, \ldots)$, and thus see the limit, that is the realization $\tilde{F}(\cdot)(\omega)$ of the random $\tilde{F}$.

2.1 Predictive characterization of exchangeability

The representation theorem allows us to specify an exchangeable probability law through the usual inferential scheme. In a predictive approach, however, we would avoid models and priors and directly specify it through the predictive rule. This is the core of predictive modeling, beyond exchangeability.

For any probability law $\mathbb{P}$ for the sequence $(X_n)_{n\geq 1}$, define the *predictive rule* as the sequence of predictive distributions $P_0(\cdot) \equiv \mathbb{P}(X_1 \in \cdot)$ and, for $n \geq 1$,

$$P_n(\cdot) \equiv \mathbb{P}(X_{n+1} \in \cdot \mid X_1, \ldots, X_n), \tag{2.2}$$

and let us denote by $P_n(\cdot \mid x_{1:n})$ its realization for $x_{1:n}$. In particular, if $\mathbb{P}$ is exchangeable, the predictive rule is obtained as $P_0(\cdot) = E(\tilde{F}(\cdot))$ and $P_n(\cdot) = E(\tilde{F}(\cdot)|X_1,\ldots,X_n)$ for $n \geq 1$.

In predictive modeling, one moves from the predictive rule to specify the probability law of the process

$(X_n)_{n\geq 1}$. More formally, one can assign a sequence $(P_n)_{n\geq 0}$ of probability kernels (or a *strategy*, [45], [12]). Then by Ionescu-Tulcea theorem (see [77] Theorem 5.17 and Corollary 5.18) there exists a unique probability law $\mathbb{P}$ for a process $(X_n)_{n\geq 1}$ such that $X_1 \sim P_0$ and for every $n \geq 1$, P_n is the conditional distribution of X_{n+1} given $(X_1, \ldots, X_n)$. (Given this equivalence, we will use the notation $(P_n)_{n\geq 0}$ to represent both the sequence of the predictive distributions from a given $\mathbb{P}$, and a strategy).

Thus, the probability law $\mathbb{P}$ of a process $(X_n)_{n\geq 1}$ is uniquely defined (characterized) by the sequence of predictive distributions $(P_n)_{n\geq 0}$. A natural question is under what conditions on the P_n one obtains an *exchangeable* law $\mathbb{P}$. This problem has been addressed in [52].

THEOREM 2.3 ([52], Proposition 3.2 and Theorem 3.1). *Let $(X_n)_{n\geq 1} \sim \mathbb{P}$ be an infinite sequence of r.v.'s, with predictive rule $(P_n)_{n\geq 0}$. Then $(X_n)_{n\geq 1}$ is exchangeable if and only if, for every $n \geq 0$, the following conditions hold, with $P_0(\cdot \mid x_{1:0})$ meant as $P_0(\cdot)$,*

i) *For every A, $P_n(A \mid x_{1:n})$ is a symmetric function of $x_1, \ldots, x_n$;*
ii) *The set function $(A, B) \to \int_A P_{n+1}(B \mid x_{1:n+1}) dP_n(x_{n+1} \mid x_{1:n})$ is symmetric in A and B.*

Condition i) requires that, for every $n \geq 1$, the predictive distribution of X_{n+1} is a function of the empirical distribution of $(x_1, \ldots, x_n)$; which is a necessary condition for exchangeability. As well, given $x_{1:n}$, the predictive distribution of $(X_{n+1}, \ldots, X_{n+k})$ should be invariant under permutations of the k future observations, since under exchangeability the joint distribution of $(X_1, \ldots, X_{n+k})$ is symmetric. Condition ii) only asks that the next $k = 2$ observations can be permuted.

2.2 Prediction, frequency, models

Although there are no formal constraints in assigning a predictive rule $(P_n)_{n\geq 0}$, we aim for our predictions to be consistent with facts. For exchangeable sequences, the following property relates prediction to frequency.

PROPOSITION 2.4. *Let $(X_n)_{n\geq 1} \sim \mathbb{P}$ be an exchangeable sequence, with predictive rule $(P_n)_{n\geq 0}$. Then, with probability one, for $n \to \infty$ the sequence of predictive distributions converges, and its limit coincides with the limit of the empirical distributions:*

$$P_n \to \tilde{F}, \quad \mathbb{P}\text{-a.s.}, \tag{2.3}$$

with $\tilde{F}$ as in Theorem 2.2.

A proof is given in [2], Lemma 8.2 page 61. In fact, the result remains valid under the less restrictive condition that $(P_n(A))_{n\geq 0}$ is a martingale for every A, without the need for $(X_n)_{n\geq 1}$ to be exchangeable [15]. We return on this point in more details in Section 2.3.

EXAMPLE 2.5. Consider an exchangeable sequence $(X_n)_{n\geq 1}$ with $X_i \in \{1, \ldots, k\}$. Then the empirical distribution is characterized by the vector of relative frequencies n_j/n, and any predictive distribution must be a function of $(n_1, \ldots, n_k)$, i.e. $p_n(j) \equiv \mathbb{P}(X_{n+1} = j \mid x_{1:n}) = \mathbb{P}(X_{n+1} = j \mid n_1, \ldots, n_k), j = 1, \ldots, k$. For any j, with probability one the relative frequency n_j/n and the predictive probability $p_n(j)$ converge to the same random limit $\tilde{p}_j$. The statistical model is a discrete distribution on $\{1, \ldots, k\}$ with masses $(\tilde{p}_1, \ldots, \tilde{p}_k)$ and the prior is the probability law of the random limit $(\tilde{p}_1, \ldots, \tilde{p}_k)$. □

In Bayesian statistics, Proposition 2.4 ensures that, with probability one, our predictions will adjust to frequencies; in other words, the predictive distribution P_n and the empirical distribution $\hat{F}_n$ will be close. Several refinements of this property are available, as well as quantitative bounds ([43], and references therein; see also [38]).

Proposition 2.4 also shows that, for exchangeable sequences, the statistical model is the limit of the predictive distribution; that is also the limit of the empirical distribution. Hence, the uncertainty on the model at a finite n is uncertainty on their common limit. It is this uncertainty that is expressed by the posterior distribution of $\tilde{F}$, as we will illustrate in Section 2.4, expanding from [56]. This is also the basic principle that underlines the interpretation of uncertainty in terms of "missing observations" in [50].

Remark. de Finetti proved the convergence property (2.3) for exchangeable binary sequences $(X_n)_{n\geq 1}$, and it is interesting to note that he used this result to give an explanation in terms of prediction of the frequentist viewpoint on probability [30]. He considers replicates of an experiment with binary outcome where a frequentist researcher assumes that $\mathbb{P}(X_{n+k} = 1 \mid x_{1:n}) = p$ for any $k \geq 1$ and, for n large, estimates p with the relative frequency $\hat{p}_n = \sum_{i=1}^n x_i/n$. Exchangeability makes the frequentist prediction, namely $\mathbb{P}(X_{n+k} = 1 \mid x_{1:n}) \simeq \hat{p}_n$, "permissible", by the result (2.3); see [23], Sect 2.3.

2.3 Asymptotic exchangeability.

A natural question is whether there is a reverse implication of Proposition 2.4. Exchangeability of $(X_n)_{n\geq 1}$ implies that $P_n \to \tilde{F}$, $\mathbb{P}$–a.s., but convergence of $(P_n)_{n\geq 0}$ to a random probability measure does not imply exchangeability. However, it does so asymptotically. A sequence $(X_n)_{n\geq 1}$ is asymptotically exchangeable with limit directing random measure $\tilde{F}$ (shortly, *$\tilde{F}$-asymptotically exchangeable*) if, for $n \to \infty$

$$(X_{n+1}, X_{n+2}, \ldots) \xrightarrow{d} (Z_1, Z_2, \ldots),$$

where the sequence $(Z_n)_{n\geq 1}$ is exchangeable and has directing random measure $\tilde{F}$. It can be proved that, if the sequence of predictive distributions $(P_n)_{n\geq 0}$ of $(X_n)_{n\geq 1}$

converges to a random probability measure $\tilde{F}$, then $(X_n)_{n\geq 1}$ is $\tilde{F}$-asymptotically exchangeable ([2], Lemma 8.2). Roughly speaking, for n large

$$X_n \mid \tilde{F} \overset{iid}{\approx} \tilde{F},$$

where $\tilde{F}$ has a prior law induced by the predictive rule.

Interestingly, convergence of the sequence $(P_n)_{n\geq 0}$ to a random probability measure, thus asymptotic exchangeability, holds if $(P_n)_{n\geq 0}$ is a martingale, or, equivalently, if the sequence $(X_n)_{n\geq 1}$ is *conditionally identically distributed* (c.i.d.); that is, if it satisfies

$$(X_1, \ldots, X_n, X_{n+1}) \overset{d}{=} (X_1, \ldots, X_n, X_{n+k}), \tag{2.4}$$

for all integers $k \geq 1$ and $n \geq 1$; i.e., conditionally on the past, all future observations are identically distributed. The property (2.4) was considered by Kallenberg as a weak invariance condition that, for stationary sequences, is equivalent to exchangeability ([75], Proposition 2.1). He also noted that (2.4) is equivalent to $(X_1, \ldots, X_n, X_{n+1}) \overset{d}{=} (X_1, \ldots, X_n, X_{n+2})$ for all $n \geq 1$. The term c.i.d. was introduced by Berti *et al.* [15], who proved, among other results, that the c.i.d. property is equivalent to $(P_n)_{n\geq 0}$ being a measure-valued martingale with respect to the natural filtration of $(X_n)_{n\geq 1}$. The martingale property means that the sequence of random measures $(P_n)_{n\geq 0}$ satisfies

$$E(P_{n+1}(A) \mid X_1, \ldots, X_n) = P_n(A)$$

for every $n \geq 0$ and every measurable set A (see [71]).

For exchangeable sequences, it is straightforward to show that the predictive rule is a martingale. But the martingale condition is weaker than exchangeability; still, remarkably, it is sufficient to prove the convergence result in Proposition 2.4: for a c.i.d. process $(X_n)_{n\geq 1}$, the sequence of the empirical distributions converges $\mathbb{P}$-a.s. to a random probability distribution $\tilde{F}$, and the sequence of predictive distributions not only converges (being a bounded martingale), but converges to the same limit $\tilde{F}$ ([15], Theorem 2.5).

Thus, a c.i.d. sequence is asymptotically exchangeable. However, it is not generally exchangeable. The property that is broken is stationarity: the researcher is assuming a temporal evolution in the process. It is however a specific form of evolution: marginally, the r.v.'s are identically distributed, and the process converges to a stationary - thus, together with the c.i.d. property, exchangeable - state (see also [60] and [56]). For more developments, we refer to [12]. We return to asymptotic exchangeability and c.i.d. sequences in Section 5.

2.4 Predictive-based approximations of the posterior distribution.

In the usual inferential setting, one computes the posterior distribution and obtains the predictive distribution as in expression (1.1). In predictive modeling, the order is reversed; here, from the predictive assumption of exchangeability of the X_i, we have obtained the implied inferential scheme. Can we also revert the order in expression (1.1), i.e. go from the predictive rule to the posterior distribution, and what would be the implications on inference? In this section we address this question and show two implications; namely, two predictive-based approximations of the posterior distribution.

For brevity, here we consider $X_i \in \mathbb{R}$. In the exchangeable setting, with no parametric restrictions, we have $X_i \mid \tilde{F} \overset{i.i.d.}{\sim} \tilde{F}$ and we are used to think of the prior and posterior distributions on $\tilde{F}$ as expressing uncertainty on the true distribution, say F_0. In fact, as seen in Section 2.2, what the prior and the posterior distributions are expressing is the uncertainty about the common limit $\tilde{F}$ of the empirical and the predictive distributions. If we knew the entire sample path $\omega = (x_1, x_2, \ldots)$, we would know the limit, namely $\tilde{F}(\cdot)(\omega)$, and there would be no uncertainty left. Given a finite sample $(x_1, \ldots, x_n)$, we are still uncertain about the limit, and this uncertainty is expressed through the posterior distribution. The following approximations of the posterior distribution are based on this principle.

A predictive-based simulation scheme. First, leveraging on Proposition 2.4, we can provide a predictive-based sampling scheme ([56], [57]) to approximate the prior and the posterior distribution of $\tilde{F}$; in practice we use $[\tilde{F}(t_1), ..., \tilde{F}(t_k)]$ for $t_1, ..., t_k$ in a grid of values. Assume that $P_0(\cdot) = E(\tilde{F}(\cdot))$ is continuous in $t_1, \ldots, t_k$, which implies that, $\mathbb{P}$-a.s., $\tilde{F}$ is continuous at those points so that $\lim_n \hat{F}_n(t_j) = \lim_n P_n(t_j) = \tilde{F}(t_j)$ for any t_j, $\mathbb{P}$-a.s.

In principle, given the predictive rule, one can generate $\omega = (x_1, x_2, \ldots)$ by sampling x_1 from P_0, then x_2 from $P_1(\cdot \mid x_1)$ and so on; and, having $\omega = (x_1, x_2, \ldots)$, can obtain $\tilde{F}(t_j)(\omega)$ for $j = 1, \ldots, k$; which is a sample from the prior law of $[\tilde{F}(t_1), \ldots, \tilde{F}(t_k)]$. Repeating M times gives a Monte Carlo sample of size M from the prior.

To simulate from the posterior law given $(x_1, \ldots, x_n)$, one can proceed similarly by generating the missing observations $(x_{n+1}, x_{n+2}, \ldots)$ from the predictive rule to complete ω, and repeat M times to obtain a sample of size M from the posterior. Of course, in practice, one would truncate ω to a finite sequence $(x_1, \ldots, x_N)$ with N large, and approximate $\tilde{F}(t)(\omega)$ with $P_N(t \mid x_{1:N})$, or with $\hat{F}_N(t)$, for each t in the grid.

A similar predictive principle is considered in the interesting developments by Fong, Holmes and Walker [50]

of the Bayesian bootstrap in a parametric setting: samples from a martingale posterior distribution are obtained by Doob theorem [44], after simulating future observations from a sequence of predictive distributions (see also [67]).

A predictive-based asymptotic approximation of the posterior distribution. One can also obtain a predictive-based analytic approximation of the posterior distribution for large n. By Proposition 2.4, $|\tilde{F}(t) - P_n(t)| \to 0$, $\mathbb{P}$-a.s., for any continuity point t of $\tilde{F}$, and because $(P_n(t))_{n\geq 0}$ is a martingale, one could use martingale central limit theorems to give asymptotic approximations of $(\tilde{F}(t) - P_n(t))$; yet, this would not inform on its behavior *conditionally* on the data. The following result uses a central limit theorem for martingales in terms of *almost sure convergence of conditional distributions* [26]. This type of convergence has been applied in probability for other aims and was used in a novel way in Bayesian statistics by [56] to inform on the asymptotic form of the *posterior* distribution. Here, we provide an asymptotic Gaussian approximation of the joint posterior distribution of $[\tilde{F}(t_1), \dots, \tilde{F}(t_k)]$, extending a result in [57]. Because $E(\tilde{F}(t) \mid X_1, \dots X_n) = P_n(t)$, the approximation is centered on P_n. Then we look at *how* the predictive rule learns from the data, introducing the notion of *predictive updates*. As a fresh observation becomes available, the predictive distribution is updated by incorporating the latest information, and for $n \geq 1$ and $t \in \mathbb{R}$ we denote by

$$\Delta_{t,n} = P_n(t) - P_{n-1}(t)$$

the n-th update of the predictive distribution function at the point t as X_n becomes available. For a given t, the predictive updates $\Delta_{t,n}$ eventually converge to zero, since $P_n \to \tilde{F}$, and the rate of convergence is generally of the order $1/n$, as discussed in [57]. The following proposition shows that the convergence of $\sqrt{n}(\tilde{F}(t) - P_n(t))$ depends on the asymptotic behaviour of $(n\Delta_{t,n})_{n\geq 1}$. For a grid of points $\mathbf{t} = (t_1, \dots, t_k) \in \mathbb{R}^k$, we denote by $\mathbf{\Delta}_{\mathbf{t},n} = [\Delta_{t_1,n} \ \dots \Delta_{t_k,n}]^T$ the column vector of the updates of $(P_n(t_1), \dots, P_n(t_k))$. The proposition below holds for exchangeable sequences, but more generally for sequences whose predictive rule is a martingale, i.e. for c.i.d. sequences.

PROPOSITION 2.6. *Let $(X_n)_{n\geq 1} \sim \mathbb{P}$ be a c.i.d. sequence of real-valued r.v.'s, with predictive rule $(P_n)_{n\geq 0}$, and take $\mathbf{t} = (t_1, \dots, t_k)$ such that $\mathbb{P}(X_1 \in \{t_1, \dots, t_k\}) = 0$. Suppose that the predictive updates satisfy*

$$E(\sup_n \sqrt{n}|\Delta_{t_i,n}|) < +\infty \qquad (i = 1, \dots, k),$$

$$\sum_{n=1}^{\infty} n^2 E(\Delta_{t_i,n}^4) < +\infty \qquad (i = 1, \dots, k),$$

$$E(n^2 \mathbf{\Delta}_{\mathbf{t},n} \mathbf{\Delta}_{\mathbf{t},n}^T \mid X_1, \dots, X_{n-1}) \to U_{\mathbf{t}} \qquad \mathbb{P}\text{-a.s.},$$

for a positive definite random matrix $U_{\mathbf{t}}$. Define, for every $n \geq 1$,

$$V_{n,\mathbf{t}} = \frac{1}{n} \sum_{m=1}^{n} m^2 \mathbf{\Delta}_{\mathbf{t},m} \mathbf{\Delta}_{\mathbf{t},m}^T. \tag{2.5}$$

Then, $\mathbb{P}$-a.s., $V_{n,\mathbf{t}}$ converges to $U_{\mathbf{t}}$ and

$$\sqrt{n}\, V_{n,\mathbf{t}}^{-1/2} \begin{bmatrix} \tilde{F}(t_1) - P_n(t_1) \\ \dots \\ \tilde{F}(t_k) - P_n(t_k) \end{bmatrix} \mid X_1, \dots, X_n \xrightarrow{d} \mathcal{N}_k(0, I)$$

as $n \to \infty$, where $\mathcal{N}_k(0, I)$ denotes the k-dimensional standard Gaussian distribution.

Informally, for n large,

$$\begin{bmatrix} \tilde{F}(t_1) \\ \vdots \\ \tilde{F}(t_k)) \end{bmatrix} \mid x_{1:n} \approx \mathcal{N}_k \left(\begin{bmatrix} P_n(t_1) \\ \vdots \\ P_n(t_k)) \end{bmatrix}, \frac{V_{n,\mathbf{t}}}{n} \right)$$

for $\mathbb{P}$-almost all sample paths $\omega = (x_1, x_2, \dots)$.
Proposition 2.6 allows to compute asymptotic credible sets. For example, a $(1-\alpha)$ marginal asymptotic credible interval for $\tilde{F}(t)$ given $x_{1:n}$ is

$$\left[P_n(t) - z_{1-\alpha/2}\sqrt{\frac{V_{n,t}}{n}}, P_n(t) + z_{1-\alpha/2}\sqrt{\frac{V_{n,t}}{n}} \right]$$

with $z_{1-\alpha/2}$ denoting the $1-\alpha/2$ quantile of the standard normal distribution and $V_{n,t} = \frac{1}{n}\sum_{m=1}^{n} m^2 \Delta_{t,m}^2$.

The proof of Proposition 2.6 is given in Section S2 of the Supplement [58], and consists of two steps. First we prove (Proposition S2.1) that, under the given conditions on the predictive updates,

$$\sqrt{n} \begin{bmatrix} \tilde{F}(t_1) - P_n(t_1) \\ \dots \\ \tilde{F}(t_k) - P_n(t_k) \end{bmatrix} \mid x_{1:n} \xrightarrow{d} \mathcal{N}_k(0, U_{\mathbf{t}}(\omega)) \tag{2.6}$$

for $\mathbb{P}$-almost all $\omega = (x_1, x_2, \dots)$. Then we prove that the asymptotic result remains valid if the matrix $U_{\mathbf{t}}$, that depends on the whole sequence $(X_1, X_2, \dots)$, is replaced by its "estimate" $V_{n,\mathbf{t}}$, that only depends on $(X_1, \dots, X_n)$.

Proposition 2.6 gives sufficient conditions that could possibly be relaxed; also, other choices of $V_{n,t}$ can be envisaged. Note that the result is given under the law $\mathbb{P}$, thus, although having a similar flavor, it differs from Bernstein-von Mises asymptotic Gaussian approximations, which are stated with respect to a law $P_{F_0}^{\infty}$ that assumes that the X_i are i.i.d. from a true distribution F_0. Moreover, here the asymptotic variance is expressed in terms of the predictive updates.

As "$\mathbb{P}$-probability one" results, our findings may rather be regarded as a refinement of Doob's theorem for inverse probabilities in the nonparametric case; see point

(ii) in Section 4 of Doob [44] (for us limited to the finite-dimensional distributions). For an exchangeable law $\mathbb{P}$, Doob's theorem ensures that, with $\mathbb{P}$-probability one i.e. for $\mathbb{P}$-almost all $\omega = (x_1, x_2, \ldots)$, the posterior expectation $E(\tilde{F}(\cdot) \mid x_{1:n})$ converges to $F = \tilde{F}(\cdot)(\omega)$ and the posterior variance goes to zero, so that the posterior distribution of $\tilde{F}$ concentrates around F. Proposition 2.6 further describes how the posterior distribution of $\tilde{F}$ concentrates around its conditional expectation: the asymptotic distribution is Gaussian, and in particular, the rescaled asymptotic variance depends on how the predictive distribution varies in response to new data, namely on the predictive updates.

Discussion. Although ours are "probability one results", they give insights on frequentist properties of the posterior distribution, from a novel perspective explicitly related to the behavior of the predictive learning rule. Roughly speaking, our results suggest that frequentist consistency at F_0, and frequentist coverage, can be read as a problem of "efficiency" of the predictive distribution: if the data are generated as i.i.d. from F_0, the predictive distribution, that is, the adopted learning rule, should be able to 'efficiently' learn that. As discussed in [57], the predictive updates should balance the convergence rate with a proper "learning rate": if P_n converges quickly, with predictive updates that quickly decrease to zero, at step n we would be rather sure about the limit $\tilde{F}$ of P_n, reflected in small uncertainty (small variance $V_{n,t}$) in the posterior distribution of $\tilde{F}$ and narrow credible intervals. On the other hand, very small predictive updates could reflect poor learning; the extreme case being a degenerate predictive distribution $P_n = P_0$ for any n, that converges immediately but does not learn from the data. An open problem we see is thus to explore conditions under which the predictive rule efficiently balances convergence and learning properties and provides asymptotic credible intervals for $\tilde{F}(t)$ with good frequentist coverage.

3. METHODS FOR PREDICTIVE CONSTRUCTIONS

The reader may be fairly convinced that predictive modeling is conceptually sound; but may still be concerned that is it difficult to apply in practice. An interpretable statistical model, when possible, incorporates valuable information, and sounds more natural. Moreover, while there is wide literature on prior elicitation, methodological guidance on "predictive elicitation" is quite fragmented. The aim of this section is to trace some available methodology, and provide a few examples. The methods include the notion of predictive sufficiency, that reconciles predictive modeling to parametric models; and the different notion of *sufficientness*, that generally leads to nonparametric constructions - a point that seems overlooked; and predictive constructions based on stochastic processes with reinforcement. Most of the examples we provide come from Bayesian nonparametric statistics and machine learning, where the predictive approach allows to overcome difficulties in assigning a prior law on infinite-dimensional random objects and has indeed been the basis of vigorous theoretical and applied developments.

3.1 Constraints on the form of point predictions

Basically all predictive constructions make some assessment on the form of the predictive distribution. If a parametric model has been already chosen, it may be enough to restrict the class of point predictions $E(X_{n+1} \mid x_{1:n})$. Diaconis and Ylvisaker's [41] characterization of conjugate priors for models in the natural exponential family (NEF) is possibly the most classic example.

EXAMPLE 3.1. (*Conjugate priors for the NEF.*) Let $X_i \mid \theta \overset{i.i.d.}{\sim} p(x \mid \theta) = e^{x^T\theta - M(\theta)}$, where $p(\cdot \mid \theta)$ is a probability density function on $\mathbb{R}^k$ with respect to a dominating measure λ whose support contains an open interval of $\mathbb{R}^k$, and $M(\theta) = \ln \int e^{x^T\theta} d\lambda(x)$, for $\theta \in \Theta = \{s \in \mathbb{R}^k : M(s) < \infty\}$. Because the model is given, the predictive rule characterizes the prior distribution π of $\tilde{\theta}$, which is assumed to be non degenerate. Let $\tilde{\mu} = E(X_1 \mid \tilde{\theta})$ denote the mean vector parameter, which is also the point prediction: $E(X_2 \mid X_1) = E(\tilde{\mu} \mid X_1)$. Diaconis and Ylvisaker ([41], Theorem 3) prove that if $E(\tilde{\mu} \mid X_1) = aX_1 + b$ with $a \in \mathbb{R}$ and $b \in \mathbb{R}^k$, then $a \neq 0$ and the prior density on the natural parameter $\tilde{\theta}$ is the conjugate prior $\pi(\theta) = c \exp(a^{-1}b^T\theta - a^{-1}(1-a)M(\theta))$. This result does not apply to discrete distributions in the NEF, since the support of the dominating measure does not include an interval of $\mathbb{R}^k$. For discrete *univariate* distributions they give an analogous characterization under the assumption $\Theta = (-\infty, \theta_0)$ with $\theta_0 < \infty$. The characterization for the Poisson distributions was already known. □

EXAMPLE 3.2. (*Conjugate prior for binary data*). Let $X_i \mid \tilde{p} = p \overset{i.i.d.}{\sim}$ Bernoulli(p). Diaconis and Ylvisaker [41] prove that, if $E(\tilde{p} \mid X_1, \ldots, X_n)$ - that coincides with $E(X_{n+1} \mid X_1, \ldots X_n)$ - is linear in $\overline{X}_n$ for every n, then the prior on $\tilde{p}$ is the conjugate Beta distribution. The result extends to the characterization of the Dirichlet as the unique family of distributions allowing linear posterior expectation for multinomial observations. □

3.2 Predictive sufficient statistics and parametric models

A natural tool for predictive elicitation is predictive sufficiency. For exchangeable sequences $(X_n)_{n\geq 1}$, the predictive distribution P_n is a function of the entire empirical distribution $\hat{F}_n$. In other words, the empirical distribution is a sufficient summary of $(X_1, \ldots, X_n)$ for prediction of future observations, which is an immediate consequence

of exchangeability. In many applications, it is natural to think that a summary $T_n = T(\hat{F}_n)$ of $\hat{F}_n$ is sufficient, i.e. that the predictive distribution is a function of T_n. The statistic T_n is said to be sufficient for prediction or *predictive sufficient.*

Predictive sufficiency has been investigated by several authors from the 1980's; see the book by Bernardo and Smith ([10], Sect 4.5) and Fortini *et al.* [52], which includes extensive references. Related notions of sufficiency have been studied by Lauritzen ([83], [84]) and Diaconis and Freedman [37]; Schervish ([115], Sect 2.4) gives a review. Several results, and relations with classical and Bayesian sufficiency, are given in [52].

The assumption of a predictive sufficient statistic is strictly connected with the assumption of a parametric model. Informally, if the predictive distribution depends on the data through a predictive sufficient statistic $T_n = T(\hat{F}_n)$ - expressed, with an abuse of notation, as $\mathbb{P}(X_{n+1} \in \cdot \mid X_1, \ldots, X_n) = P_n(\cdot \mid T(\hat{F}_n))$ - then we can expect that, under conditions on T,

$$T_n \equiv T(\hat{F}_n) \to T(\tilde{F}) \equiv \tilde{\theta},$$

(because $\hat{F}_n \to \tilde{F}$); and, under conditions on P_n as a function of T_n,

$$P_n(\cdot \mid T_n) \to F(\cdot \mid \tilde{\theta})$$

for a function F. That is, the statistical model, which is the limit of P_n, has a parametric form $F(\cdot \mid \tilde{\theta})$ where the parameter $\tilde{\theta} = T(\tilde{F})$ is the limit of the predictive sufficient statistic. This is the content of next theorem. A more general result, but technically more involved, is in [52], Theorem 7.1.

THEOREM 3.3. *Let $(X_n)_{n\geq 1} \sim \mathbb{P}$ be an exchangeable sequence with directing random measure $\tilde{F}$. Assume that there is a predictive sufficient statistic $T_n = T(\hat{F}_n)$, where $T : \mathcal{M} \to \mathbb{T} \in \mathcal{B}(\mathbb{R}^k)$ is a continuous function defined on a measurable set $\mathcal{M}$ of probability measures such that $\mathbb{P}(\tilde{F} \in \mathcal{M}) = 1$. For every $n \geq 1$, let $q_n(\cdot, t) = \mathbb{P}(X_{n+1} \in \cdot \mid T(\hat{F}_n) = t)$, $t \in \mathbb{T}$.*
If, for every A with $P_0(\partial A) = 0$, the functions $(q_n(A, \cdot))_{n\geq 0}$ are continuous on $\mathbb{T}$, uniformly in t and n, then there exists a function F such that $\tilde{F}(\cdot) = F(\cdot \mid \tilde{\theta})$, where $\tilde{\theta} = T(\tilde{F})$ is the $\mathbb{P}$-a.s. limit of T_n.

The continuity assumptions in the theorem seem reasonable as a 'robustness' requirement expressing the idea that small changes in the value of the predictive sufficient statistic T_n do not lead to abrupt changes in the prediction. The proof is in Section S3 of the Supplement [58].

EXAMPLE 3.4. Consider a Gaussian model $X_i \mid \mu \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, with $\tilde{\mu} \sim \mathcal{N}(0, 1)$ and known variance σ^2, for simplicity equal to one. Take $\mathcal{M}$ as the set of probability measures with finite first moment, $\mathbb{T} = \mathbb{R}$ and $T(m) = \int x\, dm(x)$, for $m \in \mathcal{M}$. The conditions of Theorem 3.3 hold. First, $E(\int |x| \tilde{F}(dx)) = \int |x| dP_0(x) < +\infty$, which implies that $\tilde{F} \in \mathcal{M}$, $\mathbb{P}$-a.s. The function T is continuous on $\mathcal{M}$ and, for every A, the evaluation on A of $q_n(\cdot, t) = \mathcal{N}(n/(n+1)t, (2+n)/(1+n))$ is continuous in t, uniformly with respect to t and n. □

Theorem 3.3 gives sufficient conditions under which the statistical model is parametric. Stronger conditions are needed if we want to obtain a dominated model.

PROPOSITION 3.5. *Under the assumptions of Theorem 3.3, and*

i) *the predictive distributions P_n are absolutely continuous w.r.t. a dominating measure λ,*
ii) *with probability one, the sequence $(P_n)_{n\geq 0}$ converges to the directing random measure $\tilde{F}$ in total variation,*

then the statistical model is dominated, i.e. $\mathbb{P}$-a.s., $\tilde{F}(\cdot) = F(\cdot \mid \tilde{\theta})$, with $F(\cdot \mid \theta)$ absolutely continuous with respect to λ for every θ.

The proof follows from Theorem 1 in [16], which shows that the conditions i) and ii) are necessary and sufficient for the random directing measure $\tilde{F}$ to be absolutely continuous w.r.t. λ. By Theorem 3.3, $\tilde{F}$ has parametric form $F(\cdot \mid \tilde{\theta})$, and because the limit of P_n is unique almost everywhere, we have the conclusion.

3.3 Predictive "sufficientness"

A different concept is predictive "sufficientness" [133]. The term 'sufficientness' was used by Good [64] with reference to the work by W. E. Johnson [74]. Zabell [130] notes that Good initially used 'sufficiency' but switched to 'sufficientness' to avoid confusion with the usual notion of sufficiency. Here, there is no predictive sufficient statistic beyond the empirical distribution; however, for every set A, the probability that a future observation takes value in A is assumed to depend only on $\hat{F}_n(A)$. In principle, only sufficientness assumptions of the kind above are made; it is however assumed that $(X_n)_{n\geq 1}$ is exchangeable, which introduces constraints on the permissible analytic form of P_n and may identify it.

Interestingly, since the entire empirical distribution is needed for prediction of future observations, we expect that, if no further restrictions are made beyond exchangeability and sufficientness, this type of predictive constructions leads to a *nonparametric* model.

EXAMPLE 3.6. (*Sufficientness characterization of the Dirichlet conjugate prior for categorical data*). Consider an exchangeable sequence $(X_n)_{n\geq 1}$ of categorical r.v.'s

with values in $\{1,\dots,k\}$ with $k>2$, finite. With the notation as in Example 2.5,

$$X_i \mid (\tilde{p}_1,\dots,\tilde{p}_k) \overset{i.i.d.}{\sim} \begin{cases} 1,\dots,k \\ \tilde{p}_1,\dots,\tilde{p}_k. \end{cases} \tag{3.1}$$

No parametric form is imposed on the masses $(\tilde{p}_1,\dots,\tilde{p}_k)$; in this sense, this is a "nonparametric" setting. Since the sequence $(X_n)_{n\geq 1}$ is exchangeable, the predictive distribution depends on the empirical frequencies $(n_1,\dots,n_k)$, i.e. $\mathbb{P}(X_{n+1}=j \mid x_{1:n}) = \mathbb{P}(X_{n+1}=j \mid n_1,\dots,n_k)$. The sufficientness postulate states that the predictive probability of $X_{n+1}=j$ only depends on n_j,

$$\mathbb{P}(X_{n+1}=j \mid x_{1:n}) = f_{n,j}(n_j), \quad j=1,\dots,k. \tag{3.2}$$

We stress that, to provide the predictive probabilities for *all* j, the entire vector of empirical frequencies is needed.

Formally developing an argument by [74], Zabell [130] proves that the sufficientness assumption (3.2), together with $\mathbb{P}(X_1=x_1,\dots,X_n=x_n)>0$ for every $(x_1,\dots,x_n)$, implies that $f_{n,j}(n_j)$ is linear in n_j, and more specifically that, if the X_i are not independent, there exist positive constants $(\alpha_1,\dots,\alpha_k)$ such that

$$\mathbb{P}(X_{n+1}=j \mid n_j) = \frac{\alpha_j + n_j}{\alpha + n}, \tag{3.3}$$

where $\alpha = \sum_{i=1}^k \alpha_i$. In turn, this allows to obtain the expression of all the moments of the prior distribution, which are shown to characterize the Dirichlet$(\alpha_1,\dots,\alpha_k)$ distribution as the prior for $(\tilde{p}_1,\dots,\tilde{p}_n)$. [130] also includes results for finite exchangeable sequences. □

Johnson's sufficientness postulate can be extended to the case of r.v.'s taking values in a general Polish space $\mathbb{X}$. Consider $(X_n)_{n\geq 1}$ exchangeable and assume that for any $n\geq 1$, the predictive rule states that for any measurable set A

$$P_n(A) = \mathbb{P}(X_{n+1}\in A \mid \hat{F}_n(A)). \tag{3.4}$$

Since $(X_n)_{n\geq 1}$ is exchangeable, there exists $\tilde{F}$ such that $X_i \mid \tilde{F} \overset{i.i.d.}{\sim} \tilde{F}$. Again, the entire empirical distribution is needed to obtain the predictive distribution, thus we expect to characterize a nonparametric prior on the random distribution $\tilde{F}$. This is indeed the case.

PROPOSITION 3.7. *Let $(X_n)_{n\geq 1}$ be an exchangeable sequence and assume that $X_1\sim P_0$ and, for any $n\geq 1$, the predictive distribution satisfies (3.4). If the X_i are not independent, then the directing random measure $\tilde{F}$ has a Dirichlet process distribution with parameters (α,P_0) for some $\alpha>0$, denoted $\tilde{F}\sim DP(\alpha,P_0)$.*

The proof of Proposition 3.7 is in Section S3 of the Supplement [58]. This result seems new. Doksum ([42], Corollary 2.1) proves that the Dirichlet process is the only 'non trivial' process such that the posterior distribution of $\tilde{F}(A)$ given $x_{1:n}$ only depends on the number n_A of observations in A (and not on where they fall within or outside A). This implies that the predictive distribution of X_{n+1} given $x_{1:n}$ only depends on n_A; but the latter is a weaker condition. The proposition above shows that it still implies that $\tilde{F}$ is a Dirichlet process. Other characterizations of the Dirichlet process through sufficientness use the additional assumption that the predictive distribution has a specific linear form, as e.g. in [87], or, equivalently, assume that the X_i are categorical r.v.'s. Actually, the sufficientness postulate (3.4) is reasonable only for categorical r.v.'s (for continuous data, for example, one would not fully exploit the information in the sample).

A number of nonparametric priors are characterized by forms of predictive sufficientness. Zabell [132] characterizes the two parameter Dirichlet process from sufficientness assumptions (see Example 3.11 in Section 3.5). Extensions to the class of Gibbs-type priors [28] and to hierarchical generalizations are given by [7]. Muliere and Walker [127] give a predictive characterization of Neutral to the Right processes [42] based on an extension of Johnson's sufficientness postulate. Sariev and Savov [113] provide a sufficientness characterization of exchangeable measure-valued Pólya urn sequences.

3.4 Stochastic processes with reinforcement

Stochastic processes with reinforcement, originated from an idea by Diaconis and Coppersmith [25], are perhaps the main tool used in Bayesian statistics for predictive constructions. They express the idea that, if an event occurs along time, the probability that it occurs again in the next time increases (is *reinforced*). They are of interest in probability and in many areas beyond Bayesian statistics; applications include population dynamics, network modeling (where they are often referred to as preferential attachment rules), learning and evolutionary game theory, self-organization in statistical physics and many more. A beautiful review is given by Pemantle [99].

Urn schemes are basic building blocks for random processes with reinforcement.

EXAMPLE 3.8. (*Two-color Pólya urn*) The simplest example is the two color Pólya urn ([47], [107]). One starts with an urn that contains α balls, of which α_1 are white and the others are black. At each step, a ball is picked at random and returned in the urn along with an additional ball of the same color. Denoting by X_n the indicator of a white additional ball at step n, and by Z_n, $n\geq 0$, the proportion of white balls in the urn before the $(n+1)$th draw, we have $\mathbb{P}(X_1=1)=\alpha_1/\alpha=Z_0$ and for any $n\geq 1$

$$\mathbb{P}(X_{n+1}=1 \mid X_1,\dots,X_n) = \frac{\alpha_1+\sum_{i=1}^n X_i}{\alpha+n} = Z_n.$$

The two color Pólya urn was proposed as a model for the evolution of contagion. In Bayesian statistics, Pólya sampling is not meant as describing a process that actually evolves over time (such as the spread of contagion), but describes the evolution of information; namely a learning process where the probability that the next observation is white is *reinforced* as more white balls are observed in the sample. It is well known that the sequence $(X_n)_{n\geq 1}$ so generated is exchangeable, and that both the relative frequency $\sum_{i=1}^n X_i/n$ and the proportion of white balls Z_n converge to a random limit $\tilde{\theta} \sim \text{Beta}(\alpha_1, \alpha - \alpha_1)$. Thus, from the predictive rule we get $X_i \mid \tilde{\theta} \overset{i.i.d.}{\sim} \text{Bernoulli}(\tilde{\theta})$ with a conjugate $\text{Beta}(\alpha_1, \alpha - \alpha_1)$ prior. □

The celebrated extension to a countable number of colors are Pólya sequences [17], see the following Example 3.9. Many more exchangeable predictive constructions are based on reinforced stochastic processes; we provide a few notable examples in the next section.

3.5 Examples in Bayesian nonparametrics

EXAMPLE 3.9. (*The Dirichlet process*) In Section 3.3, we have seen a characterization of the Dirichlet process in terms of sufficientness. The predictive characterization as an extension of Pólya sampling was given by Blackwell and MacQueen [17]. For data in a Polish space $\mathbb{X}$, Blackwell and MacQueen define *Pólya sequences* $(X_n)_{n\geq 1}$ as characterized by the predictive rule $X_1 \sim P_0$ and for any $n \geq 1$,

$$X_{n+1} \mid X_1, \ldots, X_n \sim P_n = \frac{\alpha}{\alpha + n} P_0 + \frac{n}{\alpha + n} \hat{F}_n, \tag{3.5}$$

where $\alpha > 0$. They prove that a Pólya sequence is exchangeable and P_n converges $\mathbb{P}$-a.s. to a *discrete* random distribution $\tilde{F}$; moreover, $\tilde{F} \sim \text{DP}(\alpha, P_0)$. It follows that $X_i \mid \tilde{F} \overset{i.i.d.}{\sim} \tilde{F}$, with a $\text{DP}(\alpha, P_0)$ prior on $\tilde{F}$.

Pólya sequences can be also described as reinforced urn processes; the interest in such characterization is that it enlightens the link with the theory of random partitions. Indeed, the discrete nature of the Dirichlet process, that follows from (3.5), implies that ties are observed in a random sample $(X_1, \ldots, X_n)$ with positive probability. This induces a random partition of $\{1, \ldots, n\}$, with i and j in the same group if $X_i = X_j$. The characterization as a reinforced urn model explicates its probability law.

Rather than an impractical urn with infinitely many balls, a proper urn metaphor is the Hoppe's urn scheme ([69], [70]), also popularly described as the Chinese Restaurant Process [2]. Consider sampling from an urn that initially only contains $\alpha > 0$ black balls. At each step, a ball is picked at random and, if colored, it is returned in the urn together with an additional ball of the same color; if black, the additional ball is of a new color. Natural numbers are used to label the colors and they are chosen sequentially as the need arises. The sampling generates a process $(L_n)_{n\geq 1}$, where L_n denotes the label of the additional ball returned after the nth draw. Clearly, the sequence $(L_n)_{n\geq 1}$ is not exchangeable. However, if one 'paints' it, picking colors, when needed, from a color distribution P_0, then the resulting sequence of colors $(X_n)_{n\geq 1}$ has predictive rule (3.5), thus it is a Pólya sequence with parameters (α, P_0). In terms of the Chinese Restaurant metaphor, where customers enter sequentially in the restaurant and are allocated either in a occupied table, or in a new one, $L_n = j$ denotes that the n customer is seated at table j, and for any $n \geq 1$, the label's configuration $(L_1, \ldots, L_n)$ gives the allocation of customers at tables, representing the random partition; then tables are painted at random from the color distribution P_0.

For any $n \geq 1$, let $\rho_n = (A_1, \ldots, A_{k_n})$ be the random partition of $\{1, 2, \ldots, n\}$ so generated (where $i \in A_j$ if $L_i = j$, k_n is the number of colors that have been created, or of the occupied tables, and the A_j are in order of appearance). The probability mass function, or *partition probability function*, of ρ_n is easily computed from the labels' sampling scheme; if P_0 is diffuse, we have

$$\mathbb{P}(\rho_n = (A_1, \ldots, A_{k_n})) = \frac{\alpha^{k_n}}{\alpha^{[n]}} \prod_{j=1}^{k_n} (n_j - 1)! \tag{3.6}$$

where $\alpha^{[n]} = \alpha(\alpha + 1) \cdots (\alpha + n - 1)$ and n_j is the number of elements of A_j, $j = 1, \ldots, k_n$. See [48], [4], [69]. □

The above characterization is an emblematic example of the potential of predictive constructions - in this case, explicating the link with random partitions theory. In Bayesian statistics, the capacity of the Dirichlet process of generating random partitions is leveraged for model-based clustering in many applications; beyond Bayesian statistics, random partitions, and in particular, *exchangeable* random partitions, are of interest in a wide range of fields such as combinatorics, genetics, population dynamics. The construction in Example 3.9 extends more generally; let us recall a few basic notions that we use in the following examples.

Given an exchangeable sequence $(X_n)_{n\geq 1}$ one can define a random partition ρ_n of $\{1, \ldots, n\}$ by letting i and j be in the same group if $X_i = X_j$. Then we have

$$\mathbb{P}(\rho_n = (A_1, \ldots, A_{k_n})) = p(n_1, \ldots, n_{k_n}) \tag{3.7}$$

for a symmetric function p of $(n_1, \ldots, n_{k_n})$, where n_j is the number of elements in A_j. A partition probability function p so generated is called the *exchangeable partition probability function* (EPPF) derived from the sequence $(X_n)_{n\geq 1}$. More formally, p is defined on the space of sequences $\mathbf{n} = (n_1, n_2, \ldots)$, identifying $(n_1, \ldots, n_{k_n})$

as $\mathbf{n} = (n_1, \dots, n_{k_n}, 0, 0, \dots)$. Let $\mathbf{n}^{j+}$ be defined from $\mathbf{n}$ by incrementing n_j by 1. Clearly an EPPF p must satisfy

$$p(1,0,0,\dots) = 1 \quad \text{and} \quad p(\mathbf{n}) = \sum_{j=1}^{k_n+1} p(\mathbf{n}^{j+}).$$

In predictive modeling, the conditional probability that the next observation X_{n+1} is in group j, given $x_{1:n}$, is

$$p_j(\mathbf{n}) = \frac{p(\mathbf{n}^{j+})}{p(\mathbf{n})} \text{ provided } p(\mathbf{n}) > 0, \tag{3.8}$$

for $j = 1, \dots, k_n + 1$. The concept of EPPF has been introduced in Pitman [103]. A fundamental result in the theory of exchangeable random partitions is Kingman's de Finetti-like representation theorem for exchangeable random partitions as mixtures of paint-box processes [82]; for extensive treatment, we refer to [81], [105], [133].

EXAMPLE 3.10. (*Species sampling priors*). Pitman [104] defines a class of predictive rules, in the framework of species sampling, that generalizes Blackwell and McQueen scheme (3.5). One underlines sequential sampling from a discrete random distribution for categorical data - in species sampling, sequential draws from a population of species labeled in the order they are discovered with tags X_j^* i.i.d. from a diffuse distributions P_0. Here X_i represents the species of the ith individual sampled and takes values in the set of tags. In a sample $x_{1:n}$, one observes k_n distinct species, labeled $x_1^*, \dots, x_{k_n}^*$ and the next observation X_{n+1} will either be one of the species already discovered in the sample, or a new one, formalized in the predictive distribution $P_n(\cdot \mid x_{1:n}) = \sum_{j=1}^{k_n} p_{j,n}(x_{1:n})\delta_{x_j^*}(\cdot) + p_{k_n+1,n}(x_{1:n})P_0(\cdot)$. In random sampling, the sequence $(X_n)_{n\geq 1}$ should be exchangeable, and a necessary condition is that $p_{j,n}$ depends on $(x_1, \dots, x_n)$ only through the sequence of counts $\mathbf{n} = (n_1, n_2, \dots)$ (terminating with a string of zeroes) of the various species in the sample in the order of appearance.

A sequence $(X_n)_{n\geq 1}$ is a *species sampling sequence* if it is exchangeable and has a predictive rule of the form

$$P_n(\cdot) = \sum_{j=1}^{k_n} p_j(\mathbf{n})\delta_{X_j^*}(\cdot) + p_{k_n+1}(\mathbf{n})P_0(\cdot), \tag{3.9}$$

for $n \geq 1$, for a diffuse distribution P_0, which is also the law of X_1. Pitman ([104], Theorem 14) shows that exchangeability holds if and only if there exists a non-negative symmetric function p that drives the probabilities $p_j(\mathbf{n})$ according to (3.8). Then the EPPF of $(X_n)_{n\geq 1}$ is the unique non-negative symmetric function p such that (3.8) holds and $p(1) = 1$.

From exchangeability, by Proposition 2.4 we have that, with probability one, P_n converges to a random distribution $\tilde{F}$; Pitman ([104], Proposition 11) proves that the convergence is in total variation norm and $\tilde{F}$ has the form

$$\tilde{F}(\cdot) = \sum_{j=1}^{k_\infty} p_j^*\delta_{X_j^*}(\cdot) + (1 - \sum_{j=1}^{k_\infty} p_j^*)P_0(\cdot), \tag{3.10}$$

where $p_j^* = \lim n_j/n$ is the random limit of the relative frequency of the j-th species discovered, the X_j^* are i.i.d. according to P_0, independently of the p_j^* and $k_\infty = \inf\{k : p_1^* + \cdots + p_k^* = 1\}$ is the number of distinct values to appear in the infinite sequence $(X_1, X_2, \dots)$.

The above results do not provide an explicit description of the distribution of the weights p_j^*, which is however available in remarkable special cases, including the Dirichlet process, that corresponds to $p_j(\mathbf{n}) = n_j/(\alpha + n)$, where $\alpha > 0$ is a fixed number; the finite Dirichlet process [72] that assumes $p_j(\mathbf{n}) = (n_j + \alpha/K)/(\alpha + n)$ for $j \leq k_n \leq K$, where $\alpha > 0$ and $K \in \mathbb{N}$ are fixed numbers; and the two parameter Poisson-Dirichlet process. □

EXAMPLE 3.11. (*Two parameter Poisson-Dirichlet process)* The two parameter Poisson-Dirichlet process, or Pitman-Yor process, introduced in [100] and further studied in [103] and [106], can be viewed both as an extension of the Dirichlet process and as the directing random measure of a species sampling sequence characterized by the predictive rule (3.9) with

$$p_j(\mathbf{n}) = \frac{n_j - \theta}{\alpha + n} \text{ and } p_{k_n+1}(\mathbf{n}) = \frac{\alpha + k_n\theta}{\alpha + n}, \tag{3.11}$$

where α and θ are real parameters satisfying $0 \leq \theta < 1$ and $\alpha > -\theta$. As it appears from (3.11), the Poisson-Dirichlet process allows for a more flexible predictive structure than the Dirichlet process (corresponding to $\theta = 0$): the predictive probability of observing a new species at time n depends on both n and the number k_n of distinct species sampled.

In analogy to Example 3.9, the sequence $(X_n)_{n\geq 1}$ can be described as a *generalized Hoppe's urn* [132] if $\alpha > 0$. Initially, the urn only contains one black ball of weight α, and balls are selected with probabilities proportional to their weights; whenever a black ball is selected, it is returned into the urn together with *two* new balls, one black, having weight θ, and one of a new color, sampled from P_0, having weight $1 - \theta$.

The two-parameter Poisson-Dirichlet process is also characterized through sufficientness [132]; namely, by postulating the sufficientness of n_j and of k_n in the predictive probabilities $p_j(\mathbf{n})$ and $p_{k_n+1}(\mathbf{n})$, respectively.

For increasing n, the predictive distribution P_n converges $\mathbb{P}$-a.s. to a discrete random measure $\tilde{F} = \sum_{j=1}^{\infty} p_j^*\delta_{X_j^*}$, where the p_j^* have the stick-breaking representation $p_j^* = \prod_{i=1}^{j-1}(1 - V_i)V_j$, with $V_i \overset{indep}{\sim} \text{Beta}(\alpha + i\theta, 1 - \theta)$. The predictive construction can be exploited to design computational strategies (see e.g. [8]). □

In some examples, the predictive construction does not characterize a novel prior, but explicates the predictive assumptions that are made when adopting a certain (already known) prior law – which is clearly important; and here is an example of a purely predictive construction, whose de Finetti-like representation and implied prior law was only found afterwards.

EXAMPLE 3.12 (*Indian Buffet Process)*. The Indian Buffet process, introduced by Griffith and Ghahramani [65], is a clever and popular predictive scheme for infinite latent features problems. Here, exchangeable objects or individuals are each described through a potentially infinite array of features, resulting in an underlying random binary matrix with rows representing the individuals, and with an unbounded number of columns, representing the features. Specifically, a 1 in the $[n, k]$ entry of the random matrix indicates that the nth individual possesses the kth feature. The predictive construction can be illustrated by imagining customers sequentially entering an Indian Buffet restaurant. In this metaphor, customers represent individuals, dishes symbolize features, and when a customer selects a dish z, it means that the corresponding individual possesses feature z. Let θ be a fixed strictly positive number. The first customer chooses a Poisson(θ) number of dishes from a non-atomic distribution F_0. Then, for $n = 1, 2, \ldots$, the $(n+1)$th customer decides, for each of the k_n dishes already served, whether to take it or not, according to its popularity, namely she chooses dish z with probability $k_{z,n}/(n+1)$, where $k_{z,n}$ is the number of customers who have already chosen dish z, independently for $z = 1, \ldots, k_n$. Then she chooses a Poisson$(\theta/(n+1))$ number of new dishes, sampling them from F_0.

This construction, which is purely predictive, characterizes an exchangeable law for the individuals' features, represented as $X_n = \sum_{k=1}^{\infty} b_{n,k}\delta_{Z_k}$, where $b_{n,k} = 1$ if the nth individual possesses feature Z_k, and zero otherwise, and with the features $(Z_k)_{k\geq 1}$ independently sampled from F_0; and enables Bayesian learning without an explicit prior law. Actually, the implied prior law was later made explicit [122], and assumes that, conditionally on a sequence $(p_k)_{k\geq 1}$ of r.v.s taking values in $(0, 1)$, the $(b_{n,k})_{n,k\geq 1}$ are sampled independently, with $b_{n,k} \sim$ Bernoulli(p_k). In turn, the $(p_k)_{k\geq 1}$ are the points of a Poisson random measure with mean intensity $\lambda(s) = \theta s^{-1}\mathbf{1}_{(0,1)}(s)$.

The Indian Buffet process has been extended for allowing different distributions on $(p_k, b_{i,k})_{i,k\geq 1}$ (see [73], [19] and references therein) or random weights [11]. □

EXAMPLE 3.13 (*Predictive constructions for continuous data)*. As already mentioned, the predictive rule (3.5) of Pólya sequences is appropriate for categorical data, but, as it appears from the underlying sufficientness postulate (3.4), it is not efficient for continuous data, failing to fully exploit the sample information. A similar remark holds for species sampling sequences. Indeed, in Bayesian statistics, the Dirichlet process and generally discrete prior laws are mostly used at the *latent* stage of hierarchical models, where, as already noted by Antoniak [4], their power in generating a random partition is an asset; see e.g. [121] and [96] for overviews. A popular example are Dirichlet process mixture models where, conditionally on a latent exchangeable sequence $(\tilde{\theta}_n)_{n\geq 1}$, the X_i are independent and the distribution of X_i only depends on $\tilde{\theta}_i$, with a slight abuse of notation written as

$$X_i \mid \tilde{\theta}_i \overset{indep}{\sim} k(\cdot \mid \tilde{\theta}_i), \tag{3.12}$$

for a kernel density k, and

$$\tilde{\theta}_i \mid \tilde{G} \overset{i.i.d.}{\sim} \tilde{G}, \quad \text{with} \quad \tilde{G} \sim DP(\alpha, G_0). \tag{3.13}$$

This gives an exchangeable mixture model:

$$X_i \mid \tilde{G} \overset{i.i.d.}{\sim} f_{\tilde{G}}(\cdot) = \int k(\cdot \mid \theta) d\tilde{G}(\theta).$$

The predictive rule of the Dirichlet process induces a parametric model on the random partition of the $\tilde{\theta}_i$'s of the form (3.6), and X_i and X_j are set in the same cluster if $\tilde{\theta}_i = \tilde{\theta}_j$. This is a powerful and popular use of predictive rules such as (3.5), which however would not be appropriate as predictive learning rules at the observation level with continuous data.

An approach to address this difficulty is to smooth the trajectories generated by discrete priors thus obtaining novel prior laws that almost surely select absolutely continuous distributions; for example, a constructive smoothing of the Dirichlet process through Bernstein polynomials was proposed, from an idea of Diaconis, by [101] and extended by [102], who obtained a general class of mixture priors. However, in these constructions, and more generally in Bayesian mixture models with a discrete prior law on the mixing distribution, the predictive distribution is not analytically tractable, requiring to average with respect to the posterior law on the huge space of partitions (see e.g. [125]).

A predictive approach may consist in directly smoothing the empirical distribution in predictive rules such as (3.5). Recent proposals are *kernel-based Dirichlet sequences* [13], that are defined as exchangeable sequences whose predictive distributions spread the point mass δ_{X_i} in (3.5) through a probability kernel K, as

$$P_n(\cdot) = \frac{\alpha}{\alpha+n} P_0(\cdot) + \frac{1}{\alpha+n}\sum_{i=1}^{n} K(\cdot \mid X_i).$$

The exchangeability condition imposes that the kernel K must satisfy $K(\cdot \mid x) = P_0(\cdot \mid \mathcal{G})(x)$ for a sigma-algebra $\mathcal{G}$ on $\mathbb{X}$ ([13], [112]). In particular, in their perhaps most natural specification, with $K(\cdot \mid x) \ll P_0$ for every $x \in \mathbb{X}$,

the underlying $\tilde{F}$ is a mixture model with kernels having known disjoint support (e.g. a histogram with known bins), see [112], Theorem 3.13; which is clearly limited for statistical applications.

This example hints that exchangeability constraints may be quite restrictive if one wants to have both a tractable predictive rule and some desired modeling features. Here is a predictive construction that is analytically simple, and gives another 'smoothed version' of (3.5). We start from $P_0(\cdot) = \int K(\cdot \mid \theta) dG_0(\theta) \equiv F_{G_0}(\cdot)$ and recursively update our prediction as

$$P_n(\cdot) = (1-\alpha_n)P_{n-1}(\cdot) + \alpha_n F_{G_{n-1}}(\cdot \mid X_n),$$

with $F_{G_{n-1}}(\cdot \mid X_n) = \int K(\cdot \mid \theta) dG_{n-1}(\theta \mid X_n)$, where $G_{n-1}(\cdot \mid X_n)$ denotes the posterior distribution obtained from the prior G_{n-1} and updated based on X_n, and $G_n(\cdot) = (1-\alpha_n)G_{n-1}(\cdot) + \alpha_n G_{n-1}(\cdot \mid X_n)$; and the α_n are real numbers in $(0,1)$ satisfying $\sum_{n=1}^{\infty} \alpha_n = +\infty$ and $\sum_{n=1}^{\infty} \alpha_n^2 < +\infty$. (We may recognize 'Newton's algorithm' [95], popularly used for fast computations in Dirichlet process mixture models). This predictive rule does not characterize an exchangeable sequence $(X_n)_{n\geq 1}$; it is however a martingale and preserves exchangeability asymptotically. More specifically, it can be shown ([53]) that P_n converges to a mixture model $F_{\tilde{G}}(\cdot)$ with a novel prior law on $\tilde{G}$, as we will expand in Section 5. □

Further examples, among many, include the class of *reinforced urn processes* ([126], [91]; see Example 4.9), and constructions aimed at addressing the rigidity of the global clustering induced by the predictive rule (3.5) of the Dirichlet process in the case of multivariate random distributions; for example, [124] obtain a nested clustering for multivariate data characterized by an *enriched* Hoppe's urn scheme. In the next section we present more predictive constructions, based on the idea of reinforcement, that characterize forms of *partial* exchangeability.

4. PARTIAL EXCHANGEABILITY FOR MORE STRUCTURED DATA

As seen, exchangeability in Bayesian statistics is the natural predictive requirement in homogeneous repeated trials; but of course data may be much more complex. Still, in many cases the data show forms of symmetry, such that exchangeability assessments, judging that the individuals' labels in some data sub-structures do not bring any information for prediction, are still natural. In this section we review the concept of *partial exchangeability*, i.e. invariance under a group of permutations. For the sake of space, we focus on the main concepts and on de Finetti-like representation theorems that again justify the Bayesian inferential model from predictive assumptions. The predictive characterization in Theorem 4.4 is new. We start with the notion of partial exchangeability in the sense of de Finetti and a point we will underline is that other forms of partial exchangeability are ultimately related to it.

4.1 de Finetti's partial exchangeability

A first notion of partial exchangeability was introduced by de Finetti in [31]. It is interesting to report some excerpt from this, perhaps less known, writing by de Finetti, as it clearly shows what are the applied contexts that suggest a partial exchangeability assessment. de Finetti [31] refers to replicates of trials of different *types*, for which "exchangeability can still be considered, but specifying that the trials are divided into a certain number of types, and what is judged exchangeable are the events of the same type."

As a simple example, he considers tossing two coins. If the two coins look exactly alike, one may judge all tosses as exchangeable; at the opposite extreme, if the coins are completely different, one would consider the corresponding tosses as two separate exchangeable sequences, completely independent of each other. However, if the coins look almost alike, then

> observations of the tosses of one coin will still be capable of influencing, although in an *less direct* manner, our probability judgment regarding the tosses of the other coin.

Again from [31]: "One can have any number of types of trials", for example different coins, or tosses of one coin by two different people, or under different conditions of temperature and atmospheric pressure.

> If the types are in a countable or continuous set, prediction would typically refer to a new type; thus, information will exclusively be *indirect*.

(de Finetti's note [31] includes several more examples, e.g. in insurances and in treatments' effects and debatable causality). In the Bayesian literature, partial exchangeability in the sense of de Finetti is usually referred to random sampling in (a finite number of) parallel experiments; as we see, it may refer more generally to fixed-design regression where the 'types' are induced by covariates. Although the experiments are run independently, each of them brings information on the other ones, and because information is described through probability (see Section 1), the joint probability law will assume a form of dependence across the experiment-specific samples, i.e. of sharing information in prediction.

Formalizing, consider a family of sequences $(X_{n,j})_{n\geq 1}$ of r.v.'s where $X_{n,j}$ describes the nth observation of type j, $j = 1, \ldots, M$; M can be finite or infinite, and the types may be taken from a continuum of types.

For more compact notation, we may arrange them in an array $[X_{n,j}]_{n\geq 1, j=1,\ldots,M}$. The family of sequences $[X_{n,j}]_{n\geq 1, j=1,\ldots,M}$ is *partially exchangeable in the sense of de Finetti* if its probability law is invariant under separate finite permutations within each column; that is, if

$$[X_{n,j}]_{n\geq 1, j=1,\ldots,M} \stackrel{d}{=} [X_{\sigma_j(n),j}]_{n\geq 1, j=1,\ldots,M}$$

for any finite permutation σ_j, $j = 1,\ldots,M$. Roughly speaking, observations are exchangeable inside each experiment, but not across experiments. Aldous ([2], page 23) refers to this symmetry property as *exchangeability over* V. A sequence $(Y_n)_{n\geq 1}$ is exchangeable over V if $(V, Y_1, Y_2, \ldots) \stackrel{d}{=} (V, Y_{\sigma(1)}, Y_{\sigma(2)}, \ldots)$ for any finite permutation σ. Partial exchangeability corresponds to each sequence $(X_{n,j})_{n\geq 1}$ being exchangeable over all the others, collected as V_j.

The representation theorem extends to partially exchangeable families of sequences.

THEOREM 4.1 (Law of large numbers and de Finetti representation theorem for partially exchangeable sequences). *Let $[X_{n,i}]_{n\geq 1, i=1,\ldots,M} \sim \mathbb{P}$ be a partially exchangeable array in the sense of de Finetti. Then:*

i) *For $n_1,\ldots,n_M \to \infty$, the vector of the marginal empirical distributions $(\hat{F}_{n_1},\ldots,\hat{F}_{n_M})$ converges weakly to a vector of random distributions $(\tilde{F}_1,\ldots,\tilde{F}_M)$, $\mathbb{P}$-a.s.;*
ii) *For any $n \geq 1$ and measurable sets $A_{i,j}$,*

$$\mathbb{P}(\cap_{j=1}^{M}(X_{1,j}\in A_{1,j},\ldots,X_{n_j,j}\in A_{n_j,j})) = \int \prod_{j=1,\ldots,M}\prod_{i=1,\ldots,n_j} F_j(A_{i,j})d\pi(F_1,\ldots,F_M),$$

where π is the joint probability law of $(\tilde{F}_1,\ldots,\tilde{F}_M)$.

A proof is in [2], pp. 23-25. The representation ii) is often phrased as: conditionally on $(\tilde{F}_1,\ldots,\tilde{F}_M)$, the sequences $(X_{n,j})_{n\geq 1}$ are independent and, within sequence j, the $X_{n,j}$ are i.i.d. according to $\tilde{F}_j$. That is, a de Finetti-partially exchangeable array is obtained by first picking $(F_1,\ldots,F_M)$ from a *joint* prior distribution and then for each $j = 1,\ldots,M$ picking $X_{n,j} \stackrel{i.i.d.}{\sim} F_j$, independently for different j.

EXAMPLE 4.2 (*Hierarchical models).* Consider random samples $(X_{1,j},\ldots,X_{n_j,j})$, $j = 1,\ldots,M$, from M independent parallel experiments, say of binary r.v.'s with experiment specific means θ_j, and the classic problem of estimating the mean vector $(\theta_1,\ldots,\theta_M)$. This problem is also described (e.g. in [46]) as predicting $X_{n_j+1,j}$ in each experiment. Bayesian hierarchical models are a powerful tool for borrowing strength across experiments and for shrinkage. In this example, a basic hierarchical model regards the parameters as r.v.'s $\tilde{\theta}_j$, sampled from a latent distribution, and assumes a hierarchical structure as follows

$$\tilde{\theta}_j \mid \lambda \stackrel{i.i.d.}{\sim} \pi(\cdot \mid \lambda), \quad \text{with } \lambda \sim h(\cdot),$$

$$(X_{1,j},\ldots,X_{n_j,j}) \mid \theta_1,\ldots,\theta_M \stackrel{i.i.d.}{\sim} \text{Bernoulli}(\theta_j),$$

independently across j (here π and h denote densities). The theoretical justification of this model comes from the assessment of partial exchangeability of the sequences $(X_{n,j})_{n\geq 1}$, and of exchangeability of the experiments. By partial exchangeability, the observations are exchangeable inside each experiment, but not across them; and the sequences $(X_{n,j})_{n\geq 1}$ are only *conditionally* independent given $(\tilde{\theta}_1,\ldots,\tilde{\theta}_M)$, which naturally implies sharing information. The dependence across experiments is introduced through the joint prior law of $(\tilde{\theta}_1,\ldots,\tilde{\theta}_M)$ - here, the joint density $\pi(\theta_1,\ldots,\theta_M) = \int \prod_{j=1}^{M} \pi(\theta_j \mid \lambda)h(\lambda)d\lambda$. □

EXAMPLE 4.3. In hierarchical models as above, the prior law π expresses the judgement that the $\tilde{\theta}_j$ - informally, the experiments - are exchangeable. But, more generally, the groups may be induced by covariates, or refer to time or space, etc., and the prior would express other forms of dependence. For example, consider clinical trials where the outcome is binary (tumor shrank or not), run in different hospitals, with patients receiving the same treatment in all hospitals. Here one would judge that the hospitals' labels do not bring information, that is, the hospitals (the corresponding model parameters $\tilde{\theta}_j$'s) are exchangeable; as in the example above. Now suppose that different treatments, say different dosages z_j, are administrated in different hospitals. Then the groups' labels are relevant, and the prior will not treat the $\tilde{\theta}_j$'s as exchangeable, but will incorporate the effect of the covariate; for example, express the idea that $\tilde{\theta}_j$ and $\tilde{\theta}_k$ are similar if the dosages z_j and z_k are close.

With no replicates inside the groups and no random effects - i.e. in a basic fixed-design regression context where the probability of success is $\tilde{\theta}_j = g(z_j;\tilde{\beta})$ for a known g and unknown $\tilde{\beta}$ - partial exchangeability reduces to conditional independence of the $X_{1,j}$ given $\tilde{\beta}$, with dependence across j modeled through the regression function.□

Marginally, the result of Theorem 4.1 is not surprising, because each sequence $(X_{n,j})_{n\geq 1}$ is exchangeable and one obtains the marginal directing random measure $\tilde{F}_j$ (the statistical model and the prior for experiment j) as seen in Section 2; in particular, from

$$P_{n,j}(\cdot) \equiv \mathbb{P}(X_{n+1,j}\in\cdot \mid X_{1,j},\ldots,X_{n,j}) \to \tilde{F}_j(\cdot). \quad (4.1)$$

But this is not enough: the theorem characterizes the *joint* distribution (the joint prior law) of $(\tilde{F}_1,\ldots,\tilde{F}_M)$.

It is this joint distribution that induces probabilistic dependence across the individual sequences, i.e. *borrowing strength* in prediction. As in Example 4.2, rather than the marginal predictive distribution $P_{n,j}$ in (4.1), a more interesting predictive distribution refers to future results in experiment j given past observations therein *and* observations in all the related experiments. Aldous's notion of exchangeability over V is particularly suited. Let $V = [(X_{n,i})_{n\geq 1; i=1,\dots,M; i\neq j}]$ collect the observations in all the experiments but the jth. Then, with $\mathbb{P}$-probability one,

$$\begin{aligned}(4.2)\quad &\lim_n \mathbb{P}(X_{n+1,j} \in \cdot \,|\, X_{1,j},\dots,X_{n,j},V)\\ &= \lim_n \mathbb{P}(X_{n+1,j} \in \cdot \,|\, (X_{k,i})_{k\leq n, i=1,\dots,M}) = \tilde{F}_j(\cdot).\end{aligned}$$

For a proof, see [2]. Informally, V does not carry additional information only in the limit, when the experiments become independent.

Note that the rows $([X_{n,1},\dots,X_{n,M}])_{n\geq 1}$ of a de Finetti partially exchangeable array are an exchangeable sequence, with directing random measure $\tilde{F}$ on the product space $(\mathbb{X}_1 \times \cdots \times \mathbb{X}_M)$ that assumes independent components, i.e. $\tilde{F} = \times_{j=1}^M \tilde{F}_j$. This implies that the relationship between variables in distinct columns of the array $[X_{n,j}]_{n\geq 1, j=1,\dots,M}$ is solely driven by the probabilistic link between the marginal directing random measures. Again, the sequences do not *physically* interact.

Also note that $\mathbb{P}(X_{n+1,j} \in \cdot \mid (X_{k,i})_{k\leq n, i=1,\dots,M}) = E(\tilde{F}_j(\cdot) \mid (X_{k,i})_{k\leq n, i=1,\dots,M})$ and, as shown in equation (4.2), approximates $\tilde{F}_j$ for n large. Since the sequence $(X_{n,j})_{n\geq 1}$ is exchangeable, an alternative approximation of $\tilde{F}_j$ is provided by the predictive distribution $P_{n,j}(\cdot) = E(\tilde{F}_j(\cdot) \mid X_{1,j},\dots,X_{n,j})$, that is only based on the past observations in experiment j. However, the latter uses less information, resulting in a less efficient approximation:

$$\begin{aligned}&E\Big(\big(\tilde{F}_j(A) - \mathbb{P}(X_{n+1,j} \in A \mid X_{1,j},\dots,X_{n,j})\big)^2\Big)\\ &\geq E\Big(\big(\tilde{F}_j(A) - \mathbb{P}(X_{n+1,j} \in A \,|\, (X_{k,i})_{k\leq n, i=1,\dots,M}\big)^2\Big).\end{aligned}$$

If the sequences $(X_{n,j})_{n\geq 1}$ are independent, both methods yield the same result; there is no gain of information in considering the entire array.

The predictive characterization of exchangeability of Theorem 2.3 can be extended to de Finetti partial exchangeability.

THEOREM 4.4. *A family of sequences $[X_{n,j}]_{n\geq 1, j=1,\dots,M}$ is partially exchangeable in the sense of de Finetti if and only if for every finite $m \leq M$ and every $n \geq 0$, the following conditions hold:*

i) *For every measurable sets $A_1,\dots,A_m$ and every $i = 1,\dots,m$*

$$P_n(A_1 \times \cdots \times A_m \mid (x_{k,j})_{k\leq n, j\leq m})$$

is symmetric in $(x_{1,i},\dots,x_{n,i})$;

ii) *The set function that maps $\{A_j, B_j : j \leq m\}$ into*

$$\int_{A_1\times\cdots\times A_m} P_{n+1}(B_1 \times \cdots \times B_m \mid (x_{k,j})_{k\leq n+1, j\leq m})\, dP_n((x_{n+1,1},\dots,x_{n+1,m}) \mid (x_{k,j})_{k\leq n, j\leq m})$$

is symmetric in (A_i, B_i) for every $i = 1,\dots,k$,

where P_n is to the conditional distribution of $(X_{n+1,j})_{j\leq m}$, given $(X_{k,j})_{k\leq n, j\leq m}$ and $P_0(\cdot \mid (x_{k,j})_{k\leq 0, j\leq m})$ is meant as P_0.

The proof is provided in Section S4 of the Supplement [58]. The predictive characterization in Theorem 4.4 is natural when at each time n, a new observation is made for each type. In fact, de Finetti's partial exchangeability can be described as invariance of the law of a sequence $(X_1, X_2, \dots)$ to the permutations acting separately on M groups of random variables, forming a partition of $(X_n)_{n\geq 1}$. In this perspective, a predictive characterization of partial exchangeability should account for the structure of the partition into groups, likely in a nontrivial way.

Statistical applications of partial exchangeability are broad; hierarchical models are one of the key strengths of Bayesian statistics. Sharing information in prediction is enabled through the prior law π, and the choice of π in parametric models is a long studied problem. Defining a nonparametric prior on the vector of random distributions $(\tilde{F}_1,\dots,\tilde{F}_M)$ has posed challenges, yet a wealth of proposals is nowadays available, many of which are defined through, or benefit from, predictive characterizations.

EXAMPLE 4.5 (*Hierarchical Dirichlet process*). The hierarchical Dirichlet process has been introduced in [120] to model shared clusters among groups of data. For example, consider the problem of modelling shared topics in a corpus of M documents, where a "topic" induces a multinomial distribution over the words of a given dictionary, and a document j is defined as an unordered - exchangeable - sequence of words $(X_{n,j})_{n\geq 1}$. For each document j, we have a latent sequence of topics $(\tilde{\theta}_{n,j})_{n\geq 1}$, and $X_{n,j} \mid \tilde{\theta}_{n,j} \sim k(\cdot \mid \tilde{\theta}_{n,j})$. The family of sequences $(\tilde{\theta}_{n,j})_{n\geq 1}$ for $j = 1,\dots,M$ is assumed to be partially exchangeable, thus conditionally independent given the vector $(\tilde{G}_1,\dots,\tilde{G}_M)$ of the random distributions of topics in the documents.

A predictive construction that allows for document-specific clustering into topics and shared topics across documents is given in [120] as a hierarchical Chinese Restaurant process, or *Chinese franchise*, which is reminiscent of the hierarchical Hoppe's urn proposed for infinite hidden Markov models by [9], as we here describe. To each document j, let us associate a Hoppe's urn $\mathcal{U}_j$, that initially only includes $\alpha_j > 0$ black balls, then sample from each urn as described in Example 3.9; however,

whenever a new color is needed, pick it from an "oracle urn" which is another Hoppe's urn, with initial number γ of black balls and color distribution G_0, for simplicity assumed to be diffuse. The draws from the oracle Hoppe's urn represent the labels of the topics available for all documents; when colored, they are an exchangeable sequence with directing random measure $\tilde{G} \sim DP(\gamma, G_0)$. Conditionally on all the draws from the oracle urn (thus on $\tilde{G}$), the colored drawings from the document-specific Hoppe's urns $\mathcal{U}_j$ are independent exchangeable sequences $(\tilde{\theta}_{n,j})_{n\geq 1}$, with

$$\tilde{\theta}_{n,j} \mid \tilde{G}_j \overset{i.i.d.}{\sim} \tilde{G}_j\,, \quad n \geq 1,$$
$$\tilde{G}_j \mid \tilde{G} \overset{indep}{\sim} DP(\alpha_j, \tilde{G})\,, \quad j = 1, \ldots, M,$$

independently across j, with $\tilde{G} \sim DP(\gamma, G_0)$. This defines a Hierarchical Dirichlet Process prior for $(\tilde{G}_1, \ldots, \tilde{G}_M)$, with parameters $(\alpha_1, \ldots, \alpha_M, \gamma, G_0)$.

This model is based on an exchangeable structure at the latent stage, where (in line with the considerations in Example 3.13), one envisages an actual random partition. Differently from Example 3.9, here the draws from the Hoppe's urns are latent variables, since, at any time, an "old" color could be picked from $\mathcal{U}_j$ or from the oracle urn. This leads to computational challenges, as we discuss in the next section.

Extensions of the hierarchical Dirichlet process include the hierarchical Pitman-Yor process [119], and hierarchies of general discrete random measures leading to interesting combinatorial structures; see Camerlenghi *et al.* [20], and [22] and references therein. □

4.2 Asymptotic partial exchangeability

In the above example, and in fact more generally with partially exchangeable data, the predictive and the posterior distributions are not available in a "closed" (ideally, conjugate) analytic form - with a sometimes significant computational cost. To give some insight on the reasons for this difficulty, suppose for brevity that one only has two partially exchangeable sequences $(X_n)_{n\geq 1}, (Y_n)_{n\geq 1}$ and aims for an analytically tractable predictive distribution $\mathbb{P}(X_{n+1} \in \cdot \mid x_{1:n}, y_{1:n}, y_{n+1})$. Assuming $X_{n+1} \perp\!\!\!\perp Y_{n+1} \mid X_1, \ldots, X_n, Y_1, \ldots, Y_n$ would help, but typically breaks partial exchangeability, except for trivial cases. On another extreme, a functionally simple inclusion of Y_{n+1} in the expression of the predictive distribution above may create *direct* dependence between the two sequences, and rather give *interacting* stochastic processes (see e.g. [3] and references therein). In fact, in partially exchangeable constructions one typically identifies a conditional independence structure of the kind $X_{n+1} \perp\!\!\!\perp Y_{n+1} \mid X_1, \ldots, X_n, Y_1, \ldots, Y_n, U$ where U is a latent random variable (in nonparametric settings with discrete priors, U is an appropriate feature of the random partition, see e.g. [20]). While this may allow approximation schemes, for example through Gibbs sampling ([120], [85], [21]), integrating out the latent U to obtain the predictive distribution $\mathbb{P}(X_{n+1} \in \cdot \mid x_{1:n}, y_{1:n}, y_{n+1})$ is not, generally, analytically manageable.

Although there has been a sensible effort to find "closed form" expressions for predictive distributions for partially exchangeable models, the above considerations highlight that it is not easy to have partial exchangeability *and* also an analytically tractable predictive rule. This raises interest for predictive structures that only preserve partial exchangeability asymptotically, but are computationally easier. [60] have proposed the notion of *partially conditionally identically distributed* (partially c.i.d.) sequences, which is equivalent to partial exchangeability for stationary data and preserves main properties of partially exchangeable sequences. In particular, partially c.i.d. processes are asymptotically partially exchangeable. Natural extensions of reinforced stochastic processes turn out to be partially c.i.d. For example, consider a family of sequences $[X_{n,j}]_{n\geq 1, j=1,\ldots,M}$ such that $\mathbb{P}(X_{1,j} \in \cdot) = P_{0,j}(\cdot)$ and for any $n \geq 1$

$$\mathbb{P}(X_{n+1,j} \in \cdot \mid (X_{k,i})_{k\leq n, i=1,\ldots,M}) = \frac{\alpha_{0,j} P_{0,j}(\cdot) + \sum_{k=1}^{n} W_{k,j} \delta_{X_{k,j}}(\cdot)}{\alpha_{0,j} + \sum_{k=1}^{n} W_{k,j}},$$

where the random weights $W_{k,j}$ are positive r.v.'s and may be functions of the observed values of the other sequences. It is proved in [60] that if, conditionally on $(X_{k,j}, W_{k,j})_{k\leq n, j\leq M}$, the future observations $(X_{n+1,j})_{j\leq M}$ are mutually independent and $W_{n,j}$ is independent of $X_{n,j}$, $j = 1, \ldots, M$, then the sequences $[X_{n,j}]_{n\geq 1, j=1,\ldots,M}$ are partially c.i.d.

4.3 Markov exchangeability

The representation theorem 2.2 for exchangeable sequences gives the conceptual justification of the Bayesian inferential setting for random sampling. A natural question is if there is a symmetry notion and a de Finetti-like representation theorem that justify the Bayesian inferential setting for Markov chains. In this section we recall the notion of Markov exchangeability [36] and its predictive characterization [55], and review Diaconis and Freedman's representation theorem and a different representation that relates Markov exchangeability to partial exchangeability in the sense of de Finetti. Many models, for instance state-space models for nonstationary time series, are based on Markov chains; thus, these results also give insights on predictive constructions for Bayesian learning with temporal data, beyond Markov chains.

Let $\mathbb{X}$ be a finite or countable set that includes at least two points, and $(X_n)_{n\geq 0}$ be a sequence of r.v.'s taking values in $\mathbb{X}$, and with probability law $\mathbb{P}$. The process $(X_n)_{n\geq 0}$ is *partially exchangeable in the sense of*

Diaconis and Freedman, or, following the terminology of [134] and [131], *Markov exchangeable*, if its probability law is invariant under finite permutations that do not alter the number of transitions between any two states; more precisely, if $\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) = \mathbb{P}(X_0 = x'_0, \dots, X_n = x'_n)$ whenever $(x_0, \dots, x_n)$ and $(x'_0, \dots, x'_n)$ have the same initial value (i.e. $x_0 = x'_0$) and exhibit the same number of transitions from state i to state j, for every $i, j \in \mathbb{X}$.

Under a recurrence condition, Diaconis and Freedman prove a de Finetti-like representation theorem for Markov exchangeable sequences. The process $(X_n)_{n\geq 0}$ is recurrent if the initial state x_0 is visited infinitely many times with probability one. Let us also define, for any $i, j \in \mathbb{X}$, the transition counts $T_{i,j}^{(n)}$ as the number of transitions from state i to state j in $(X_0, \dots, X_n)$, and the matrix of normalized transition counts as the matrix with elements $\hat{T}_{i,j}^{(n)} = T_{i,j}^{(n)} / \sum_{k\in\mathbb{X}} T_{i,k}^{(n)}$ if the sum is different from zero, and zero otherwise.

THEOREM 4.6 (Diaconis and Freedman [36], Theorem 7 and Remark 25). *Suppose that the process $(X_n)_{n\geq 0}$, starting at x_0, is recurrent. If $(X_n)_{n\geq 0}$ is Markov exchangeable, then*

i) *With probability one, the matrix of normalized transition counts converges (in the topology of coordinate convergence) to a random limit $\tilde{Q}$;*
ii) *conditionally on $\tilde{Q}$, the process $(X_n)_{n\geq 0}$ is a Markov chain with transition matrix $\tilde{Q}$.*

In applications in Bayesian statistics, the probability law of the random limit $\tilde{Q}$, which is uniquely determined by $\mathbb{P}$, plays the role of the prior.

The following result gives a predictive characterization of Markov exchangeable processes; it parallels Theorems 2.3 and 4.4.

THEOREM 4.7 ([55]). *A predictive rule $(P_n)_{n\geq 0}$ for a process $(X_n)_{n\geq 0}$ with $X_0 = x_0$ and $X_n \in \mathbb{X}$ characterizes a Markov exchangeable process if and only if for every $n \geq 0$ and every $(x_1, \dots, x_n)$ the following conditions hold:*

i) *For every $j \in \mathbb{X}$, $\mathbb{P}(X_{n+1} = j \mid x_{0:n})$ depends on $x_{0:n}$ only through x_0 and its transition counts;*
ii) *For every $k \geq 1$ and all strings $\mathbf{y}$, $\mathbf{y}'$ and $\mathbf{z}$ of elements in $\mathbb{X}$ that do not contain x_n and have no common elements, the function that maps $(\mathbf{y}, \mathbf{y}')$ into $\mathbb{P}((X_{n+1}, \dots, X_{n+k}) = (\mathbf{y}, \mathbf{z}, x_n, \mathbf{y}', \mathbf{z}, x_n) \mid x_{0:n})$ is symmetric in $(\mathbf{y}, \mathbf{y}')$.*

This is proved in [55], where a predictive condition for recurrence is also given. The characterization becomes much simpler when the predictive distribution of X_{n+1} only depends on the last visited state x_n and on the x_nth row $\mathbf{t}_{x_n}$ of the matrix of transition counts: $\mathbb{P}(X_{n+1}=y \mid x_{0:n})=p(y|x_n, \mathbf{t}_{x_n})$. In this case, $(X_n)_{n\geq 0}$ is Markov exchangeable if and only if

$$p(y\,|x, \mathbf{t})p(z\,|x, \mathbf{t} + \mathbf{e}_y) = p(z\,|x, \mathbf{t})p(y\,|x, \mathbf{t} + \mathbf{e}_z) \tag{4.3}$$

for every $\mathbf{t}$ and every $x, y, z \in \mathbb{X}$, where $\mathbf{e}_y$ and $\mathbf{e}_z$ have a 1 at positions y and z, respectively, and 0 elsewhere.

EXAMPLE 4.8 (*Reinforced urn scheme)*. Let $\mathbb{X}$ be finite or countable, and let $(X_n)_{n\geq 0}$ satisfy $X_0 = x_0$ and

$$\mathbb{P}(X_{n+1} = y \mid x_{0:n}) = \frac{\alpha_{x_n} q_{x_n}(y) + t_{x_n,y}}{\alpha_{x_n} + \sum_{j\in\mathbb{X}} t_{x_n,j}}, \tag{4.4}$$

where for every x, α_x is a positive number and $q_x(\cdot)$ is a probability mass function on $\mathbb{X}$. It is easy to verify that the predictive rule (4.4) satisfies (4.3). Moreover, by the Lévy extension of the Borel-Cantelli lemma, the state x_0 is visited infinitely many times (see [55] for the details). Hence, $(X_n)_{n\geq 0}$ is a mixture of Markov chains.

For a finite state space, Zabell [131] derived the predictive rule (4.4) from Johnson's sufficiency postulate and assuming that $(X_n)_{n\geq 0}$ is recurrent and Markov exchangeable, characterizing independent Dirichlet prior distributions on the rows of the random transition matrix. □

By the result i) in Theorem 4.6, the random transition matrix $\tilde{Q}$ has an empirical meaning as the limit of the matrix of normalized transition counts. An interesting question is whether it also has an interpretation in terms of prediction, in the spirit of Proposition 2.4. We can show it does by leveraging on an alternative characterization of Markov exchangeable processes, hinted in [32] and [131] and developed in [51], in terms of partial exchangeability of the matrix of successor states.

The nth successor state of a state x is defined as the state visited by the process $(X_n)_{n\geq 0}$ just after the nth visit to state x. Denoting by $\tau_n(x)$ the time of the nth visit to state x, with $\tau_n(x) = \infty$ if x is not visited n times, we can define the nth successor state of x as

$$S_{x,n} = X_{\tau_n(x)+1}$$

if $\tau_n(x)$ is finite. Let us collect the successor states for all x in an array $[S_{x,n}]_{x\in\mathbb{X},n\geq 1}$. Note that the xth row of $[S_{x,n}]$ has infinite length if the state x is visited infinitely many times, otherwise it is of finite length. The set of states that are visited infinitely many times depends on the path ω, so does the length of the row $(S_{x,n})_n$. To avoid rows of finite length, [51] enlarge the state space, by adding an external point ∂, and define $S_{x,n}(\omega) = \partial$ if $\tau_n(x)(\omega) = \infty$.

It is proved in [51] that $(X_n)_{n\geq 0}$ is a mixture of recurrent Markov chains if and only if the array of successor states $[S_{x,n}]_{x\in\mathbb{X},n\geq 1}$ is partially exchangeable by rows in

the sense of de Finetti, (see Theorem 1 in [51] for more details and for the extension to uncountable state spaces). This allows us to use the results in Section 4.1 for the array $[S_{x,n}]_{x\in\mathbb{X},n\geq 1}$ of successor states. For each x, the xth row $(S_{x,n})_{n\geq 1}$ is exchangeable, thus the successors $S_{x,n}$ of state x are conditionally i.i.d. given a random probability mass function $\tilde{Q}_x$ on $\mathbb{X}$. The probability masses $\tilde{Q}_{x,i} \equiv \tilde{Q}_x(i)$ are the limits of the empirical frequencies $\sum_{k=1}^{n} \delta_{S_{x,k}}(i)/n$, $i \in \mathbb{X}$, and correspond to the xth row of the random transition matrix $\tilde{Q}$. Partial exchangeability also implies that the rows of the array of successor states are not independent sequences: probabilistic dependence across them is introduced through the joint prior law of the vector $(\tilde{Q}_x, x \in \mathbb{X})$, i.e. of the rows of the random transition matrix $\tilde{Q}$.

Moreover, the random transition matrix is the limit of the predictive distributions, in the sense that, for all x, i,

$$\lim_n \mathbb{P}(S_{x,n+1} = i \mid S_{x,1}, \ldots, S_{x,n}, V)$$
$$= \lim_n \mathbb{P}(S_{x,n+1} = i \mid S_{x,1}, \ldots, S_{x,n}) = \tilde{Q}_{x,i}$$

where V collects all the rows of the matrix of successors states but the xth. This result refers to the successor states. In terms of the sequence $(X_n)_{n\geq 0}$, see Theorem 1 in [55].

Stochastic processes with reinforcement are again powerful tools in predictive constructions of Markov exchangeable sequences. An elegant construction, through *edge reinforced random walks* on a graph, is Diaconis and Rolles' [40] characterization of a conjugate prior for the transition matrix of a reversible Markov chain. Developments for variable order reversible Markov chains are in [6]. The reinforced urn schemes in the following examples could also be read in terms of reinforced random walk on a graph (by associating urns to the vertices).

EXAMPLE 4.9 (*Reinforced Hoppe urn processes).* The predictive rule of Example 4.8 was obtained by [53] through a class of 'reinforced Hoppe urn processes', that includes other constructions in the literature as special cases. Let the sample space (or color space) be finite or countable. To each $x \in \mathbb{X}$, associate a Hoppe urn $\mathcal{U}_x$, with α_x black balls and discrete color distribution q_x on $\mathbb{X}$. Balls are extracted from each urn by Hoppe sampling as in Example 3.9, but we now move across urns as follows. Pick x_0 from an initial distribution q on $\mathbb{X}$, set $X_0 = x_0$, go to urn U_{x_0} and pick a ball from it. Since the ball will be black, a color x_1 is sampled from q_{x_0} and a ball of color x_1 is added in the urn, together with the black ball. Set $X_1 = x_1$ and move to Hoppe urn U_{x_1}, and proceed simlarly. Let $(X_n)_{n\geq 0}$ be the process so obtained. In this construction, the draws from the state-specific Hoppe urns $\mathcal{U}_x$ represent the successors of state x and are Pólya sequences, independent across x; thus, the process $(X_n)_{n\geq 0}$ is Markov exchangeable. Under mild conditions it is also recurrent (see [55] for details). It follows that a recurrent reinforced Hoppe urn process is conditionally Markov, and the prior on the random transition matrix $\tilde{Q}$ is such that the rows of $\tilde{Q}$, regarded as random measures on the state space $\mathbb{X}$, are independent, with $\tilde{Q}_x \sim \mathrm{DP}(\alpha_x, q_x)$ (or Dirichlet distributions in the case of a finite state space).

As a special case, with a finite state space $\mathbb{X}$, suppose that, for each $x \in \mathbb{X}$, the color distribution q_x has finite support in $\mathbb{X}$; then the process $(X_n)_{n\geq 0}$ reduces to the *reinforced urn process* by [91].

If $\mathbb{X} = \{0, 1, 2, \ldots\}$ and for each $x \in \mathbb{X}$, the color distribution q_x of urn $\mathcal{U}_x$ has positive masses only on $x + 1$ and $x_0 = 0$, the process (X_n) corresponds to the reinforced urn process proposed by [126] for Bayesian survival analysis. In this case, the exchangeable sequence of the lengths of the x_0 blocks characterizes a novel *Beta-Stacy* prior can be used as a conjugate prior with exchangeable censored data. A version of this predictive construction allows a generalization of the finite population Bayesian bootstrap [86] to include censored observations [92]. □

A hierarchical version of the reinforced Hoppe urn process gives the popular *infinite hidden Markov model* proposed by Beal, Ghahramani and Rasmussen [9] for Bayesian learning in hidden Markov models with an *unbounded* number of states.

EXAMPLE 4.10 (*infinite Hidden Markov Model).* Suppose that the state space $(\theta_1^*, \theta_2^*, \ldots)$ is countable and *unknown*. In [9], this is the state space of the *latent state process* $(X_n)_{n\geq 0}$ of a hidden Markov model where a new state may be added as the need occurs. The authors construct $(X_n)_{n\geq 0}$ through a predictive scheme that again envisages a reinforced Hoppe urn process; but, differently from Example 4.9, and also from the construction in Example 4.5, here Hoppe's urns are created as a new state (color) is discovered; and colors are drawn when the need occurs from a common 'oracle' Hoppe urn with an initial number γ of black balls and diffuse color distribution P_0. The process starts by picking a ball from the oracle urn; since the ball will be black, a first color, say θ_1^*, is picked from P_0; and the black ball and an additional ball of color θ_1^* are returned in the oracle urn. Then one sets $X_0 = \theta_1^*$ and creates a Hoppe urn $\mathcal{U}_{\theta_1^*}$ with α black balls, picks a ball from it, and proceeds similarly. This generates a process $(X_n)_{n\geq 0}$ that is recurrent and Markov exchangeable; thus, there exist $\tilde{Q}$ conditionally on which $(X_n)_{n\geq 0}$ is a Markov chain with transition matrix $\tilde{Q}$, and the construction characterizes the prior law on $\tilde{Q}$. The draws from the oracle urn generate the states of the process, and are a Pólya sequence with directing random measure $\tilde{P} \sim \mathrm{DP}(\gamma, P_0)$. Conditionally on all the draws $(\theta_1^*, \theta_2^*, \ldots)$ from the oracle urns, thus on $\tilde{P} = P$,

the process $(X_n)_{n\geq 0}$ is a reinforced Hoppe urn process as in Example 4.9, with state space $(\theta_1^*, \theta_2^*, \ldots)$; thus, the rows of $\tilde{Q}$, regarded as random distributions on the state space $(\theta_1^*, \theta_2^*, \ldots)$, have independent DP$(\alpha, P)$ distributions. Therefore, the prior law on the rows of $\tilde{Q}$, regarded as random distributions, is a hierarchical Dirichlet process with parameters (α, γ, P_0). Here, the construction of the prior is purely predictive; the hierarchical Dirichlet process was introduced later [120].

The predictive distribution of X_{n+1} given $x_{0:n}$ is analytically complex; however, an advantage of the predictive construction is to allow for efficient computational strategies (see [61]). □

4.4 Row-column exchangeability

Many data are in the form of arrays, graphs, matrices, and forms of partial exchangeability are developed for general random structures. In this section we briefly review Aldous' notion of *row-column exchangeability*, or partial exchangeability for random arrays, and refer to Aldous [2] and Kallenberg [78] for extensive treatment. An excellent review paper that also includes Bayesian models for exchangeable random structures in statistics and machine learning is [96]. We do not even try to review the wide and growing literature on row-column exchangeable arrays, and related theory of exchangeable random graphs and more recent theory for sparse graphs. We just recall basic concepts and the analogue of de Finetti representation theorem for row-column exchangeable arrays. In the predictive perspective of this paper, it would be interesting to include basic properties of the predictive distributions, in analogy to Proposition 2.4 for exchangeable sequences. However, results of this nature for row-column exchangeable arrays seem lacking. To our knowledge, a first result, that relates the problem with de Finetti's concept of partial exchangeability and holds for a fairly general class of row-column exchangeable arrays, is given in unpublished work by [49].

In studying exchangeability, we have regarded the data $(X_1, \ldots, X_n)$ as elements of an infinite sequence $(X_n)_{n\geq 1}$. Similarly, here we consider an observed finite array $[X_{i,j}]_{i,j=1,\ldots,n}$ as a sub-array of an infinite random array $X = [X_{i,j}]_{i,j\geq 1}$; the $X_{i,j}$ are $\mathbb{X}$-valued random variables, where $\mathbb{X}$ is a Polish space.

DEFINITION 4.11. *An infinite random array $X = [X_{i,j}]_{i,j\geq 1}$ is* separately exchangeable *if*

$$X \stackrel{d}{=} [X_{\sigma_1(i),\sigma_2(j)}]_{i,j\geq 1} \tag{4.5}$$

for all finite permutations σ_1, σ_2 of $\mathbb{N}$. It is jointly exchangeable *if the above holds in the special case $\sigma_1 = \sigma_2$.*

Condition (4.5) is equivalent to requiring that the rows of X are exchangeable and the columns are exchangeable; it is thus referred to as *row-and-column exchangeability* (RCE). We will use the terminolgy *RCE array* to mean that the array is either separately or jointly exchangeable. Separate exchangeability is an appropriate assumption if rows and columns of the array correspond with two distinct sets of entities; for example, rows correspond to users and columns to movies. If there is a single set of entities, for example the vertices of a graph, one may require invariance under permutations of the entities, that is, joint exchangeability.

Binary jointly exchangeable arrays give a representation of exchangeable random graphs. A random infinite graph (with known vertices, labeled by $\mathbb{N}$, and random edges) is exchangeable if its probability law is invariant under every finite permutation of its vertices. Equivalently, if and only if the corresponding adjacency matrix $X = [X_{i,j}]_{i,j\geq 1}$, where $X_{i,j}$ is the indicator of there being an edge (i, j) in the graph, is jointly exchangeable. Actually, theoretical results for RCE arrays have been rediscovered in the developments of the limiting theory for large graphs initiated by Lovász and Szegedy [89]. The connection between graph limits and RCE arrays is given by Diaconis and Janson [39]. We refer to the monograph by Lovász [88] for the graph limit theory.

Proving a de Finetti-like representation for RCE arrays has been more delicate than expected. The representation theorem was independently given by Hoover [68] and Aldous [1] and developed more systematically by Kallenberg, culminating in his 2005 monograph [78]. The proof that appears in Aldous ([1]; see also [2], Theorem 14.11) uses the concept of 'coding'. The way this is used may be unfamiliar to some readers; to introduce it, note that de Finetti's representation theorem can be given (e.g. [2], page 129) as follows. A sequence of r.v.'s $(X_n)_{n\geq 1}$ is exchangeable if and only if there exists a measurable function $H : [0,1]^2 \to \mathbb{X}$ such that $(X_n)_{n\geq 1}$ can be *coded* through i.i.d. uniform r.v.'s $U, U_i, i \geq 1$ with a *representing function* H, that is, $(X_n)_{n\geq 1} \stackrel{d}{=} (H(U_n, U))_{n\geq 1}$. For example, binary r.v.'s $(X_n)_{n\geq 1}$ are exchangeable if and only if they can be generated by first picking θ from a prior law π (through $\theta = \pi^{-1}(U)$, where π^{-1} is the generalized inverse of the prior distribution function π), then sampling $X_i \stackrel{i.i.d.}{\sim}$ Bernoulli(θ) (through $X_i = \mathbf{1}_{(U_i\leq\theta)}$).

THEOREM 4.12 (Aldous-Hoover representation theorem for separately exchangeable arrays). *An infinite random array $X = [X_{i,j}]_{i,j\geq 1}$ is separately exchangeable if and only if there exists $H : [0,1]^4 \to \mathbb{X}$ such that X can be coded by i.i.d. Uniform$(0,1)$ independent r.v.'s $U; U_i, i \geq 1; V_j, j \geq 1, U_{i,j}, i, j \geq 1$, with representing function H, that is*

$$[X_{i,j}]_{i,j\geq 1} \stackrel{d}{=} [X^*_{i,j}]_{i,j\geq 1}, \text{where } X^*_{i,j} = H(U, U_i, V_j, U_{i,j}).$$

The natural statistical interpretation is that $X_{i,j}$ is determined by a row effect U_i, a column effect V_j, an individual effect $U_{i,j}$ and an overall effect U.

Binary arrays. To simplify, let us consider binary arrays. The representation theorem can be rephrased by saying that an infinite binary random array X is separately exchangeable if and only if there exists a probability measure π on the space of (measurable) functions from $[0,1]^2 \to [0,1]$ such that X can be generated as follows (the r.v.'s $U_i, V_i, U_{i,j}$ are as in the theorem). Each row i is assigned a latent feature U_i and each column j is assigned a feature V_j. Independently generate a function $W(\cdot,\cdot)$ from the probability distribution π (through the uniform r.v. U). Given the features assignment and W, set $X_{i,j} = 1$ with probability $W(U_i, V_j)$ (that is, $X_{i,j} = 1$ if $U_{i,j} \leq W(U_i, V_j)$). Note that if W is fixed (not picked from π), the resulting array $[X_{i,j}]_{i,j\geq 1}$ is separately exchangeable by the symmetry of the construction. Denote by P_W its probability law. Aldous-Hoover representation theorem proves that any separately exchangeable binary array is a mixture of such arrays.

THEOREM 4.13 (Aldous-Hoover; binary arrays). *Let $X = [X_{i,j}]_{i,j\geq 1}$ be an infinite separately exchangeable binary random array. Then, there is a probability distribution π such that*

$$\mathbb{P}(X \in \cdot) = \int P_W(\cdot) d\pi(W). \tag{4.6}$$

Borrowing from the language of random graphs, a (measurable) map $W : [0,1]^2 \to [0,1]$ is called a *graphon*. A graphon defines a probability law P_W as above, however this parametrization is not unique; in other words, in statistical sense, W is not identifiable. Indeed, if W' is obtained from W by a measure-preserving transformation of each variable, then clearly the associated process $[X'_{i,j}]_{i,j\geq 1}$ has the same joint distribution as $[X_{i,j}]_{i,j\geq 1}$. It has been proved that this is the only source of non-uniqueness [78]. A unique parametrization can be obtained by substituting the graphons W by equivalence classes. The results by Orbanz and Szegedy [97] imply that this parametrization is measurable. See [39] and [96] for a more extensive treatement.

For *jointly* exchangeable arrays, there is an analogous representation result as Theorem 4.12, with $X^*_{i,j} = H(U, U_i, U_j, U_{\{i,j\}})$, where $(U_i)_{i\geq 1}$ and $[U_{\{i,j\}}]_{i,j\geq 1}$ are, respectively, a sequence and an array of independent uniform r.v.'s and H is symmetric in (U_i, U_j); see [2], Theorem 14.21. Note that the indexes of the $U_{\{i,j\}}$ are unordered and the array $[U_{\{i,j\}}]$ may be thought of as an upper-triangular matrix with i.i.d. uniform entries.

Let us consider *binary* arrays; in particular, binary arrays representing the adjacency matrix of an infinite simple graph (undirected and with no multiple edges and self-loops), thus, $[X_{i,j}]_{i,j\geq 1}$ symmetric with a zero diagonal. A binary *jointly exchangeable* array can be constructed in a similar way as before, by now assigning features to vertices. Namely, each vertex $i \in \mathbb{N}$ is assigned a latent feature U_i, with $U_i \overset{i.i.d.}{\sim} \text{Uniform}(0,1)$; given the latent features and a graphon W, we set $X_{i,j} = 1$ with probability $W(U_i, U_j)$, independently for all i, j. The array $[X_{i,j}]_{i,j\geq 1}$ so constructed is *jointly* exchangeable by construction (and is symmetric if W is such). Denote by $P_W^{(joint)}$ its probability law. The Aldous-Hoover representation theorem shows that any binary jointly exchangeable array can be constructed as a mixture of these $P_W^{(joint)}$. In other words, if $[X_{i,j}]_{i,j\geq 1}$ is an infinite jointly exchangeable binary array, then conditionally on the features and on the graphon, the $X_{i,j}$ are independent Bernoulli($\tilde{\theta}_{i,j}$) where $\tilde{\theta}_{i,j} = W(U_i, U_j)$.

As given, the Aldous-Hoover theorem does not provide an empirical link for the elements of the representation. For exchangeable sequences, de Finetti's representation theorem is complemented by a law of large numbers, that gives an empirical meaning to the random directing measure $\tilde{F}$ as the limit of the sequence of empirical distributions; moreover, $\tilde{F}$ is also the limit of the predictive distributions P_n (see Proposition 2.4). For RCE arrays, the notion of an empirical distribution and a law of large numbers are given by Kallenberg [76], Theorem 3. The asymptotic theory is also thoroughly explained in [96]. Instead, no result seems available on convergence of predictive distributions, that relate the Aldous-Hoover representation to prediction. To our knowledge, a first result is given in [49]; but we do not expand this further here.

5. RECURSIVE ALGORITHMS AND PREDICTIONS

The predictive approach has been shown to be powerful in many contexts. A last but important (to us) point we want to make in this paper is that a Bayesian predictive approach can also be taken in less 'classic' contexts, in particular to evaluate predictive algorithms, possibly arising from other fields, in order to obtain better awareness of their implicit assumptions and provide probabilistic quantification of uncertainty. We discuss this point for two recursive procedures. The first example is from [56]; the second one is new.

Recursive computations are particularly convenient in sequential learning from streaming data, where it is crucial to have predictions that can be quickly updated as new observations become available, at a constant computational cost and with limited storage of information; and recursive procedures have been developed since at least the work of Kalman [79]. Recent directions in a Bayesian predictive approach include, among others, [66], [50] and [14]. In Sections 5.1 and 5.2 below, we examine two

recursive algorithms for prediction with streaming data; and, in line with the principles at the basis of this paper, exposed in the Introduction, we show how they can be read as Bayesian predictive learning rules (although not exchangeable), unveiling the implied statistical model and obtaining Bayesian uncertainty quantification. In the examples, the implied model is asymptotically exchangeable, thus for n large the 'algorithm' provides a computationally simple approximation of an exchangeable Bayesian procedure.

In some more detail, we can read the algorithms as particular cases of a broad class of Bayesian recursive predictive rules of the following form: $X_1 \sim P_0$ and for every $n \geq 1$ $X_{n+1} \mid X_1, \ldots, X_n \sim P_n$, with

$$\begin{cases} P_n = p_n(T_n), \\ T_n = h_n(T_{n-1}, X_n), \end{cases} \tag{5.1}$$

where p_n and h_n are given functions, and T_n is a predictive sufficient summary (Sect. 3.2) of $X_1, \ldots, X_n$. The form of P_n allows storage of only the sufficient summaries and straightforward updating. Suitable specifications lead to desirable properties for the sequence $(X_n)_{n\geq 1}$ (in the examples, asymptotic exchangeability).

In an exchangeable parametric setting, many common models, for example the Beta-Bernoulli scheme, have a recursive rule of the form (5.1). In a nonparametric setting, this holds for Pólya sequences, whose predictive rule (3.5) can be written recursively as

$$P_n = \frac{\alpha + n - 1}{\alpha + n} P_{n-1} + \frac{1}{\alpha + n} \delta_{X_n}.$$

In this case $T_n \equiv P_n$, and the recursive rule applies directly to the predictive distributions. This extends to other discrete nonparametric schemes; it is however quite delicate in the continuous case. Here, a general class of sequences $(X_n)_{n\geq 1}$ that satisfy (5.1) are *measure-valued Pólya sequences* (MVPS; [112]), characterized by

$$P_n(\cdot) = \frac{\gamma P_0(\cdot) + \sum_{i=1}^n R_{X_i}(\cdot)}{\gamma + \sum_{i=1}^n R_{X_i}(\mathbb{X})}, \quad n \geq 1,$$

where R is a non-null finite transition kernel on the sample space $\mathbb{X}$ and γ is a positive constant. Letting $\mu_0(\cdot) = \gamma P_0(\cdot)$, we can write the above predictive distributions as in (5.1), with

$$\begin{cases} P_n(\cdot) = \dfrac{\mu_n(\cdot)}{\mu_n(\mathbb{X})} \\ \mu_n(\cdot) = \mu_{n-1}(\cdot) + R_{X_n}(\cdot). \end{cases}$$

The predictive sufficient statistic T_n in (5.1) is, in this case, the random measure μ_n, which is updated by simply adding the random measure R_{X_n} to μ_{n-1}. This scheme extends the Hoppe's urn characterization of Pólya sequences shown in Example 3.9: any set of colors B in $\mathbb{X}$ has initially mass $\mu_0(B)$; then, at each step n, the mass of B is reinforced with a mass $R_{x_n}(B)$. As proved in [112], a measure-valued Pólya sequence $(X_n)_{n\geq 1}$ is exchangeable if and only if it coincides with a kernel-based Dirichlet sequence; unfortunately, as seen in Example 3.13, the latter seems quite limited for statistical applications; and so are *exchangeable* specifications of MVPS. Moreover, natural extensions, for example allowing for random reinforcement (see e.g. [59], [111]) also require to go beyond exchangeability. Indeed, MVPS can be *asymptotically* exchangeable. The procedure we consider in the next section 5.1 will be shown to be an asymptotically exchangeable generalized MVPS.

Remark. As seen in Section 2, asymptotic exchangeability holds for c.i.d. sequences. A class of recursive predictive rules that, under mild assumptions, meets the c.i.d. condition, is presented in [12], (Sect. 4.1, Eqn (5)). Although this class is rather general, not all the predictive rules of the form (5.1) - in particular, not those arising in the following sections - are included in it. We need the generality and the predictive features of the class (5.1).

5.1 Newton's algorithm and recursive prediction in mixture models

Michael Newton and collaborators ([94], [95], [93]) proposed a recursive procedure for unsupervised sequential learning in mixture models, that extends an earlier proposal by Smith and Makov [118] and is referred to as the *Newton's algorithm* in the Bayesian nonparametric literature. Let $(X_n)_{n\geq 1}$ be a sequence of r.v.'s taking values in $\mathbb{X} \subseteq \mathbb{R}^d$, and consider a mixture model

$$X_i \mid G \overset{i.i.d.}{\sim} f_G(x) \equiv \int k(x \mid \theta) dG(\theta),$$

where $k(x \mid \theta)$, $\theta \in \Theta \subseteq \mathbb{R}^k$, is a kernel density of known parametric form, and G is the unknown mixing distribution. Let us assume that the mixture model is identifiable. The Newton's algorithm estimates G starting from an initial guess G_0 and recursively updating the estimate, as $x_1, x_2, \ldots$ become available, as

$$G_n(\cdot) = (1 - \alpha_n) G_{n-1}(\cdot) + \alpha_n G_{n-1}(\cdot \mid x_n), \tag{5.2}$$

where α_n and $G_{n-1}(\cdot \mid x_n)$ are as described in Example 3.13 of Section 3, and a simple choice for α_n is $\alpha_n = 1/(\alpha + n)$ for some $\alpha > 0$. At step n, the algorithm returns $G_n(\cdot)$ as the estimate of $G(\cdot)$.

This recursive procedure was suggested as a simple and computationally fast approximation of the intractable Bayesian solution in a Dirichlet process mixture model. In the latter, the mixing distribution is random and is assigned a DP(αG_0) prior; then the prior guess is $E(\tilde{G}(\cdot)) = G_0(\cdot)$, and the first update, based on x_1, gives the Bayesian estimate

$$E(\tilde{G}(\cdot) \mid x_1) = \frac{\alpha}{\alpha + 1} G_0(\cdot) + \frac{1}{\alpha + 1} G_0(\cdot \mid x_1).$$

Newton's algorithm (5.2) replicates the same updating form for any $n > 1$. The resulting estimate G_n deviates from the Bayesian solution $E(\tilde{G}(\cdot) \mid x_{1:n})$, but is computationally much simpler; and, in practice, may give a surprisingly good approximation. Based on G_n, one can also obtain a plug-in estimate of the mixture density f_G, as $f_{G_n}(x) = \int k(x \mid \theta) dG_n(\theta)$. Again, this differs from the Bayesian density estimate $E(f_G(x) \mid x_{1:n})$ in the Dirichlet process mixture model. Note that $E(f_G(x) \mid x_{1:n})$ is also the predictive density of X_{n+1} given $x_{1:n}$. Our point is thus that Newton's algorithm is using a different learning rule, namely the predictive density

$$X_{n+1} \mid x_{1:n} \sim f_{G_n}(x) = \int k(x \mid \theta) dG_n(\theta), \tag{5.3}$$

with G_n as in (5.2). This is of the form (5.1), with sufficient statistic $T_n = G_n$; and gives a generalized measure-valued Pólya sequence. Because, as seen in Section 2.1, the predictive rule characterizes the probability law of the process $(X_n)_{n\geq 1}$, reading the algorithm as a *probabilistic* predictive rule allows us to reveal the probability law that the researcher is implicitly assuming for the process. It is easy to see that (5.3) characterizes a probability law $\mathbb{P}$ for $(X_n)_{n\geq 1}$ that is no longer exchangeable. However, one still has $X_{n+2} \mid x_{1:n} \stackrel{d}{=} X_{n+1} \mid x_{1:n}$. Thus, the sequence $(X_n)_{n\geq 1}$ is c.i.d. (see section 2.3), therefore asymptotically exchangeable. Actually, [56] prove stronger results: the asymptotic directing random measure of $(X_n)_{n\geq 1}$ has precisely density $f_{\tilde{G}}(x) = \int k(x \mid \theta) d\tilde{G}(\theta)$, where, $\mathbb{P}$-a.s, the random distribution $\tilde{G}$ is the limit of the sequence G_n, and $f_{\tilde{G}}$ is the limit in L^1 of the predictive density $f_{G_n}(x)$.

The above results imply that for $n \geq N$ large

$$X_n \mid \tilde{G} \stackrel{i.i.d.}{\approx} f_{\tilde{G}},$$

with a novel prior on the random mixing distribution $\tilde{G}$, that, interestingly, can select absolutely continuous distributions [56]; but is not known explicitly. However, one can sample from it, through the 'sampling from the future' algorithm described in Section 2.4. Moreover, in the same spirit as in Proposition 2.6, but referring to the mixing distribution, one can obtain an asymptotic Gaussian approximation of the posterior distribution of $[\tilde{G}(t_1), \ldots, \tilde{G}(t_M)]$ given $x_{1:n}$. Our predictive methodology also allows to naturally obtain principled extensions, that otherwise would mostly be heuristic; see again [56].

5.2 Online gradient descent and prediction

Consider the problem of classifying items as 'type 0' or 'type 1' based on a d-dimensional vector of features, for example through a neural network or a generalized linear model. Let the items arrive sequentially, and, for every $n \geq 1$, let Y_n and X_n represent the 'type' and features of the nth item, respectively. Typically, the relationship between X_n and Y_n is modelled through $\mathbb{P}(Y_n = 1 \mid x_n) = g(x_n, \beta)$, where g is a known function and β is an unknown d-dimensional parameter, and the X_i are assumed to be i.i.d. from a distribution (known or unknown) P_X. Given a sample, or 'training set', $(x_i, y_i)_{i=1,\ldots,n}$, an estimate of the parameter β can be obtained by minimizing, with respect to β, a loss function $L(\beta; x_1, y_1, \ldots, x_n, y_n)$ measuring the difference between the actual values of $y_1, \ldots, y_n$ and the ones predicted by the model. While efficient algorithms exist to solve this optimization problem, the computational cost becomes substantial when β is high-dimensional. Additionally, if data arrive sequentially, the process must be restarted from scratch with each new data point. In this context, β can be estimated by an *online learning* [117] procedure, based on the stochastic approximation [110] of the gradient descent dynamic: β is initialised at time zero as β_0 (which can be random or deterministic), and then updated, at each new data (x_n, y_n), by "moving" it along the direction that minimizes $L(\beta_{n-1}; x_n, y_n)$, (that is opposite to the direction of the gradient with respect to β_{n-1}):

$$\beta_n = \beta_{n-1} - \frac{1}{n} \nabla_\beta L(\beta_{n-1}; x_n, y_n). \tag{5.4}$$

Although the results of this section hold for other choices of L (for example quadratic loss) and g, here we consider the typical case of binary cross entropy loss $L(\beta; x_1, y_1, \ldots, x_n, y_n) = -\sum_{i=1}^n [y_i \log_2(g(x_i, \beta)) + (1 - y_i) \log_2(1 - g(x_i, \beta))]$ and logistic function

$$g(x, \beta) = \frac{e^{x^T \beta}}{1 + e^{x^T \beta}}. \tag{5.5}$$

In this case, (5.4) becomes

$$\beta_n = \beta_{n-1} + \frac{1}{n \log 2} (y_n - g(x_n, \beta_{n-1})) x_n. \tag{5.6}$$

If P_X is known, we can reinterpret the algorithm as a Bayesian predictive learning rule of the form (5.1), where $T_n = \beta_n$ is updated at each new observation (x_n, y_n) as in (5.6), and we assume that, for $y = 0, 1$,

$$\begin{aligned} &\mathbb{P}(X_{n+1} \in dx, Y_{n+1} = y | x_{1:n}, y_{1:n}) \\ &\quad = g(x, \beta_n)^y (1 - g(x, \beta_n))^{1-y} P_X(dx), \end{aligned} \tag{5.7}$$

with g as in (5.5). In fact, the assumption that P_X is known (or has been estimated separately) is only instrumental for the theoretical results; our final result does not require to know P_X.

This predictive rule is not consistent with exchangeability of the sequence $((X_n, Y_n))_{n\geq 1}$; however, under mild assumptions, exchangeability holds asymptotically, as shown in the following proposition. All the proofs are in Section S5 of the Supplement [58].

PROPOSITION 5.1. *Let $((X_n, Y_n))_{n\geq 1}$ have probability law $\mathbb{P}$ characterized by the predictive rule* (5.6)-(5.7)*, where g is given by* (5.5)*, $E(||\beta_0||^2) < \infty$, and P_X has bounded support. Then:*

i) *The sequence of random vectors* $(\beta_n)_{n\geq 0}$ *converges* $\mathbb{P}$*-a.s. to a random limit* $\tilde{\beta}$ *and, for every* $n \geq 0$, $\beta_n = E(\tilde{\beta} \mid \beta_0, X_1, Y_1, \ldots, X_n, Y_n)$*;*

ii) *The sequence of random vectors* $((X_n, Y_n))_{n\geq 1}$ *is* $\tilde{P}_{X,Y}$*-asymptotically exchangeable, with the random measure* $\tilde{P}_{X,Y}$ *such that the conditional distribution* $\tilde{P}_{Y|X=x}$ *of* Y *given* $X = x$ *is Bernoulli*$(g(x, \tilde{\beta}))$*.*

Informally, this implies that, for n large,

$$Y_n \mid \tilde{\beta}, x_n \overset{indep}{\approx} \text{Bernoulli}(g(x_n, \tilde{\beta})).$$

The posterior distribution of the random vector $\tilde{\beta}$ remains unknown. However, for n large, it can be approximated by a multivariate Normal distribution centered in β_n.

PROPOSITION 5.2. *Under the assumptions of Proposition 5.1, with* P_X *being non-degenerate on any linear subspace of* $\mathbb{R}^d$*, the conditional distribution of* $\sqrt{n}(\tilde{\beta} - \beta_n)$*, given* $\beta_0, X_1, Y_1, \ldots, X_n, Y_n$*, converges* $\mathbb{P}$*-a.s., as* $n \to \infty$*, to a multivariate Normal distribution with mean zero and random covariance matrix*

$$U = (\log 2)^{-2} \int xx^T g(x, \tilde{\beta})(1 - g(x, \tilde{\beta}))P_X(dx). \tag{5.8}$$

The random matrix U, that depends on the unknown parameter $\tilde{\beta}$, can be approximated by replacing $\tilde{\beta}$ with β_n. Thus, for n large $\tilde{\beta} \mid x_{1:n}, y_{1:n} \approx \mathcal{N}_d(\beta_n, U_n/n)$, with $U_n = (\log 2)^{-2} \int xx^T g(x, \beta_n)(1 - g(x, \beta_n))dP_X(x)$.

The following alternative approximation of U does not require to know P_X.

PROPOSITION 5.3. *Under the assumptions of Proposition 5.2, as* $n \to \infty$*,*

i) *The statistic* $V_n = \frac{1}{n}\sum_{k=1}^{n} k^2(\beta_k - \beta_{k-1})(\beta_k - \beta_{k-1})^T$ *converges* $\mathbb{P}$*-a.s. to the random matrix* U *in* (5.8)*;*

ii) *The conditional distribution of* $\sqrt{n}V_n^{-1/2}(\tilde{\beta} - \beta_n)$*, given* $\beta_0, X_1, Y_1, \ldots, X_n, Y_n$ *converges* $\mathbb{P}$*-a.s. to the standard multivariate Normal distribution.*

Thus, for n large, for $\mathbb{P}$-almost all sample paths,

$$\tilde{\beta} \mid x_{1:n}, y_{1:n} \approx \mathcal{N}_d(\beta_n, V_n/n),$$

which can be used, in particular, to provide asymptotic credible sets.

Remark. The proofs of Propositions 5.1, 5.2 and 5.3 are based on a key martingale property of the sequence $(\beta_n)_{n\geq 0}$. These results can be generalized to other algorithms as long as the martingale property holds and certain moment bounds are met. Although our techniques are not directly applicable without the martingale property, extending the Bayesian interpretation beyond martingale-based learning appears feasible, since many algorithms are based on stochastic approximations with well-understood limit theorems and convergence rates. Also, computational strategies such as Approximate Bayesian Computation or Variational Bayes might be read as using a predictive learning rule whose properties could be studied in our predictive approach.

6. FINAL REMARKS

We have offered a review, from foundations to some recent directions, of principles and methods for Bayesian predictive modeling; and of course a lot could not be covered. We barely mentioned that prediction is not, in fact, the ultimate goal, but the basis for decisions to be taken under risk. Also, the paper is, somehow unavoidably, mostly theoretical, aiming at discussing fundamental concepts; but a predictive approach involves the perspective we adopt in inference and in any statistical problem, with evident practical implications; ultimately, the basic principle is that, differently from inferential conclusions, predictions can be checked with facts.

ACKNOWLEDGMENTS

We thank the three reviewers for their valuable comments, and are grateful to Sara Wade for interesting suggestions. Both authors acknowledge funding by the European Union grant 'Next Generation EU Funds, PRIN 2022 (2022CLTYP4)'.

SUPPLEMENTARY MATERIAL

Supplement to "Exchangeability, prediction and predictive modeling in Bayesian statistics". The Supplement collects the proofs for the results in the paper.

REFERENCES

[1] ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivar. Anal.* **11** 581-598.

[2] ALDOUS, D. J. (1985). Exchangeability and related topics. *École d'Été de Probabilités de Saint-Flour XIII 1983* **1117** 1-198.

[3] ALETTI, G., CRIMALDI, I. and GHIGLIETTI, A. (2023). Interacting innovation processes. *Nature, Scientific Reports* **13**.

[4] ANTONIAK, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2** 1152–1174.

[5] AUGENBLICK, N. and RABIN, M. (2021). Belief Movement, Uncertainty Reduction, and Rational Updating. *Q. J. Econ.* **136** 933–985.

[6] BACALLADO, S. (2011). Bayesian analysis of variable-order, reversible Markov chains. *Ann. Statist.* **39** 838–864.

[7] BACALLADO, S., BATTISTON, M., FAVARO, S. and TRIPPA, L. (2017). Sufficientness Postulates for Gibbs-Type Priors and Hierarchical Generalizations. *Statist. Sci.* **32** 487-500.

[8] BACALLADO, S., FAVARO, S., POWER, S. and TRIPPA, L. (2022). Perfect Sampling of the Posterior in the Hierarchical Pitman–Yor Process. *Bayesian Anal.* **17** 685–709.

[9] BEAL, M. J., GHAHRAMANI, Z. and RASMUSSEN, C. E. (2002). The infinite Hidden Markov Model. In *Mach. Learn.* 239–245. MIT Press.

[10] BERNARDO, J. M. and SMITH, A. F. M. (1994). *Bayesian Theory*. Wiley, Chichester, England.

[11] BERTI, P., CRIMALDI, I., PRATELLI, L. and RIGO, P. (2015). Central limit theorems for an Indian buffet model with random weights. *Ann. Appl. Probab.* **25** 523 – 547.

[12] BERTI, P., DREASSI, E., LEISEN, F., PRATELLI, L. and RIGO, P. (2023). A Probabilistic View on Predictive Constructions for Bayesian Learning. *Statist. Sci., Advance Publication* 1-15.

[13] BERTI, P., DREASSI, E., LEISEN, F., PRATELLI, L. and RIGO, P. (2023). Kernel based Dirichlet sequences. *Bernoulli* **29** 1321–1342.

[14] BERTI, P., DREASSI, E., LEISEN, F., PRATELLI, L. and RIGO, P. (2023). Bayesian predictive inference without a prior. *Statist. Sinica* **33** 2405–2429.

[15] BERTI, P., PRATELLI, L. and RIGO, P. (2004). Limit theorems for a class of identically distributed random variables. *Ann. Probab.* **32** 2029-2052.

[16] BERTI, P., PRATELLI, L. and RIGO, P. (2013). Exchangeable sequences driven by an absolutely continuous random measure. *Ann. Probab.* **41** 2090–2102.

[17] BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353-355.

[18] BREIMAN, L. (2001). Statistical Modeling: The Two Cultures (with discussion). *Statist. Sci.* **16** 199-231.

[19] CAMERLENGHI, F., FAVARO, S., MASOERO, L. and BRODERICK, T. (2024). Scaled Process Priors for Bayesian Nonparametric Estimation of the Unseen Genetic Variation. *J. Amer. Statist. Assoc.* **119** 320–331.

[20] CAMERLENGHI, F., LIJOI, A., ORBANZ, P. and PRÜNSTER, I. (2019). Distribution theory for hierarchical processes. *Ann. Statist.* **47** 67–92.

[21] CAMERLENGHI, F., LIJOI, A. and PRÜNSTER, I. (2017). Bayesian prediction with multiple-samples information. *J. Multivar. Anal.* **156** 18-28.

[22] CATALANO, M., SOLE, C. D., LIJOI, A. and PRÜNSTER, I. (2023). A Unified Approach to Hierarchical Random Measures. *Sankhya A*.

[23] CIFARELLI, D. M. and REGAZZINI, E. (1996). De Finetti's contribution to probability and statistics. *Statist. Sci.* **11** 253 – 282.

[24] CLARKE, B. S. and CLARKE, J. L. (2018). *Predictive Statistics: Analysis and Inference beyond Models*. Cambridge University Press, Cambridge.

[25] COPPERSMITH, D. and DIACONIS, P. (1986). Random walk with reinforcement. *Unpublished manuscript*.

[26] CRIMALDI, I. (2009). An almost sure conditional convergence result and an application to a generalized Pólya urn. *Int. Math. Forum* **4** 1139-1156.

[27] DAWID, A. P. (1984). Present Position and Potential Developments: Some Personal Views. Statistical Theory. The Prequential Approach (with Discussion). *J. Roy. Statist. Soc. Ser. A* **147** 278–292.

[28] DE BLASI, P., FAVARO, S., LIJOI, A., MENA, R. H., PRÜNSTER, I. and RUGGIERO, M. (2015). Are Gibbs-type priors the most natural beneralization of the Dirichlet Process? *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 212–229.

[29] DE FINETTI, B. (1931). Sul significato soggettivo della probabilità. *Fundamenta Mathematicae* **17** 298-329.

[30] DE FINETTI, B. (1937). La prévision, ses lois logiques, ses sources subjectives. *Ann. Henri Poincaré* **7** 1-68.

[31] DE FINETTI, B. (1937). Sur la condition de "equivalence partielle". In *Colloque consacré á la théorie des probabilités, Université de Genéve, 12-16 octobre, 1937* 5–18. Hermann et C.ie, Paris, l938-39, Actual. sci. industr. 730, vol. VI [English translation: On the Condition of Partial Exchangeability, in *Studies in Inductive Logic and Probability*, vol.2 (1980) p.193-205].

[32] DE FINETTI, B. (1959). La probabilità e la statistica nei rapporti con l'induzione secondo i diversi punti di vista. In *Centro Internazionale Matematico estivo (CIME), Induzione e Statistica* 1–115. Cremones [English translation in B. de Finetti (1972) *Probability, Induction and Statistics*. Wiley, New York, 147–227.].

[33] DE FINETTI, B. (1970). *Teoria delle Probabilità*. Einaudi, Turin Italy. English translation: Theory of Probability by Antonio Machi and Adrian Smith, Wiley (London, England) Vol. 1 1974, Vol. 2 1975.

[34] DEGROOT, M. H. and FIENBERG, S. E. (1982). The comparison and evaluation of forecasters. *The Statistician* **32** 12-22.

[35] DIACONIS, P. (1988). Recent progress on de Finetti's notions of exchangeability. In *Bayesian Statistics 3* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.) Oxford University Press, Oxford, Oxford, UK.

[36] DIACONIS, P. and FREEDMAN, D. (1980). de Finetti theorem for Markov chains. *Ann. Probab.* **8** 115–130.

[37] DIACONIS, P. and FREEDMAN, D. (1984). Partial exchangeability and sufficiency. In *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference, Calcutta, 1981* (J. K. Ghosh and J. Roy, eds.) 205–274.

[38] DIACONIS, P. and FREEDMAN, D. (1990). On the Uniform Consistency of Bayes Estimates for Multinomial Probabilities. *Ann. Statist.* **18** 1317-1327.

[39] DIACONIS, P. and JANSON, S. (2008). Graphs limits and exchangeable random graphs. *Rendiconti di Matematica, Serie VII* **28** 33–61.

[40] DIACONIS, P. and ROLLES, S. W. W. (2006). Bayesian analysis for reversible Markov chains. *Ann. Statist.* **34** 1270–1292.

[41] DIACONIS, P. and YLVISAKER, D. (1979). Conjugate Priors for Exponential Families. *Ann. Statist.* **7** 269–281.

[42] DOKSUM, K. (1974). Tailfree and Neutral Random Probabilities and Their Posterior Distributions. *Ann. Probab.* **2** 183-201.

[43] DOLERA, E. and REGAZZINI, E. (2019). Uniform rates of the Glivenko–Cantelli convergence and their use in approximating Bayesian inferences. *Bernoulli* **25** 2982 – 3015.

[44] DOOB, J. L. (1949). Application of the theory of martingales. *Le Calcul de Probabilités et ses Applications, Colloques Internationaux du Centre National de la Recherche Scientifique* 23-27.

[45] DUBINS, L. E. and SAVAGE, L. J. (1965). *How to Gamble if You Must. Inequalities for Stochastic Processes*. McGraw-Hill, New York.

[46] EFRON, B. and HASTIE, T. (2016). *Computer Age Statistical Inference. Algorithms, Evidence and Data Science*. Cambridge University Press.

[47] EGGENBERGER, F. and PÓLYA, G. (1923). Über die Statistik verketteter Vorgänge. *Z. Angew. Math. Mech. (Appl. Math. Mech.)* **3** 279–289.

[48] EWENS, W. J. (1972). The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3** 87–112.

[49] FASANO, A. and PETRONE, S. (2022). A predictive approach to Aldous-Hoover representation theorem for exchangeable random graphs. Manuscript.

[50] FONG, E., HOLMES, C. and WALKER, S. G. (2023). Martingale Posterior Distributions. *https://imstat.org/journals-and-publications/statistical-science/statistical-science-resources-for-authors/* **85** 1357–1391, with discussion.

[51] FORTINI, S., LADELLI, L., PETRIS, G. and REGAZZINI, E. (2002). On mixtures of distributions of Markov chains. *Stoch. Process. Appl.* **100** 147-165.

[52] FORTINI, S., LADELLI, L. and REGAZZINI, E. (2000). Exchangeability, predictive distributions and parametric models. *Sankya, Series A* **62** 86–109.

[53] FORTINI, S. and PETRONE, S. (2012). Hierarchical reinforced urn processes. *Stat. Probab. Lett.* **82** 1521-1529.

[54] FORTINI, S. and PETRONE, S. (2016). *Predictive Distribution (de Finetti's View)* In *Wiley StatsRef: Statistics Reference Online* 1-9. John Wiley & Sons, Ltd.

[55] FORTINI, S. and PETRONE, S. (2017). Predictive characterization of mixtures of Markov chains. *Bernoulli* **23** 1538 – 1565.

[56] FORTINI, S. and PETRONE, S. (2020). Quasi-Bayes properties of a procedure for sequential learning in mixture models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **82** 1087-1114.

[57] FORTINI, S. and PETRONE, S. (2023). Prediction-based uncertainty quantification for exchangeable sequences. *Phil. Trans. R. Soc. A* **381:20220142**.

[58] FORTINI, S. and PETRONE, S. (2024). Supplement to "Exchangeability, prediction and predictive modeling in Bayesian statistics". doi:.

[59] FORTINI, S., PETRONE, S. and SARIEV, H. (2021). Predictive constructions based on measure-valued Pólya urn processes. *Mathematics* **9** 2845.

[60] FORTINI, S., PETRONE, S. and SPORYSHEVA, P. (2018). On a notion of partially conditionally identically distributed sequences. *Stoch. Process. Appl.* **128** 819–846.

[61] GAEL, J. V. and GHAHRAMANI, Z. (2011). *Nonparametric hidden Markov models* In *Bayesian Time Series Models* 317–340. Cambridge University Press.

[62] GEISSER, S. (1993). *Predictive Inference*. Chapman and Hall/CRC, London, UK.

[63] GOLDSTEIN, M. (1999). Bayes Linear Analysis. In *Encyclopedia of Statistical Sciences.*, (S. Kotz, C. B. Read and D. L. Banks, eds.) **3** 29-34. Wiley, New York.

[64] GOOD, I. J. (1967). A Bayesian Significance Test for Multinomial Distributions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **29** 399–431. (with discussion).

[65] GRIFFITHS, T. L. and GHAHRAMANI, Z. (2005). Infinite latent feature models and the Indian Buffet Process. In *Advances in Neural Information Processing Systems (NIPS)* 475–482.

[66] HAHN, P. R., MARTIN, R. and WALKER, S. G. (2018). On recursive Bayesian predictive distributions. *J. Amer. Statist. Assoc.* **113** 1085–1093.

[67] HOLMES, C. C. and WALKER, S. G. (2023). Statistical inference with exchangeability and martingales. *Philos. Trans. R. Soc. A* **381** 1-17.

[68] HOOVER, D. N. (1979). Relations on probability spaces and arrays of random variables. Technical Report, Institute of Advanced Study, Princeton.

[69] HOPPE, F. M. (1984). Pólya-like urns and the Ewens' sampling formula. *J. Math. Biol.* **20** 91-94.

[70] HOPPE, F. M. (1987). The sampling theory of neutral alleles and an urn model in population genetics. *J. Math. Biol.* **25** 123-159.

[71] HOROWITZ, J. (1985). Measure-valued random processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **70** 213–236.

[72] ISHWARAN, H. and JAMES, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *J. Amer. Statist. Assoc.* **96** 161–173.

[73] JAMES, L. F. (2017). Bayesian Poisson calculus for latent feature modeling via generalized Indian Buffet Process priors. *Ann. Statist.* **45** 2016 – 2045.

[74] JOHNSON, W. E. (1932). Probability: The Deductive and Inductive Problems. *Mind* **41** 409–423.

[75] KALLENBERG, O. (1988). Spreading and predictable sampling in exchangeable sequences and processes. *Ann. Probab.* **16** 508-534.

[76] KALLENBERG, O. (1999). Multivariate sampling and the estimation problem for exchangeable arrays. *J. Theoret. Probab.* **12** 859–883.

[77] KALLENBERG, O. (2002). *Foundations of modern probability*, second ed. *Probability and its Applications (New York)*. Springer-Verlag, New York.

[78] KALLENBERG, O. (2005). *Probabilistic Symmetries and Invariance Principles*. Springer, New York.

[79] KALMAN, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng.* **82** 35-45.

[80] KINGMAN, J. F. C. (1978). Uses of exchangeability. *Ann. Probab.* 183–197.

[81] KINGMAN, J. F. C. (1980). *The Mathematics of Genetic Diversity*. SIAM, Philadelphia.

[82] KINGMAN, J. F. C. (1978). The representation of partition structures. *J. London Math. Soc.* **2** 374–380.

[83] LAURITZEN, S. L. (1984). Extreme point models in statistics (with discussion). *Scand. J. Stat.* **11** 65–91.

[84] LAURITZEN, S. L. (1988). *Extremal Families and Systems of Sufficient Statistics. Lecture Notes in Statistics 49*. Springer, New York.

[85] LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20**.

[86] LO, A. Y. (1988). A Bayesian Bootstrap for a Finite Population. *Ann. Statist.* **16** 1684–1695.

[87] LO, A. Y. (1991). A characterization of the Dirichlet process. *Stat. Probab. Lett.* **12** 185-187.

[88] LOVÁSZ, L. (2013). *Large Networks and Graph Limits*. American Mathematical Society.

[89] LOVÁSZ, L. and SZEGEDY, B. (2006). Limits of dense graph sequences. *J. Combin. Theory Ser. B* **96** 933–957.

[90] MOYA, B. and WALKER, S. G. (2023). Martingale Posterior Distributions for Time Series Models. *Statist. Sci., Advance Publication* 1-15.

[91] MULIERE, P., SECCHI, P. and WALKER, S. G. (2000). Urn schemes and reinforced random walks. *Stoch. Process. Their Appl.* **88** 59–78.

[92] MULIERE, P. and WALKER, S. (1998). Extending the Family of Bayesian Bootstraps and Exchangeable Urn Schemes. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 175–182.

[93] NEWTON, M. A. (2002). On a nonparametric recursive estimator of the mixing distribution. *Sankyā* **64** 306–322.

[94] NEWTON, M. A., QUINTANA, F. A. and ZHANG, Y. (1998). *Nonparametric Bayes methods using predictive updating* In *Practical Nonparametric and Semiparametric Bayesian Statistics* 45–61. Springer, New York.

[95] NEWTON, M. A. and ZHANG, Y. (1999). A recursive algorithm for nonparametric analysis with missing data. *Biometrika* **86** 15–26.

[96] ORBANZ, P. and ROY, D. M. (2015). Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures. *IEEE Transactions of Pattern Analysis and Machine Intelligence* **37** 437–461.

[97] ORBANZ, P. and SZEGEDY, B. (2016). Borel liftings of graph limits. *Electron. Commun. Probab.* **21** 1-4.

[98] PARMIGIANI, G. and INOUE, L. (2009). *Decision Theory: Principles and Approaches*. John Wiley & Sons,Chichester, UK.

[99] PEMANTLE, R. (2007). A survey on random processes with reinforcement. *Probab. Surv.* **4** 1–79.

[100] PERMAN, M., PITMAN, J. and YOR, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields* **92** 21—39.

[101] PETRONE, S. (1999). Random Bernstein Polynomials. *Scand. J. Stat.* **26** 373–393.

[102] PETRONE, S. and VERONESE, P. (2010). Feller operators and mixture priors in Bayesian nonparametrics. *Statist. Sinica* **20** 379–404.

[103] PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* **102** 145–158.

[104] PITMAN, J. (1996). Some developments of the Blackwell–MacQueen urn scheme. In *Statistics, Probability and Game Theory. Papers in honor of David Blackwell*, (T. S. Ferguson, L. S. Shapley and J. B. MacQueen, eds.). *IMS Lecture Notes - Monograph Series* **30** 245–267. Institute of Mathematical Statistics, Hayward, California.

[105] PITMAN, J. (2002). *Combinatorial Stochastic Processes*. *Ecole d'Eté de Probabilités de Saint-Flour, Lecture Notes in Mathematics* **XXXII**. Springer-Verlag.

[106] PITMAN, J. and YOR, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25** 855–900.

[107] PÓLYA, G. (1931). Sur quelques points de la théorie des probabilités. *Ann. Inst. H. Poincaré* **1** 117–161.

[108] RAMSEY, F. P. (1926). Truth and Probability. In *Philosophy of Probability: Contemporary Readings.* (A. Eagle, ed.) 52–94. Routledge.

[109] REGAZZINI, E. (1999). Old and Recent Results on the Relationship Between Predictive Inference and Statistical Modelling either in Nonparametric or Parametric Form. In *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting June 6-10, 1998*. Oxford Universiy Press.

[110] ROBBINS, H. and MONRO, S. (1951). A Stochastic Approximation Method. *Ann. Math. Stat.* **22** 400 – 407.

[111] SARIEV, H., FORTINI, S. and PETRONE, S. (2023). Infinite-color randomly reinforced urns with dominant colors. *Bernoulli* **29** 132 – 152.

[112] SARIEV, H. and SAVOV, M. (2024). Characterization of exchangeable measure-valued Pólya urn sequences. *Electron. J. Probab.* **29** 1 – 23.

[113] SARIEV, H. and SAVOV, M. (2025). Sufficientness postulates for measure-valued Pólya urn sequences. *J. Appl. Probab.* In press.

[114] SAVAGE, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.

[115] SCHERVISH, M. J. (1995). *Theory of Statistics*. Springer-Verlag, New York.

[116] SHMUELI, G. (2010). To Explain or to Predict? *Statist. Sci.* **25** 289-310.

[117] SMALE, S. and YAO, Y. (2006). Online Learning Algorithms. *Foundations of Computational Mathematics* **6** 145-170.

[118] SMITH, A. F. M. and MAKOV, U. E. (1978). A quasi-Bayes sequential procedure for mixtures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **40** 106–112.

[119] TEH, Y. W. (2006). A hierarchical Bayesian language model based on Pitman–Yor processes. In *Proceedings of the International Conference on Computational Linguistics* 985-992.

[120] TEH, Y. W., JORDAN, M. I., BEAL, M. J. and BLEI, D. M. (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101** 1566–1581.

[121] TEH, Y. W. and JORDAN, M. I. (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics* (N. L. Hjort, C. C. Holmes, P. Muller and S. G. Walker, eds.) 158–207. Cambridge University Press.

[122] THIBAUX, R. and JORDAN, M. I. (2007). Hierarchical Beta Processes and the Indian Buffet Process. In *AISTATS Proceedings,* 564–571.

[123] VON PLATO, J. (1982). The Generalization of de Finetti's Representation Theorem to Stationary Probabilities. In *Philosophy of Science Association: Proceedings of the Biennial Meeting of the Philosophy of Science* 137-144. The University of Chicago Press.

[124] WADE, S. K., MONGELLUZZO, S. and PETRONE, S. (2011). An enriched conjugate prior for Bayesian nonparametric inference. *Bayesian Anal.* **6** 359–499.

[125] WADE, S., WALKER, S. G. and PETRONE, S. (2014). A predictive study of Dirichlet process mixture models for curve fitting. *Scand. J. Stat.* **41** 580–605.

[126] WALKER, S. and MULIERE, P. (1997). Beta-Stacy processes and a generalization of the Pólya-urn scheme. *Ann. Statist.* **25** 1762–1780.

[127] WALKER, S. and MULIERE, P. (1999). A characterization of a neutral to the right prior via an extension of Johnson's sufficientness postulate. *Ann. Statist.* **27** 415–781.

[128] WILLIAMS, P. M. (1976). *Indeterminate Probabilities* In *Formal Methods in the Methodology of Empirical Sciences: Proceedings of the Conference for Formal Methods in the Methodology of Empirical Sciences, Warsaw, June 17–21, 1974* 229–246. Springer Netherlands, Dordrecht.

[129] YIU, A., FONG, E., WALKER, S. and HOLMES, C. (2022). Causal predictive inference and target trial emulation. *arXiv preprint arXiv: 2207.12479*.

[130] ZABELL, S. L. (1982). W. E. Johnson's "Sufficientness" Postulate. *Ann. Statist.* **10** 1090–1099.

[131] ZABELL, S. L. (1995). Characterizing Markov exchangeable sequences. *J. Theor. Probab.* **8** 175–178.

[132] ZABELL, S. L. (1997). *The continuum of inductive methods revisited* In *The Cosmos of Science: Essays in Exploration*. Earman, J. and Norton, J.D. Eds. Universty of Pittsburgh Press.

[133] ZABELL, S. L. (2005). *Symmetry and Its Discontents: Essays on the History of Inductive Probability*. Cambridge Univ. Press, New York.

[134] ZAMAN, A. (1984). Urn Models for Markov Exchangeability. *Ann. Probab.* **12** 223–229.