A probabilistic view on predictive constructions for Bayesian learning

A Probabilistic View on Predictive Constructions for Bayesian Learning

Patrizia Berti, Emanuela Dreassi, Fabrizio Leisen, Luca Pratelli and Pietro Rigo

Abstract. Given a sequence $X = (X_1, X_2, \ldots)$ of random observations, a Bayesian forecaster aims to predict X_{n+1} based on $(X_1, \ldots, X_n)$ for each $n \geq 0$. To this end, in principle, she only needs to select a collection $\sigma = (\sigma_0, \sigma_1, \ldots)$, called "strategy" in what follows, where $\sigma_0(\cdot) = P(X_1 \in \cdot)$ is the marginal distribution of X_1 and $\sigma_n(\cdot) = P(X_{n+1} \in \cdot | X_1, \ldots, X_n)$ the nth predictive distribution. Because of the Ionescu–Tulcea theorem, σ can be assigned directly, without passing through the usual prior/posterior scheme. One main advantage is that no prior probability is to be selected. In a nutshell, this is the predictive approach to Bayesian learning. A concise review of the latter is provided in this paper. We try to put such an approach in the right framework, to make clear a few misunderstandings, and to provide a unifying view. Some recent results are discussed as well. In addition, some new strategies are introduced and the corresponding distribution of the data sequence X is determined. The strategies concern generalized Pólya urns, random change points, covariates and stationary sequences.

Key words and phrases: Bayesian inference, conditional identity in distribution, exchangeability, predictive distribution, sequential predictions, stationarity.

1. INTRODUCTION

This paper has been written having the following interpretation of Bayesian inference in mind. (We declare this interpretation from the outset just to make transparent our point of view and easier the understanding of the paper). Let us call $\mathcal{O}$ the object of inference. Roughly speaking, $\mathcal{O}$ denotes whatever we ignore but would like to know. For instance, $\mathcal{O}$ could be a parameter (finite or infinite dimensional), a set of future observations, an unknown probability distribution, the effect of some action, or something else. According to us, the distinguishing feature of the Bayesian approach is to regard $\mathcal{O}$ as the realization of a random element, and not as an unknown but fixed constant. As a consequence, the main goal of any Bayesian inferential procedure is to determine the conditional distribution of $\mathcal{O}$ given the available information.

Note that, unless $\mathcal{O}$ itself is a parameter, no other parameter is necessarily involved.

Prediction of unknown observable quantities is a fundamental part of statistics. Initially, it was probably the most prevalent form of statistical inference. The wind changed at the beginning of the 20th century when statisticians' attention shifted to other issues, such as parametric estimation and testing; see, for example, [36]. Nowadays, prediction is back in the limelight again, and plays a role in modern topics including machine learning and data mining; see, for example, [17, 18, 27, 43].

This paper deals with prediction of future observations, based on the past ones, from the Bayesian point of view. Precisely, we focus on a sequence

$$X = (X_1, X_2, \ldots)$$

of random observations and, at each time n, we aim to predict X_{n+1} based on $(X_1, \ldots, X_n)$. Hence, for each n,

Patrizia Berti is Professor of Probability, Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Università di Modena e Reggio-Emilia, via Campi 213/B, 41100 Modena, Italy (e-mail: patrizia.berti@unimore.it). Emanuela Dreassi is Professor of Statistics, Dipartimento di Statistica, Informatica, Applicazioni, Università di Firenze, viale Morgagni 59, 50134 Firenze, Italy (e-mail: emanuela.dreassi@unifi.it). Fabrizio Leisen is Professor of Statistics, School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK (e-mail: fabrizio.leisen@gmail.com). Luca Pratelli is Professor of Probability, Accademia Navale, viale Italia 72, 57100 Livorno, Italy (e-mail: pratel@mail.dm.unipi.it). Pietro Rigo is Professor of Probability, Dipartimento di Scienze Statistiche "P. Fortunati," Università di Bologna, via delle Belle Arti 41, 40126 Bologna, Italy (e-mail: pietro.rigo@unibo.it).

the object of inference is $\mathcal{O} = X_{n+1}$, the available information is $(X_1, \ldots, X_n)$, and the target is the *predictive distribution* $P(X_{n+1} \in \cdot | X_1, \ldots, X_n)$. We point out that, apart from technicalities, most of our considerations could be generalized to the case where $\mathcal{O}$ is an arbitrary (measurable) function of the future observations, say

$$\mathcal{O} = f(X_{n+1}, X_{n+2}, \ldots).$$

This case is recently object of increasing attention; see, for example, [29, 40].

No parameter θ plays a role at this stage. The forecaster may involve some θ, if she thinks it helps, but she is not interested in θ as such. To involve θ means to model the probability distribution of X as depending on θ, and then to exploit this fact to calculate the predictive distributions $P(X_{n+1} \in \cdot | X_1, \ldots, X_n)$.

To better address our prediction problem, it is convenient to introduce the notion of *strategy*. Let $(S, \mathcal{B})$ be a measurable space, with S to be viewed as the set where the observations X_n take values. Following Dubins and Savage [26], a strategy is a sequence

$$\sigma = (\sigma_0, \sigma_1, \sigma_2, \ldots)$$

such that

- σ_0 and $\sigma_n(x)$ are probability measures on $\mathcal{B}$ for all $n \geq 1$ and $x \in S^n$;
- The map $x \mapsto \sigma_n(x, A)$ is $\mathcal{B}^n$-measurable for fixed $n \geq 1$ and $A \in \mathcal{B}$.

Here, σ_0 should be regarded as the marginal distribution of X_1 and $\sigma_n(x)$ as the conditional distribution of X_{n+1} given that $(X_1, \ldots, X_n) = x$. Moreover, $\sigma_n(x, A)$ denotes the value taken at A by the probability measure $\sigma_n(x)$. We also note that strategies are often called *prediction rules* in the framework of species sampling sequences; see [54], p. 251.

Strategies are a natural tool to frame a prediction problem from the Bayesian standpoint. In fact, a strategy σ can be regarded as the collection of all predictive distributions (including the marginal distribution of X_1) in the sense that $\sigma_n(x, \cdot) = P(X_{n+1} \in \cdot | (X_1, \ldots, X_n) = x)$ for all $n \geq 0$ and $x \in S^n$. Thus, in a sense, everything a Bayesian forecaster has to do is to select a strategy σ. Obviously, the problem is how to do it. A related problem is whether, in order to choose σ, involving a parameter θ is convenient or not.

An important special case is exchangeability. In fact, if X is assumed to be exchangeable, there is natural way to involve a parameter θ. To see this, take the parameter space Θ as

$$\Theta = \{\text{all probability measures on } \mathcal{B}\}.$$

Moreover, for each $\theta \in \Theta$, denote by P_θ a probability measure which makes X i.i.d. with common distribution θ, that is,

$$P_\theta(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \theta(A_i)$$

for all $n \geq 1$ and $A_1, \ldots, A_n \in \mathcal{B}$. Then, under mild conditions on $(S, \mathcal{B})$, de Finetti's theorem yields

$$P(X \in \cdot) = \int_\Theta P_\theta(X \in \cdot)\pi(d\theta)$$

for some (unique) prior probability π on Θ. Thus, conditionally on $\theta \in \Theta$, the observations are i.i.d. with common distribution θ. This suggests calculating the strategy σ as follows:

(i) Select a prior π on Θ;
(ii) For each $n \geq 1$ and $x \in S^n$, evaluate the posterior of θ given x, namely, the conditional distribution of θ given that $(X_1, \ldots, X_n) = x$;
(iii) Calculate σ as

$$\sigma_n(x, A) = \int_\Theta \theta(A)\pi_n(d\theta | x) \quad \text{for all } A \in \mathcal{B},$$

where $\pi_n(\cdot | x)$ is the posterior and $\pi_0(\cdot | x)$ is meant as $\pi_0(\cdot | x) = \pi$.

Steps (i), (ii) and (iii) are familiar in a Bayesian framework. Henceforth, if σ is selected via (i), (ii) and (iii), the forecaster is said to follow the *inferential approach* (I.A.).

1.1 Predictive Approach to Bayesian Modeling

There is another approach to Bayesian prediction, usually called the *predictive approach* (P.A.), which is quite recurrent in the Bayesian literature and recently gained increasing attention. (Such an approach, incidentally, has been referred to as the "nonstandard approach" in [5, 6]). According to P.A., the forecaster directly selects her strategy σ. Merely, for each $n \geq 0$, she selects the predictive σ_n without passing through the prior/posterior scheme described above. Among others, P.A. is supported by de Finetti, Savage, Dubins [22, 23, 26] and more recently by Diaconis and Regazzini [10, 16, 24, 25, 31]. P.A. is also strictly connected to Dawid's prequential approach [19–21] and to Pitman's treatment of species sampling sequences [54–56]. In addition, several prediction procedures arising in nonnecessarily Bayesian frameworks, such as machine learning and data mining, are consistent with P.A.; see, for example, [17, 18, 27, 43]. Some further related references are [5, 6, 29, 30, 32, 40, 41, 44].

The theoretical foundation of P.A. is the Ionescu–Tulcea theorem; see, for example, [46], p. 159. Roughly speaking, this theorem states that, to assign the joint distribution of X, it suffices to choose, *in an arbitrary way*, the marginal distribution of X_1, the conditional distribution of X_2 given X_1, the conditional distribution of X_3 given (X_1, X_2), and so on. Note that this fact would be obvious if X would be replaced by a finite-dimensional random

vector $(X_1, \ldots, X_m)$. So, in a sense, the Ionescu–Tulcea theorem extends to infinite sequences a straightforward property of finite-dimensional vectors. In any case, a formal statement of the theorem is as follows.

THEOREM 1 (Ionescu–Tulcea). *For each $n \geq 1$, let X_n be the nth coordinate random variable on $(S^\infty, \mathcal{B}^\infty)$. Then, for any strategy σ, there is a unique probability measure P_σ on $(S^\infty, \mathcal{B}^\infty)$ such that*

$$P_\sigma(X_1 \in \cdot) = \sigma_0(\cdot) \quad \text{and}$$
$$P_\sigma\big(X_{n+1} \in \cdot \,|\, (X_1, \ldots, X_n) = x\big) = \sigma_n(x, \cdot) \tag{1}$$

for all $n \geq 1$ and P_σ-almost all $x \in S^n$.

Because of Theorem 1, to make predictions on the sequence X, the forecaster is free to select an arbitrary strategy σ. In fact, for any σ, there is a (unique) probability distribution for X, denoted above by P_σ, whose predictives $P_\sigma(X_{n+1} \in \cdot\,|X_1, \ldots, X_n)$ agree with σ in the sense of equation (1).

The strengths and weaknesses of I.A. versus P.A. are discussed in a number of papers; see, for example, [6, 18, 27, 36, 58] and references therein. Here, we summarize this issue (from our point of view) under the assumption that prediction is the main target.

I.A. is not motivated by prediction alone. The main goal of I.A. is to make inference on other features of the data distribution (typically some parameters) and in this case the prior π is fundamental. It should be added that π often provides various meaningful information on the data generating process. However, to assess π is not an easy task. In addition, once π is selected, to evaluate the posterior $\pi_n(\cdot|x)$ is quite difficult as well. Frequently, $\pi_n(\cdot|x)$ cannot be written in closed form but only approximated numerically. In short, I.A. is a cornerstone of Bayesian inference, but when prediction is the main target, it is actually quite involved.

In turn, P.A. has essentially four merits. First, P.A. allows to avoid an explicit choice of the prior π. Indeed, when prediction is the main target, why select π explicitly? Rather than wondering about π, it seems reasonable to reflect on how the information in $(X_1, \ldots, X_n)$ is conveyed in the prediction of X_{n+1}. Second, the data sequence X is not required any distributional assumption. This point is developed in Sections 1.2 and 1.3. By now, we stress a consequence of such a point. The Bayesian nature of a prediction procedure does not depend on the data distribution. For instance, a forecaster applying P.A. is certainly Bayesian independently of the distribution attached to X. Third, P.A. requires the assignment of probabilities on observable facts only. The value of X_{n+1} is actually observable, while π and π_n (being probabilities on Θ) do not necessarily deal with observable facts. Fourth, the strategy σ may be assigned stepwise. At each time n, the forecaster has observed $x = (x_1, \ldots, x_n) \in S^n$ and has already selected $\sigma_0, \sigma_1(x_1), \ldots, \sigma_{n-1}(x_1, \ldots, x_{n-1})$. Then, to predict X_{n+1}, she is still free to select $\sigma_n(x)$ as she wants. No choice of $\sigma_n(x)$ is precluded. This is consistent with the Bayesian view, where the observed data are fixed and one should condition on them. In spite of these advantages, P.A. has an obvious drawback. In fact, assigning a strategy σ directly may be very difficult, in principle as difficult as selecting a prior π.

A last (basic) remark is that, if X is exchangeable, both I.A. and P.A. completely determine the probability distribution of X. Selecting a prior π or choosing a strategy σ are just equivalent routes to fix the distribution of X. In particular, selecting σ uniquely determines π. An intriguing line of research is in fact to identify the prior corresponding to a given σ; see, for example, [4, 24, 25, 31].

1.2 Characterizations

Recall that, for any strategy σ, there is a unique probability measure P_σ on $(S^\infty, \mathcal{B}^\infty)$ satisfying condition (1).

In principle, when applying P.A., the data sequence X is free to have any probability distribution. Nevertheless, in most applications, it is reasonable (if not mandatory) to impose some conditions on X. For instance, the forecaster may wish X to be exchangeable, or stationary, or Markov or a martingale, and so on. In these cases, σ is subjected to some constraints. If X is required to be exchangeable, for instance, σ should be such that P_σ is exchangeable. Hence, those strategies σ which make P_σ exchangeable should be characterized.

More generally, fix any collection $\mathcal{C}$ of probability measures on $(S^\infty, \mathcal{B}^\infty)$ and suppose the data distribution is required to belong to $\mathcal{C}$. Then P.A. gives rise to the following problem:

Problem (*): Characterize those strategies σ such that $P_\sigma \in \mathcal{C}$.

Sometimes, Problem (*) is trivial (Markov, martingales) but sometimes it is not (stationarity, exchangeability). To illustrate, we mention three examples (which correspond to the three dependence forms examined in the sequel).

In the exchangeable case, Problem (*) admits a solution [31], Theorem 3.1, but the conditions on σ are quite hard to check in real problems. Hence, applying P.A. to exchangeable data is usually difficult (even if there are some exceptions; see Section 2).

A condition weaker than exchangeability is conditional identity in distribution. Say that X is *conditionally identically distributed* (c.i.d.) if $X_2 \overset{d}{=} X_1$ and, for each $n \geq 1$, the conditional distribution of X_k given $(X_1, \ldots, X_n)$ is the same for all $k > n$; see Section 3. It can be shown that

$$X \text{ is exchangeable} \quad \Leftrightarrow \quad X \text{ is stationary and c.i.d.};$$

see [7, 47]. Hence, conditional identity in distribution can be regarded as one of the two basic ingredients of exchangeability (the other being stationarity). Now, in the

c.i.d. case, Problem (*) has been solved [8], Theorem 3.1, and the conditions on σ are quite simple. The class of admissible strategies includes several meaningful elements which cannot be used if X is required to be exchangeable. As a consequence, P.A. works quite well for c.i.d. data; see [5, 6].

The stationary case is more involved. In fact, to our knowledge, there is no general characterization of the strategies σ which make P_σ stationary. However, such a characterization is available in some meaningful special cases (for instance, when P_σ is also required to be Markov); see Section 4.

Finally, Problem (*) is usually easier in a few (meaningful) special cases. For instance, Problem (*) is simpler if P_σ is also asked to be Markov; see, for example, [33] and Section 4. Or else, if the strategy σ is required to be dominated.

Dominated strategies: Let λ be a σ-finite measure on $(S, \mathcal{B})$. Say that a strategy σ is dominated by λ if each $\sigma_n(x)$ admits a density $f_n(\cdot|x)$ with respect to λ, namely,

$$\sigma_0(dy) = f_0(y)\lambda(dy) \quad \text{and}$$
$$\sigma_n(x, dy) = f_n(y|x)\lambda(dy)$$

for all $n \geq 1$ and $x \in S^n$. Here, $f_0 : S \to \mathbb{R}^+$ and $f_n : S \times S^n \to \mathbb{R}^+$ are nonnegative measurable functions.

For instance, if $S = \mathbb{R}$ and $\sigma_n(x)$ is a nondegenerate normal distribution for all n and x, then σ is dominated by $\lambda =$ Lebesgue measure. Or else, if S is countable, any strategy is dominated by $\lambda =$ counting measure. Instead, if S is uncountable, a nondominated strategy is $\sigma_n(x_1, \ldots, x_n) = \delta_{x_n}$ where δ_{x_n} denotes the unit mass at the point x_n. Another nondominated strategy is the empirical measure $\sigma_n(x_1, \ldots, x_n) = (1/n)\sum_{i=1}^n \delta_{x_i}$.

In a sense, dominated strategies play an analogous role to the usual dominated models in parametric statistical inference. The main advantage is that one can use the conditional density $f_n(\cdot|x)$ instead of the conditional measure $\sigma_n(x)$. A related advantage is that, if one fixes λ and restricts to strategies dominated by λ, Problem (*) becomes simpler. However, even in applied data analysis, various familiar strategies are not dominated. In the framework of species sampling sequences, for instance, most strategies are not dominated. Therefore, in this paper, we focus on general strategies while the dominated ones are regarded as an important special case.

1.3 Content of This Paper and Further Notation

This is a review paper on P.A., which also includes some (minor) new results. Our perspective is mainly on the probabilistic aspects of Bayesian predictive constructions. Moreover, we tacitly assume that the major target is to predict future observations (and not to make inference on other random elements, such as random parameters).

Essentially, we aim to achieve three goals. First, we try to put P.A. in the right framework, to provide a unifying view, and to make clear a few misunderstandings. This has been done in the Introduction. Second, in Section 2 and Section 3.1, we report some known results. Third, we provide some new strategies and we prove a few related results. The strategies, introduced by means of examples, deal with generalized Pólya urns, random change points, covariates and stationary sequences. The results consist in determining the distribution of the data sequence X under such strategies. To our knowledge, Examples 7, 9, 12, 14 and Theorems 8, 11, 13 are actually new, while Theorem 6 makes precise a claim contained in [29]. Moreover, as far as we know, Section 4 is the first attempt to develop P.A. for stationary data. It provides a brief discussion of Problem (*) and introduces two large classes of stationary sequences.

As already noted, even if X could be potentially given any distribution, in most applications X is required some conditions. There is obviously a number of such conditions. Among them, we decided to focus on exchangeability, stationarity and conditional identity in distribution. This choice seems reasonable to keep the paper focused, but of course it leaves out various interesting conditions, such as partial exchangeability. To write a paper of reasonable length, however, some choice was necessary.

To defend our choice, we note that, in addition to be natural in various practical problems, exchangeability is the usual assumption in Bayesian prediction. Hence, taking exchangeability into account is more or less mandatory. Moreover, since X is exchangeable if and only if it is stationary and c.i.d., the other two conditions can be motivated as the basic components of exchangeability. But there are also other reasons for dealing with them. Stationarity is in fact a routine assumption in the classical treatment of time series, and it is reasonable to consider it from the Bayesian point of view as well. Conditional identity in distribution, even if not that popular, seems to be quite suitable for P.A.; see Section 3.

The rest of the paper is organized in three sections, each concerned with a specific assumption on X, plus a final section of open problems. All the proofs are gathered in the Appendix.

We close this Introduction with some further notation.

As usual, δ_u is the unit mass at the point u. For each $x \in S^n$, where n is a positive integer or $n = \infty$, we denote by x_i the ith coordinate of x. Moreover, we take X to be the sequence of coordinate random variables on S^∞, namely,

$$X_i(x) = x_i \quad \text{for all } i \geq 1 \text{ and } x \in S^\infty.$$

From now on, we fix a strategy σ and we assume

$$X \overset{d}{=} P_\sigma.$$

We write ν instead of σ_0 (i.e., we let $\sigma_0 = \nu$). Hence, ν is a probability measure on $\mathcal{B}$ to be regarded as the distribution of X_1 under the strategy σ. Finally, to avoid technicalities, S is assumed to be a Borel subset of a Polish space and $\mathcal{B}$ the Borel σ-field on S.

2. EXCHANGEABLE DATA

A permutation of S^n is a map $\phi : S^n \to S^n$ of the form

$$\phi(x) = (x_{j_1}, \ldots, x_{j_n}) \quad \text{for all } x \in S^n,$$

where $(j_1, \ldots, j_n)$ is a fixed permutation of $(1, \ldots, n)$. A sequence $Y = (Y_1, Y_2, \ldots)$ of random variables is *exchangeable* if

$$\phi(Y_1, \ldots, Y_n) \overset{d}{=} (Y_1, \ldots, Y_n)$$

for all $n \geq 2$ and all permutations ϕ of S^n.

As noted in Section 1.2, if X is required to be exchangeable, applying P.A. is usually hard. But there are a few exceptions and two of them are discussed in this section. We first recall that X is a Dirichlet sequence (or a Pólya sequence, see [11]) if

$$\sigma_n(x) = \frac{c\nu + \sum_{i=1}^n \delta_{x_i}}{n + c} \quad \text{for all } n \geq 0 \text{ and } x \in S^n,$$

where $c > 0$ is a constant, ν a probability measure on $\mathcal{B}$, and $\sigma_0(x)$ is meant as $\sigma_0(x) = \nu$. The role of Dirichlet sequences is actually huge in various frameworks, including Bayesian nonparametrics, population genetics, ecology, combinatorics and number theory; see, for example, [28, 37, 45, 54–56]. From our point of view, however, two facts are to be stressed. First, a Dirichlet sequence is exchangeable. Second, being defined through its predictive distributions, a Dirichlet sequence is a natural candidate for P.A.

2.1 Species Sampling Sequences

For $n \geq 1$ and $x = (x_1, \ldots, x_n) \in S^n$, denote by $k_n = k_n(x)$ the number of distinct values in the vector x and by $x_1^*, \ldots, x_{k_n}^*$ such distinct values (in the order that they appear). Say that X is a *species sampling sequence* if it is exchangeable, $\sigma_0 = \nu$ is nonatomic, and

$$\sigma_n(x) = \sum_{j=1}^{k_n} p_{j,n}(x)\delta_{x_j^*} + q_n(x)\nu$$

$$\text{for all } n \geq 1 \text{ and } x \in S^n,$$

where the $p_{j,n}$ are nonnegative measurable functions on S^n and $q_n = 1 - \sum_{j=1}^{k_n} p_{j,n}$. Under this strategy, quoting from [42], p. 253, X can be regarded as: "the sequence of species of individuals in a process of sequential random sampling from some hypothetical infinite population of individuals of various species. The species of the first individual to be observed is assigned a random tag $X_1 = X_1^*$ distributed according to ν. Given the tags $X_1, \ldots, X_n$ of the first n individuals observed, it is supposed that the next individual is one of the jth species observed so far with probability $p_{j,n}$, and one of a new species with probability q_n."

A nice consequence of the definition is that $p_{j,n}(x)$ depends on x only through the vector $(N_{1,n}, \ldots, N_{k_n,n})$, where

$$N_{j,n} = N_{j,n}(x) = \text{card}\{k : 1 \leq k \leq n, x_k = x_j^*\}$$

is the number of times that x_j^* appears in the vector x; see [42, 54].

The most popular example of species sampling sequence is probably the *two-parameter Poisson–Dirichlet*, introduced by Pitman in [53], which corresponds to the weights

$$p_{j,n}(x) = \frac{N_{j,n} - b}{n + c} \quad \text{and} \quad q_n(x) = \frac{bk_n + c}{n + c},$$

where b and c are constants such that: either (i) $0 \leq b < 1$ and $c > -b$ or (ii) $b < 0$ and $c = -mb$ for some integer $m \geq 2$. In this model, if L denotes the number of distinct values appearing in the sequence X, one obtains $L \overset{a.s.}{=} \infty$ under (i) and $L \overset{a.s.}{=} m$ under (ii). Note also that X reduces to a Dirichlet sequence in the special case $b = 0$.

Another example, due to [38], is

$$p_{j,n}(x) = \frac{(N_{j,n} + 1)(n - k_n + b)}{n^2 + bn + c}$$

$$\text{and} \quad q_n(x) = \frac{k_n^2 - bk_n + c}{n^2 + bn + c},$$

where $b > 0$ and c is such that $k^2 + bk + c > 0$ for all integers $k > 0$. This time, unlike the two-parameter Poisson–Dirichlet, L is a finite but nondegenerate random variable.

In general, to obtain a species sampling sequence, the forecaster needs to select ν and the weights $p_{j,n}$. While the choice of ν is free (apart from nonatomicity), the $p_{j,n}$ are subjected to the constraint that X should be exchangeable. (Incidentally, the choice of $p_{j,n}$ is a good example of the difficulty of applying P.A. when X is required to be exchangeable). The usual method to select $p_{j,n}$ involves *exchangeable random partitions*. Let $\mathbb{N} = \{1, 2, \ldots\}$ and let Π be a random partition of $\mathbb{N}$. For each $n \geq 1$, call Π_n the restriction of Π to $\{1, \ldots, n\}$, namely, the random partition of $\{1, \ldots, n\}$ whose elements are of the form $\{1, \ldots, n\} \cap A$ for some $A \in \Pi$. Say that Π is exchangeable if

$$\varphi(\Pi_n) \overset{d}{=} \Pi_n$$

for all $n \geq 1$ and all permutations φ of $(1, \ldots, n)$, where $\varphi(\Pi_n)$ denotes the random partition $\varphi(\Pi_n) =$

$\{\varphi(B) : B \in \Pi_n\}$. For instance, given any sequence $Y = (Y_1, Y_2, \ldots)$ of random variables, define Π to be the random partition of $\mathbb{N}$ induced by the equivalence relation $i \sim j \Leftrightarrow Y_i = Y_j$. Then Π is exchangeable provided Y is exchangeable. Now, the weights $p_{j,n}$ of a species sampling sequence correspond, in a canonical way, to the probability law of an exchangeable partition; see [53, 54]. Hence, choosing the $p_{j,n}$ essentially amounts to choosing an exchangeable partition. We stop here since a detailed discussion of exchangeable partitions is bejond the scopes of this paper. The interested reader is referred to [38, 39, 48, 49, 53, 55] and references therein.

A last remark is that the definition of species sampling sequences can be generalized. In particular, nonatomicity of ν can be dropped (as in [3] and [13]) and exchangeability can be replaced by some weaker condition (as in [1] and [2]).

2.2 Kernel-Based Dirichlet Sequences

In [4], to generalize Dirichlet sequences while preserving their main properties, a class of strategies has been introduced. Among other things, such strategies make X exchangeable.

A *kernel* α on $(S, \mathcal{B})$ is a collection

$$\alpha = \{\alpha(\cdot|x) : x \in S\}$$

such that $\alpha(\cdot|x)$ is a probability measure on $\mathcal{B}$, for each $x \in S$, and the map $x \mapsto \alpha(A|x)$ is measurable for each $A \in \mathcal{B}$. Sometimes, to make the notation easier, we will write α_x instead of $\alpha(\cdot|x)$. A straightforward example of kernel is $\alpha_x = \delta_x$ for each $x \in S$.

Fix a probability measure ν on $\mathcal{B}$, a constant $c > 0$, a kernel α on $(S, \mathcal{B})$, and define the strategy

$$\sigma_n(x) = \frac{c\nu + \sum_{i=1}^n \alpha_{x_i}}{n + c} \tag{2}$$

for all $n \geq 0$ and $x \in S^n$. Clearly, X reduces to a Dirichlet sequence if $\alpha = \delta$. In this case, we also say that X is a *classical* Dirichlet sequence.

If α is an arbitrary kernel, X may fail to be exchangeable. However, a useful sufficient condition for exchangeability is available. In fact, X is exchangeable if α agrees with the conditional distribution for ν given some sub-σ-field $\mathcal{G} \subset \mathcal{B}$. For instance, if $\mathcal{G} = \mathcal{B}$, then $\alpha = \delta$ and X is a classical Dirichlet sequence. At the opposite extreme, if $\mathcal{G}$ is the trivial σ-field, then $\alpha_x = \nu$ for all $x \in S$ and X is i.i.d. with common distribution ν. In general, for fixed ν and c, a strategy σ which makes X exchangeable can be associated with any sub-σ-field $\mathcal{G} \subset \mathcal{B}$. It suffices to take α as the conditional distribution for ν given $\mathcal{G}$.

EXAMPLE 2 (Countable partitions). Let $\mathcal{H}$ be a (nonrandom) countable partition of S such that $H \in \mathcal{B}$ and $\nu(H) > 0$ for all $H \in \mathcal{H}$. For $x \in S$, denote by H_x the only $H \in \mathcal{H}$ such that $x \in H$. The conditional distribution for ν given the sub-σ-field generated by $\mathcal{H}$ is

$$\alpha(\cdot|x) = \sum_{H \in \mathcal{H}} 1_H(x)\nu(\cdot|H) = \nu(\cdot|H_x) \quad \text{for all } x \in S.$$

Hence, X is exchangeable whenever

$$\sigma_n(x) = \frac{c\nu + \sum_{i=1}^n \nu(\cdot|H_{x_i})}{n + c} \quad \text{for all } n \geq 0 \text{ and } x \in S^n.$$

Some remarks on the above strategy σ are in order.

- σ may be reasonable when the basic information provided by each observation x_i is H_{x_i}, namely, the element of the partition $\mathcal{H}$ including x_i.
- If S is countable, each sub-σ-field $\mathcal{G} \subset \mathcal{B}$ is generated by a partition $\mathcal{H}$ of S. Hence, α is necessarily as above.
- $\sigma_n(x)$ is absolutely continuous with respect to ν for all n and x. This is a striking difference with classical Dirichlet sequences. To make an example, call σ^* the strategy obtained by σ replacing α with δ. Under σ^*, X is a classical Dirichlet sequence. Moreover, suppose ν is nonatomic and define the set $B(x) = \{x_1, \ldots, x_n\}$ for each $x = (x_1, \ldots, x_n) \in S^n$. Since ν is nonatomic and $B(x)$ is finite,

  $$P_\sigma\big(X_{n+1} = X_i \text{ for some } i \leq n | (X_1, \ldots, X_n) = x\big) = \sigma_n\big(x, B(x)\big) = 0.$$

  On the other hand, since $\delta_{x_i}(B(x)) = 1$ for each $i = 1, \ldots, n$,

  $$P_{\sigma^*}\big(X_{n+1} = X_i \text{ for some } i \leq n | (X_1, \ldots, X_n) = x\big) = \sigma_n^*\big(x, B(x)\big) = n/(n + c).$$

  As a consequence, one obtains

  $$P_\sigma(\text{all the observations are distinct}) = 1,$$
  $$P_{\sigma^*}(\text{all the observations are distinct}) = 0.$$

- σ can be generalized replacing α with

  $$\beta(\cdot|x) = 1_A(x)\delta_x + 1_{A^c}(x)\nu\big(\cdot|A^c \cap H_x\big),$$

  where $A \in \mathcal{B}$ is a suitable set. Note that β reduces to α if $A = \varnothing$. Roughly speaking, β is reasonable in those problems where there is a set A such that x_i is informative about the future observations only if $x_i \in A$. Otherwise, if $x_i \notin A$, the only relevant information provided by x_i is H_{x_i}. As a trivial example, take $S = \mathbb{R}$ and

  $$\mathcal{H} = \big\{(-\infty, 0), \{0\}, (0, \infty)\big\}, \quad A = [-u, u]$$

  for some $u > 0$. Then β is reasonable if x_i is informative only if $|x_i| \leq u$. Otherwise, if $|x_i| > u$, the only meaningful information provided by x_i is its sign.

EXAMPLE 3 (Pólya urns). Some Pólya urns are covered by Example 2. It follows that, for such urns, the sequence X of observed colors is exchangeable. To our knowledge, this fact was previously unknown.

As an example, consider sequential draws from an urn and denote by X_n the color of the ball extracted at time $n \geq 1$. At time $n = 0$, the urn contains m_j balls of color j where $j \in \{1, \ldots, k\}$. Define

$$S = \{1, \ldots, k\}, \quad m = \sum_{j=1}^{k} m_j \quad \text{and} \quad \nu\{j\} = \frac{m_j}{m}$$

for each $j \in S$. The sampling scheme is as follows. Fix a partition $\mathcal{H}$ of S and define

$$m_j^* = m\nu(\{j\}|H_j) = \frac{mm_j}{\sum_{i \in H_j} m_i}.$$

For each $n \geq 1$, one obtains $X_n \in H$ for some unique $H \in \mathcal{H}$. In this case (i.e., if $X_n \in H$) the extracted ball is replaced together with m_j^* more balls of color j for each $j \in H$. In other terms, if the observed color belongs to H, *each* color in H is reinforced (and not only the observed color). In particular, after each draw, m new balls are added to the urn. Hence, denoting by σ the strategy of Example 2 with $c = 1$, one obtains

$$\begin{aligned} &P\big(X_{n+1} = j|(X_1, \ldots, X_n) = x\big) \\ &= \frac{m_j + \sum_{i=1}^{n} 1_{H_j}(x_i) m_j^*}{m + mn} \\ &= \frac{\nu\{j\} + \sum_{i=1}^{n} 1_{H_j}(x_i)\nu(\{j\}|H_j)}{1 + n} \\ &= \frac{c\nu\{j\} + \sum_{i=1}^{n} \nu(\{j\}|H_{x_i})}{c + n} = \sigma_n(x)\{j\}. \end{aligned}$$

If σ is the strategy (2), in addition to exchangeability, X satisfies various other properties of classical Dirichlet sequences. We refer to [4] for details. Here, we just note that the prior π and the posterior π_n can be explicitly determined. In particular, up to replacing δ with α, the Sethuraman's representation of π (see [57]) is still true. Precisely, π is the probability distribution of a random probability measure μ of the form

$$\mu(\cdot) = \sum_j V_j \alpha(\cdot|Z_j),$$

where:

- (Z_j) and (V_j) are independent sequences of random variables;
- (Z_j) is i.i.d. with common distribution ν;
- $V_j = U_j \prod_{i=1}^{j-1}(1 - U_i)$ for all $j \geq 1$, where (U_i) is i.i.d. with common distribution beta$(1, c)$. Namely, (V_j) has the *stick breaking distribution* with parameter c.

3. CONDITIONALLY IDENTICALLY DISTRIBUTED DATA

A sequence $Y = (Y_1, Y_2, \ldots)$ of random variables is *conditionally identically distributed* (c.i.d.) if $Y_2 \overset{d}{=} Y_1$ and

$$P(Y_k \in \cdot|Y_1, \ldots, Y_n) = P(Y_{n+1} \in \cdot|Y_1, \ldots, Y_n) \quad \text{a.s.}$$

for all $k > n \geq 1$. A c.i.d. sequence Y is identically distributed. It is also asymptotically exchangeable in the sense that, as $n \to \infty$, the probability distribution of the shifted sequence $(Y_n, Y_{n+1}, \ldots)$ converges weakly to an exchangeable law. Moreover, as already stressed, Y is exchangeable if and only if it is stationary and c.i.d.

Conditionally identically distributed sequences have been introduced in [7, 47] and then investigated or applied in various papers; see, for example, [1, 2, 5, 6, 8, 9, 14, 15, 29, 30, 35].

There are reasons for taking c.i.d. data into account in Bayesian prediction. In fact, in a sense, c.i.d. sequences have been introduced having prediction in mind. If X is c.i.d., at each time n, the future observations $(X_k : k > n)$ are identically distributed given the past, and this is reasonable in several prediction problems. Examples arise in clinical trials, generalized Pólya urns, species sampling models, survival analysis and disease surveillance; see [1, 2, 5–7, 14, 15, 29, 30, 34]. A further reason for assuming X c.i.d. is that the asymptotics is very close to that of exchangeable sequences. As a consequence, a meaningful part of the usual Bayesian machinery can be developed under the sole assumption that X is c.i.d.; see [29]. Finally, the strategies which make X c.i.d. can be easily characterized; see Theorem 15 in the Appendix. Hence, unlike the exchangeable case, P.A. can be easily implemented for c.i.d. data. A number of interesting strategies, which cannot be used if X is required to be exchangeable, become available if X is only asked to be c.i.d.; see, for example, [5, 6].

As a concrete example, fix a constant $q \in (0, 1)$ and define

$$\sigma_n(x) = q^n \nu + (1 - q) \sum_{i=1}^{n} q^{n-i} \delta_{x_i} \tag{3}$$

for all $n \geq 0$ and $x \in S^n$. Using σ to make predictions corresponds to exponential smoothing. It may be reasonable when the forecaster has only vague opinions on the dependence structure of the data, and yet she feels that the weight of the ith observation x_i should be a decreasing function of $n - i$. In this case, X is not exchangeable, since $\sigma_n(x)$ is not invariant under permutation of x, but it can be easily seen to be c.i.d.; see [6], Example 7.

In this section, following [5, 6], P.A. is applied to c.i.d. data. We first report some known strategies (Section 3.1) and then we introduce two new strategies which make X c.i.d. (Section 3.2).

3.1 Fast Recursive Update of Predictive Distributions

A possible condition for a strategy σ is

$$\text{(4)} \qquad \sigma_{n+1}(x, y) \text{ is a function of } \sigma_n(x) \text{ and } y$$

for all $n \geq 0$, $x \in S^n$ and $y \in S$, where y denotes the $(n+1)$-th observation and

$$(x, y) = (x_1, \ldots, x_n, y).$$

Under (4), the predictive $\sigma_{n+1}(x, y)$ is just a recursive update of the previous predictive $\sigma_n(x)$ and the last observation y. Recursive properties of this type are useful in applications. They have a long history (see, e.g., [51, 52, 59]) and have been recently investigated in [41].

For each $n \geq 0$, let $q_n : S^n \to [0, 1]$ be a measurable function (with q_0 constant) and α_n a kernel on $(S, \mathcal{B})$. Define a strategy σ through the recursive equations

$$\text{(5)} \qquad \begin{aligned} &\sigma_0 = \nu \quad \text{and} \\ &\sigma_{n+1}(x, y) = q_n(x)\sigma_n(x) + \big(1 - q_n(x)\big)\alpha_n(\cdot|y) \end{aligned}$$

for all $n \geq 0$, $x \in S^n$ and $y \in S$. Since $\sigma_{n+1}(x, y)$ is a convex combination of the previous predictive $\sigma_n(x)$ and the kernel $\alpha_n(\cdot|y)$, which depends only on y, the strategy σ satisfies condition (4). The obvious interpretation is that, at time $n+1$, after observing (x, y), the next observation is drawn from $\sigma_n(x)$ with probability $q_n(x)$ and from $\alpha_n(\cdot|y)$ with probability $1 - q_n(x)$.

An example of strategy satisfying equation (5) is Newton's algorithm [51, 52]. More precisely, Newton's algorithm aims to estimate the latent distribution in a mixture model rather than to make predictions. However, if reinterpreted as a predictive rule, Newton's algorithm corresponds to a strategy σ and such a σ meets equation (5) for a suitable choice of q_n and α_n; see, for example, [34], p. 1095. Moreover, as shown in [34], σ makes X c.i.d.

The strategies satisfying equation (5) are investigated in [5]. Under such strategies, X is usually not exchangeable but it is c.i.d. under some conditions on the kernels α_n. Precisely, X is c.i.d. if α_n is the conditional distribution for ν given $\mathcal{G}_n$ for each $n \geq 0$, where

$$\mathcal{G}_0 \subset \mathcal{G}_1 \subset \mathcal{G}_2 \subset \cdots \subset \mathcal{B}$$

is any filtration (i.e., any increasing sequence of sub-σ-fields of $\mathcal{B}$). This condition is trivially true if $\alpha_n(\cdot|y) = \delta_y$ for all $y \in S$ (just take $\mathcal{G}_n = \mathcal{B}$ for all $n \geq 0$).

EXAMPLE 4 (Finer countable partitions). For each $n \geq 0$, let $\mathcal{H}_n$ be a countable partition of S such that $H \in \mathcal{B}$ and $\nu(H) > 0$ for all $H \in \mathcal{H}_n$. Suppose that $\mathcal{H}_{n+1}$ is finer than $\mathcal{H}_n$ for all $n \geq 0$. Define σ through equation (5) with

$$\alpha_n(\cdot|y) = \sum_{H \in \mathcal{H}_n} 1_H(y)\nu(\cdot|H) = \nu\big(\cdot|H_y^n\big),$$

where H_y^n denotes the only $H \in \mathcal{H}_n$ such that $y \in H$. The kernel α_n is the conditional distribution for ν given $\mathcal{G}_n$, where $\mathcal{G}_n$ is the σ-field generated by $\mathcal{H}_n$. Since $\mathcal{H}_{n+1}$ is finer than $\mathcal{H}_n$, one obtains $\mathcal{G}_n \subset \mathcal{G}_{n+1}$. Hence, X is c.i.d. Note also that the $\mathcal{H}_n$ could be chosen such that

$$\{y\} = \bigcap_n H_y^n \quad \text{for all } y \in S.$$

In this case, as $n \to \infty$, the partitions $\mathcal{H}_n$ shrink to the partition of S in the singletons.

For instance, in Example 2, suppose the forecaster wants to replace the fixed partition $\mathcal{H}$ with a sequence $\mathcal{H}_n$ of finer partitions. This is possible at the price of having X c.i.d. instead of exchangeable. In fact, with $q_n = \frac{n+c}{n+1+c}$, one obtains

$$\begin{aligned} \sigma_n(x) &= \frac{c\nu + \sum_{i=1}^n \alpha_{i-1}(\cdot|x_i)}{n+c} \\ &= \frac{c\nu + \sum_{i=1}^n \nu(\cdot|H_{x_i}^{i-1})}{n+c}. \end{aligned}$$

Similarly, to decrease the impact of the observed data while preserving the c.i.d. condition, the strategy (3) could be modified as

$$\sigma_n(x) = q^n\nu + (1-q)\sum_{i=1}^n q^{n-i}\nu\big(\cdot|H_{x_i}^{i-1}\big).$$

We next turn to a strategy introduced in [41]. Once again, under this strategy, the data are c.i.d. but not necessarily exchangeable.

EXAMPLE 5 (Hahn, Martin and Walker; Copulas). In this example, $S = \mathbb{R}$ and "density function" means "density function with respect to Lebesgue measure." A bivariate *copula* is a distribution function on $\mathbb{R}^2$ whose marginals are uniform on $(0, 1)$. The density function of a bivariate copula, provided it exists, is said to be a *copula density*.

In [41], in order to realize condition (4), the following updating rule is introduced. Fix a density f_0 and a sequence $c_1, c_2, \ldots$ of bivariate copula densities. For the sake of simplicity, we assume $f_0 > 0$ and $c_n > 0$ for all $n \geq 1$. For $n = 0$, define $\sigma_0(dz) = f_0(z)\,dz$ and call F_0 the distribution function corresponding to σ_0. Then, for each $y \in \mathbb{R}$, define

$$\begin{aligned} &\sigma_1(y, dz) = f_1(z|y)\,dz, \quad \text{where} \\ &f_1(z|y) = c_1\big\{F_0(z), F_0(y)\big\}f_0(z). \end{aligned}$$

In general, for each $n \geq 0$ and $x \in \mathbb{R}^n$, suppose $\sigma_n(x)$ has been defined and denote by $f_n(\cdot|x)$ and $F_n(\cdot|x)$ the density and the distribution function of $\sigma_n(x)$. Then, for all $y \in \mathbb{R}$, one can define

$$\text{(6)} \qquad \begin{aligned} &\sigma_{n+1}(x, y, dz) = f_{n+1}(z|x, y)\,dz, \quad \text{where} \\ &f_{n+1}(z|x, y) = c_{n+1}\big\{F_n(z|x), F_n(y|x)\big\}f_n(z|x). \end{aligned}$$

Equation (6) defines a strategy σ dominated by the Lebesgue measure.

In [41] (but not here) the c_n are also required to be symmetric. Furthermore, in [41], equation (6) is not necessarily viewed as a method for obtaining a strategy but is *deduced* as a consequence of exchangeability. From our point of view, instead, equation (6) defines a strategy σ which we call HMW's strategy.

Under HMW's strategy, X is not necessarily exchangeable, even if the c_n are symmetric and $c_n \to 1$ (in some sense) as $n \to \infty$. To see this, recall that X is i.i.d. if and only if it is exchangeable and X_1 is independent of X_2. In turn, X_1 is independent of X_2 if c_1 is the independence copula density (i.e., $c_1(u, v) = 1$ for all $(u, v) \in [0, 1]^2$). Therefore, X fails to be exchangeable whenever c_1 is the independence copula density and $c_2 \neq c_1$. However, as noted in [29], X turns out to be c.i.d.

THEOREM 6. *If σ is HMW's strategy, then X is c.i.d.*

A proof of Theorem 6 is provided in the Appendix. We note that, for Theorem 6 to hold, the positivity assumption on f_0 and c_n may be dropped and the c_n can be taken to be conditional copula densities; see Remark 16.

3.2 Further Examples

In the next example, the data are exchangeable until a stopping time T and then go on so as to form a c.i.d. sequence. The time T should be regarded as the first time when something meaningful happens, possibly something modifying the nature of the observed phenomenon. Even if apparently involved, the example could find some applications. For instance, to model censored survival times, with $T - 1$ the first time when a given number of survival times is observed.

EXAMPLE 7 (Change points). A predictable stopping time is a function T on S^∞, with values in $\{2, 3, \ldots, \infty\}$, satisfying

$$\{T = n + 1\} = \big\{(X_1, \ldots, X_n) \in A_n\big\} \tag{7}$$

for some set $A_n \in \mathcal{B}^n$. Basically, condition (7) means that the event $\{T = n + 1\}$ depends only on $(X_1, \ldots, X_n)$. Similarly, $\{T \leq n + 1\} = \bigcup_{j=2}^{n+1}\{T = j\}$ depends only on $(X_1, \ldots, X_n)$. Therefore, for all $x \in S^n$ and $y \in S$, the indicators of $\{T \leq n + 1\}$ and $\{T > n + 1\}$ depend on x but not on y.

Fix a predictable stopping time T and a strategy $\beta = (\beta_0, \beta_1, \ldots)$, which makes X exchangeable. Moreover, as in Section 3.1, fix the measurable functions $q_n : S^n \to [0, 1]$. Then define $\sigma_0 = \beta_0$, $\sigma_1 = \beta_1$, and

$$\begin{aligned}\sigma_{n+1}(x, y) &= 1_{\{T>n+1\}}(x)\beta_{n+1}(x, y) \\ &\quad + 1_{\{T\leq n+1\}}(x)\big\{q_n(x)\sigma_n(x) + \big(1 - q_n(x)\big)\delta_y\big\}\end{aligned}$$

for all $n \geq 1$, $x \in S^n$ and $y \in S$. The next result is proved in the Appendix.

THEOREM 8. *The above strategy σ makes X c.i.d. Moreover, if*

$$A_n \text{ is invariant under permutations of } S^n \text{ for all } n \geq 1,$$

where A_n is the set involved in condition (7), then $(X_1, \ldots, X_n)$ is exchangeable conditionally on $T > n$. Precisely,

$$\begin{aligned}&P_\sigma\big(\phi(X_1, \ldots, X_n) \in \cdot\,|T > n\big) \\ &\quad = P_\sigma\big((X_1, \ldots, X_n) \in \cdot\,|T > n\big)\end{aligned}$$

for all n such that $P_\sigma(T > n) > 0$ and all permutations ϕ of S^n.

Theorem 8 is still valid if σ is defined differently at the times subsequent to T. For instance, given a countable partition $\mathcal{H}$ of S, the conclusions of Theorem 8 are true even if

$$\sigma_{n+1}(x, y) = q_n(x)\sigma_n(x) + \big(1 - q_n(x)\big)\sigma_n(x, \cdot\,|H_y)$$

for all $x \in S^n$ and $y \in S$ such that $T \leq n + 1$ and $\sigma_n(x, H_y) > 0$. Here, $\sigma_n(x, \cdot\,|H_y)$ denotes the probability measure

$$\sigma_n(x, A|H_y) = \frac{\sigma_n(x, A \cap H_y)}{\sigma_n(x, H_y)} \quad \text{for all } A \in \mathcal{B}.$$

Censored survival times are a possible application of σ. Suppose that $S = \{0, 1\} \times (0, \infty)$ and the ith observation is a pair $x_i = (j_i, t_i)$ where t_i is the survival time of item i, or the time when item i leaves the trial, according to whether $j_i = 1$ or $j_i = 0$. In this framework, $T - 1$ could be the first time when a fixed number k of survival times is observed, namely,

$$T = 1 + \inf\left\{n : \sum_{i=1}^{n} j_i = k\right\}$$

with the usual convention $\inf \varnothing = \infty$. Finally, the strategy β could be as in Section 2.2. In fact, classical Dirichlet sequences are a quite popular model to describe censored survival times but have the drawback of ties. This drawback may be overcome if β is of the form

$$\beta_n(x) = \frac{c\nu + \sum_{i=1}^{n} \alpha_{x_i}}{n + c},$$

where the kernel α satisfies the conditions of Section 2.2 and ν and α_x are nonatomic for all $x \in S$.

So far, the nth predictive distribution has been meant as the conditional distribution of X_{n+1} given $(X_1, \ldots, X_n)$. But the information available at time n is often strictly larger than $(X_1, \ldots, X_n)$. To model this situation, we suppose to observe the sequence

$$Y = (X_1, Z_1, X_2, Z_2, \ldots),$$

where $Z = (Z_1, Z_2, \ldots)$ is any sequence of random variables. The Z_n can be regarded as covariates. At each

time n, the forecaster aims to predict X_{n+1} based on $(X_1, Z_1, \ldots, X_n, Z_n)$. She is not interested in Z_{n+1} as such, but $Z_1, \ldots, Z_n$ cannot be neglected since they are informative on X_{n+1}. Moreover, she wants X to be c.i.d. and Z unconstrained as much as possible. One solution could be a strategy, which makes Y c.i.d. However, if Y is c.i.d., both X and Z are marginally c.i.d., and having Z c.i.d. may be unwelcome. In the next example, X is c.i.d. but Z is not. In addition, X satisfies a condition stronger than the c.i.d. one, that is, $X_2 \overset{d}{=} X_1$ and

$$P(X_k \in \cdot | X_1, Z_1, \ldots, X_n, Z_n) = P(X_{n+1} \in \cdot | X_1, Z_1, \ldots, X_n, Z_n) \tag{8}$$

a.s. for all $k > n \geq 1$; see [7].

EXAMPLE 9 (Covariates). Let $S = \mathbb{R}^2$ and

$$0 = b_0 < b_1 < b_2 < \cdots, \quad \sup_n b_n \leq 1,$$

a bounded strictly increasing sequence of real numbers. Take σ_0 as the probability distribution of $(U + V, V)$ where

$$U \text{ independent of } V, \quad U \overset{d}{=} \mathcal{N}(0, b_1), \quad V \overset{d}{=} \mathcal{N}(0, 1 - b_1).$$

Similarly, for each $n \geq 1$ and

$$y = (y_1, \ldots, y_n) = (x_1, z_1, \ldots, x_n, z_n),$$

take $\sigma_n(y)$ as the probability distribution of $(U_n(y) + V_n(y), V_n(y))$ where

$$U_n(y) \text{ independent of } V_n(y),$$
$$U_n(y) \overset{d}{=} \mathcal{N}(x_n - z_n, b_{n+1} - b_n),$$
$$V_n(y) \overset{d}{=} \mathcal{N}(0, 1 - b_{n+1}).$$

Then Z is not c.i.d. while X satisfies condition (8). Furthermore, arguing as in [5], Section 4, the normal distribution could be replaced by any symmetric stable law.

To see that Z is not c.i.d., just note that Z fails to be identically distributed. To prove condition (8), take a collection $\{T_n, W_n : n \geq 1\}$ of independent standard normal random variables and define the sequence

$$Y^* = (X_1^*, Z_1^*, X_2^*, Z_2^*, \ldots),$$

where $Z_n^* = \sqrt{1 - b_n} W_n$ and

$$X_n^* = \sum_{j=1}^{n} \sqrt{b_j - b_{j-1}} T_j + Z_n^*.$$

It is not hard to verify that $Y^* \overset{d}{=} Y$. Hence, it suffices to prove (8) with Y^* in the place of Y, and this can be done as in [7], Example 1.2. We omit the explicit calculations.

4. STATIONARY DATA

A sequence $Y = (Y_1, Y_2, \ldots)$ of random variables is *stationary* if

$$(Y_2, \ldots, Y_{n+1}) \overset{d}{=} (Y_1, \ldots, Y_n) \quad \text{for all } n \geq 1.$$

In the non-Bayesian approaches to prediction, stationarity is a classical assumption. In a Bayesian framework, instead, stationarity seems to be less popular. In particular, to our knowledge, there is no systematic treatment of P.A. for stationary data. This section aims to fill this gap and begins an investigation of P.A. when X is required to be stationary. It is just a preliminary step and much more work is to be done.

After some general remarks on Problem (*), two large classes of stationary sequences will be introduced. Incidentally, these two classes may look unusual for a Bayesian forecaster. We do not know whether this is true, but we recall that P.A. is consistent with any probability distribution for X. Hence, in a Bayesian framework, using data coming from such classes is certainly admissible.

If X is required to be stationary, for P.A. to apply, the strategies which make X stationary should be characterized. Hence, one comes across Problem (*) with $\mathcal{C}$ the class of stationary probability measures on $(S^\infty, \mathcal{B}^\infty)$. This version of Problem (*) is quite hard and we are not aware of any general solution; see, for example, [12, 50] and references therein. Fortunately, however, Problem (*) is simple (or even trivial) in a few special cases. As an example, a strategy σ makes X a stationary (first-order) Markov chain if and only if

$$\int \sigma_1(x, \cdot)\sigma_0(dx) = \sigma_0(\cdot) \quad \text{and} \quad \sigma_n(x) = \sigma_1(x_n)$$

for all $n \geq 1$ and P_σ-almost all $x \in S^n$. Even if obvious, this fact has a useful practical consequence. If the data are required to be stationary and Markov, in order to make Bayesian predictions, applying P.A. is straightforward.

Another remark is that, unlike the exchangeable case, a finite-dimensional stationary random vector can be always extended to an (infinite) stationary sequence. To formalize this fact, we first recall that the probability distribution of the random vector $(X_1, \ldots, X_n)$ is completely determined by $\sigma_0, \sigma_1, \ldots, \sigma_{n-1}$.

LEMMA 10. *Fix $n \geq 1$, select $\sigma_0, \sigma_1, \ldots, \sigma_{n-1}$ and define*

$$\sigma_j(u, x) = \sigma_{n-1}(x)$$

for all $j > n - 1$, $u \in S^{j-n+1}$ and $x \in S^{n-1}$. Then X is stationary provided $(X_2, \ldots, X_n) \overset{d}{=} (X_1, \ldots, X_{n-1})$.

Lemma 10 is probably well known, but again we do not know of any explicit reference. Anyway, the proof is

straightforward. It suffices to note that, under the strategy of Lemma 10, X_{j+1} is conditionally independent of $(X_1, \ldots, X_{j-n+1})$ given $(X_{j-n+2}, \ldots, X_j)$.

A last remark is that Problem (*) admits an obvious solution for dominated strategies. In this case, incidentally, Problem (*) can be easily solved even for exchangeable data.

THEOREM 11. *Let λ be a σ-finite measure on $(S, \mathcal{B})$ and σ a strategy dominated by λ, say*

$$\sigma_0(dy) = f_0(y)\lambda(dy) \quad \text{and} \quad \sigma_n(x, dy) = f_n(y|x)\lambda(dy)$$

for all $n \geq 1$ and $x \in S^n$. Define

$$g_n(x) = f_0(x_1) f_1(x_2|x_1) \cdots f_{n-1}(x_n|x_1, \ldots, x_{n-1})$$

for all $n \geq 1$ and $x \in S^n$. Then:

- *P_σ is stationary if and only if*

$$g_n(x) = \int g_{n+1}(u, x)\lambda(du)$$

for all $n \geq 1$ and P_σ-almost all $x \in S^n$.

- *P_σ is exchangeable if and only if*

$$g_n\big(\phi(x)\big) = g_n(x)$$

for all $n \geq 2$, all permutations ϕ of S^n and P_σ-almost all $x \in S^n$.

The proof of Theorem 11 is given in the Appendix.

We finally give two examples. In both, X is a stationary Markov sequence, possibly of order greater than 1.

EXAMPLE 12 (Generalized autoregressive sequences). Let $S = \mathbb{R}$. Fix a probability measure μ on $\mathcal{B}$ and a measurable function $f : \mathbb{R} \to \mathbb{R}$. Define

$$\sigma_1(x, A) = P\big(f(x) + U \in A\big) \quad \text{for all } x \in \mathbb{R} \text{ and } A \in \mathcal{B},$$

where U is a real random variable such that $U \overset{d}{=} \mu$. *Suppose* now that

$$\int \sigma_1(x, A)\nu(dx) = \nu(A), \quad A \in \mathcal{B}, \tag{9}$$

for some probability measure ν on $\mathcal{B}$. Then X is a stationary Markov chain provided

$$\sigma_0 = \nu \quad \text{and} \quad \sigma_n(x) = \sigma_1(x_n) \quad \text{for all } n \geq 2 \text{ and } x \in \mathbb{R}^n.$$

Note that $Y \overset{d}{=} P_\sigma$ for any sequence $Y = (Y_1, Y_2, \ldots)$ such that

$$Y_1 \overset{d}{=} \nu \quad \text{and} \quad Y_n = f(Y_{n-1}) + U_n \quad \text{for } n \geq 2,$$

where $(U_n : n \geq 2)$ is i.i.d., independent of Y_1, and $U_2 \overset{d}{=} \mu$. Thus, μ can be regarded as the distribution of the "errors" U_n and ν as the marginal distribution of the observations Y_n. For instance, the usual Gaussian (first-order) autoregressive processes correspond to $f(x) = cx$, $\mu = \mathcal{N}(0, b)$ and $\nu = \mathcal{N}(0, b/(1-c^2))$, where $c \in (-1, 1)$ and $b > 0$ are constants.

To make the above argument concrete, the following problem is to be solved: *For fixed f and μ, give conditions for the existence of ν satisfying equation* (9). *More importantly, give an explicit formula for ν provided it exists*. We next focus on this problem in the (meaningful) special case where μ is a symmetric stable law.

Let $\gamma \in (0, 2]$ be a constant and Z a real random variable with characteristic function

$$E\big\{\exp(itZ)\big\} = \exp\left(-\frac{|t|^\gamma}{2}\right) \quad \text{for all } t \in \mathbb{R}.$$

(The exponent γ is usually denoted by α, but this notation cannot be adopted in this paper since α denotes a kernel). For $a \in \mathbb{R}$ and $b > 0$, denote by $\mathcal{S}(a, b)$ the probability distribution of $a + b^{1/\gamma} Z$, namely,

$$\mathcal{S}(a, b; A) = P\big(a + b^{1/\gamma} Z \in A\big) \quad \text{for all } A \in \mathcal{B}.$$

The probability measure $\mathcal{S}(a, b)$ is said to be a symmetric stable law with exponent γ. Note that $\mathcal{S}(a, b) = \mathcal{N}(a, b)$ if $\gamma = 2$ and $\mathcal{S}(a, b) = \mathcal{C}(a, b)$ if $\gamma = 1$, where $\mathcal{C}(a, b)$ is the Cauchy distribution with density $f(x) = \frac{2b}{\pi} \frac{1}{b^2 + 4(x-a)^2}$ (the standard Cauchy distribution corresponds to $a = 0$ and $b = 2$).

THEOREM 13. *Let $c \in (-1, 1)$ be a constant. If $\mu = \mathcal{S}(a, b)$ and $f(x) = -a + cx$, then equation* (9) *is satisfied by*

$$\nu = \mathcal{S}\left(0, \frac{b}{1 - |c|^\gamma}\right).$$

By Theorem 13, which is proved in the Appendix, one obtains (first-order) stationary autoregressive processes with any symmetric stable marginal distribution.

EXAMPLE 14 (Markov sequences of arbitrary order). Let λ be a σ-finite measure on $(S, \mathcal{B})$. Fix $n \geq 2$ and a measurable function h on S^n such that $h > 0$ and $\int h\, d\lambda^n = 1$. Given h, define a further function g via cyclic permutations of h, namely,

$$g(x) = \frac{1}{n}\big\{h(x_1, \ldots, x_n) + h(x_2, \ldots, x_n, x_1) + \cdots + h(x_n, x_1, \ldots, x_{n-1})\big\}$$

for all $x \in S^n$. Such a g is still a density with respect to λ^n (since $\int g\, d\lambda^n = 1$) and satisfies

$$g(x, y) = g(y, x) \quad \text{for all } x \in S^{n-1} \text{ and } y \in S. \tag{10}$$

Next, define

$$f_0(x) = \int g(x, v)\lambda^{n-1}(dv) \quad \text{for all } x \in S,$$

$$f_{n-1}(x_n|x_1, \ldots, x_{n-1}) = \frac{g(x)}{\int g(x_1, \ldots, x_{n-1}, v)\lambda(dv)}$$

for all $x \in S^n$, and

$$f_{j-1}(x_j|x_1,\ldots,x_{j-1}) = \frac{\int g(x,v)\lambda^{n-j}(dv)}{\int g(x_1,\ldots,x_{j-1},v)\lambda^{n-j+1}(dv)}$$

for all $2 \le j \le n-1$ and $x \in S^j$. Finally, define a strategy σ dominated by λ as

$$\sigma_0(dz) = f_0(z)\lambda(dz),$$
$$\sigma_j(x,dz) = f_j(z|x)\lambda(dz)$$

if $1 \le j \le n-1$ and $x \in S^j$, and

$$\sigma_j(u,x) = \sigma_{n-1}(x)$$

if $j > n-1$, $u \in S^{j-n+1}$ and $x \in S^{n-1}$. Under σ, a density of $(X_1,\ldots,X_n)$ is given by g. By equation (10),

$$\int g(v,x)\lambda(dv) = \int g(x,v)\lambda(dv) \quad \text{for all } x \in S^{n-1}$$

and this in turn implies

$$(X_2,\ldots,X_n) \overset{d}{=} (X_1,\ldots,X_{n-1}).$$

Therefore, X is stationary because of Lemma 10. Note also that X is a Markov sequence of order $n-1$.

5. CONCLUDING REMARKS AND OPEN PROBLEMS

When prediction is the main target, P.A. has some advantages with respect to I.A. This is only our opinion, obviously, and we tried to support it along this paper. Even if one agrees, however, some further work is to be done to make P.A. a concrete tool. We close this paper with a brief list of open problems and possible hints for future research.

- In various applications, the available information strictly includes the past observations on the variable to be predicted. For instance, as in Example 9, suppose one aims to predict X_{n+1} based on $(X_1, Z_1, \ldots, X_n, Z_n)$ where $Z_1,\ldots,Z_n$ are any random elements. Suppose also that $Z_1,\ldots,Z_n$ cannot be neglected for they are informative on X_{n+1}. In this case, one needs the conditional distribution of X_{n+1} given $(X_1, Z_1, \ldots, X_n, Z_n)$. Situations of this type are practically meaningful and should be investigated further.
- Section 4 should be expanded. It would be nice to have a general solution of Problem (*) for both the stationary and the stationary-ergodic cases. Further examples of stationary sequences (possibly, non-Markovian) would be welcome as well.
- Obviously, P.A. could be investigated under other distributional assumptions, in addition to exchangeability, stationarity and conditional identity in distribution. In particular, partial exchangeability should be taken into account.
- A question, related to Example 5, is: Under what conditions X is exchangeable when σ is HMW's strategy?
- While probably hard, the problem raised in Example 12 looks intriguing. In Theorem 13, such a problem has been addressed when μ is a symmetric stable law and f has a special form. What happens if μ and f are arbitrary?
- In case of I.A., the empirical Bayes point of view (where the prior is allowed to depend on the data) may be problematic. In case of P.A., instead, this point of view is certainly admissible. In fact, suppose a strategy σ depends on some unknown constants, and an empirical Bayes forecaster decides to estimate these constants based on the available data. Acting in this way, she is merely replacing a strategy with another. Instead of σ, she is working with $\hat{\sigma}$, where $\hat{\sigma}$ is the strategy obtained from σ estimating the unknown constants. This empirical form of P.A. looks reasonable and could be investigated.

APPENDIX

This Appendix contains the proofs of some claims scattered throughout the text. We will need the following characterization of c.i.d. sequences in terms of strategies.

THEOREM 15 (Theorem 3.1 of [8]). *Let σ be a strategy. Then, P_σ is c.i.d. if and only if*

$$\sigma_n(x,A) = \int \sigma_{n+1}(x,y,A)\sigma_n(x,dy) \tag{11}$$

for all $n \ge 0$, all $A \in \mathcal{B}$ and P_σ-almost all $x \in S^n$.

PROOF OF THEOREM 6. In this proof, "density function" stands for "density function with respect to Lebesgue measure." We first recall a well-known fact.

Let C be a bivariate copula and F_1, F_2 distribution functions on $\mathbb{R}$. Suppose that C, F_1 and F_2 all have densities, say c, f_1 and f_2, respectively. Then

$$F(x,y) = C\{F_1(x), F_2(y)\}$$

is a distribution function on $\mathbb{R}^2$ and

$$f(x,y) = c\{F_1(x), F_2(y)\}f_1(x)f_2(y)$$

is a density of F. Therefore, for all $y \in \mathbb{R}$ with $f_2(y) > 0$, one obtains

$$\int c\{F_1(x), F_2(y)\}f_1(x)\,dx = \int \frac{f(x,y)}{f_2(y)}\,dx = 1.$$

We next show that equation (6) actually defines a strategy σ. Fix a density $f_0 > 0$ and a sequence $c_1, c_2, \ldots$ of strictly positive bivariate copula densities. For each $y \in \mathbb{R}$,

$$\int f_1(z|y)\,dz = \int c_1\{F_0(z), F_0(y)\}f_0(z)\,dz = 1$$

since $f_0(y) > 0$. Moreover, $f_1(z|y) > 0$ for all z due to $f_0 > 0$ and $c_1 > 0$. Next, suppose that $f_n(\cdot|x)$ is a strictly

positive density for some $n \geq 1$ and $x \in \mathbb{R}^n$. Then, for all $y \in \mathbb{R}$,

$$\int f_{n+1}(z|x, y)\, dz = \int c_{n+1}\{F_n(z|x), F_n(y|x)\} f_n(z|x)\, dz = 1$$

since $f_n(y|x) > 0$. Furthermore, $f_{n+1}(z|x, y) > 0$ for all z since $f_n(\cdot|x) > 0$ and $c_{n+1} > 0$. By induction, this proves that $f_n(\cdot|x)$ is a density for all $n \geq 1$ and $x \in \mathbb{R}^n$. Therefore, equation (6) defines a strategy σ (called HMW's strategy in Example 5).

Finally, we prove that P_σ is c.i.d. if σ is HMW's strategy. By Theorem 15, it suffices to prove condition (11). In turn, since σ is dominated by the Lebesgue measure, condition (11) reduces to

$$f_n(z|x) = \int f_{n+1}(z|x, y) f_n(y|x)\, dy$$

for all $n \geq 0$, almost all $z \in \mathbb{R}$ and P_σ-almost all $x \in \mathbb{R}^n$. Such a condition follows directly from the definition of σ. In fact, for all $n \geq 0$ an $x \in \mathbb{R}^n$, one obtains

$$\begin{aligned}\int f_{n+1}(z|x, y) f_n(y|x)\, dy &= \int c_{n+1}\{F_n(z|x), F_n(y|x)\} f_n(z|x) f_n(y|x)\, dy \\ &= f_n(z|x) \quad \text{for almost all } z.\end{aligned}$$

This concludes the proof. □

REMARK 16. HMW's strategy σ has been defined under the assumption that $f_0 > 0$ and $c_n > 0$ for all $n \geq 1$. Such an assumption is superfluous and has been made only to avoid annoying complications in the definition of σ. Similarly, X is c.i.d. even if the c_n are conditional copulas, in the sense that they are allowed to depend on past data. Precisely, for each $n \geq 1$ and $x \in \mathbb{R}^n$, fix a bivariate copula density $c_{n+1}(\cdot|x)$. Then the proof Theorem 6 still applies if $f_{n+1}(z|x, y)$ is rewritten as

$$f_{n+1}(z|x, y) = c_{n+1}\{F_n(z|x), F_n(y|x)|x\} f_n(z|x).$$

PROOF OF THEOREM 8. We show that X is c.i.d. via Theorem 15. Fix $A \in \mathcal{B}$ and $n \geq 0$. Since P_β is exchangeable (and thus c.i.d.) Theorem 15 yields

$$\beta_n(x, A) = \int \beta_{n+1}(x, y, A)\beta_n(x, dy) \tag{12}$$

for P_β-almost all $x \in S^n$. Hence, up to changing β on a P_β-null set, equation (12) can be assumed to hold for all $x \in S^n$. If $n = 0$,

$$\int \sigma_1(y, A)\sigma_0(dy) = \int \beta_1(y, A)\beta_0(dy) = \beta_0(A) = \sigma_0(A),$$

where the first equality is because $\sigma_0 = \beta_0$ and $\sigma_1 = \beta_1$ while the second follows from (12). Next, suppose $n \geq 1$ and take $x \in S^n$ and $y \in S$. By assumption, the events $\{T > n + 1\}$ and $\{T \leq n + 1\}$ depend on x but not on y. If $T > n + 1$, one obtains $\sigma_{n+1}(x, y) = \beta_{n+1}(x, y)$ and $\sigma_n(x) = \beta_n(x)$. Hence, equation (12) implies again

$$\int \sigma_{n+1}(x, y, A)\sigma_n(x, dy) = \int \beta_{n+1}(x, y, A)\beta_n(x, dy) = \beta_n(x, A) = \sigma_n(x, A).$$

Similarly, if $T \leq n + 1$,

$$\begin{aligned}&\int \sigma_{n+1}(x, y, A)\sigma_n(x, dy) \\ &= \int \{q_n(x)\sigma_n(x, A) + (1 - q_n(x))\delta_y(A)\}\sigma_n(x, dy) \\ &= q_n(x)\sigma_n(x, A) + (1 - q_n(x)) \int \delta_y(A)\sigma_n(x, dy) \\ &= \sigma_n(x, A).\end{aligned}$$

In view of Theorem 15, this proves that X is c.i.d.

Finally, suppose that A_n is invariant under permutations of S^n for each $n \geq 1$. We have to show that $(X_1, \ldots, X_n)$ is exchangeable conditionally on $T > n$. Fix n, a set $C \in \mathcal{B}^n$, and a permutation ϕ of S^n. For each $j \geq n$, it is easily seen that

$$\begin{aligned}&P_\sigma(T = j + 1, \phi(X_1, \ldots, X_n) \in C) \\ &\quad = P_\beta(T = j + 1, \phi(X_1, \ldots, X_n) \in C).\end{aligned}$$

Therefore,

$$\begin{aligned}&P_\sigma(T = j + 1, \phi(X_1, \ldots, X_n) \in C) \\ &\quad = P_\beta(T = j + 1, \phi(X_1, \ldots, X_n) \in C) \\ &\quad = P_\beta((X_1, \ldots, X_j) \in A_j, \phi(X_1, \ldots, X_n) \in C) \\ &\quad = P_\beta((X_1, \ldots, X_j) \in A_j, (X_1, \ldots, X_n) \in C),\end{aligned}$$

where the last equality is because P_β is exchangeable and A_j is invariant under permutations of S^j. In turn, this implies

$$\begin{aligned}&P_\sigma(T > n, \phi(X_1, \ldots, X_n) \in C) \\ &\quad = \sum_{j \geq n} P_\sigma(T = j + 1, \phi(X_1, \ldots, X_n) \in C) \\ &\quad = \sum_{j \geq n} P_\beta(T = j + 1, (X_1, \ldots, X_n) \in C) \\ &\quad = \sum_{j \geq n} P_\sigma(T = j + 1, (X_1, \ldots, X_n) \in C) \\ &\quad = P_\sigma(T > n, (X_1, \ldots, X_n) \in C).\end{aligned}$$

This concludes the proof. □

PROOF OF THEOREM 11. Just note that g_n is a density of $(X_1, \ldots, X_n)$ with respect to λ^n. Therefore, Theorem 11 follows from the very definitions of stationarity

and exchangeability, after noting that $\int g_{n+1}(u,\cdot)\lambda(du)$ is a density of $(X_2,\dots,X_{n+1})$ with respect to λ^n. □

PROOF OF THEOREM 13. We first recall that

$$\int \mathcal{S}(x,b;A)\mathcal{S}(0,r;dx) = \mathcal{S}(0,b+r;A)$$

for all $A \in \mathcal{B}$ and $b, r > 0$. This can be checked by a direct calculation; see the Claim contained in the proof of Theorem 3 of [5]. Having noted this fact, define

$$\mu = \mathcal{S}(a,b), \quad f(x) = -a + cx, \quad \nu = \mathcal{S}\left(0, \frac{b}{1-|c|^\gamma}\right),$$

and denote by Z a real random variable such that $Z \stackrel{d}{=} \mathcal{S}(0,1)$. Define also

$$r = \frac{b|c|^\gamma}{1-|c|^\gamma}, \quad h(x) = cx,$$

and call ν^* the probability distribution of h under ν. On noting that

$$a + b^{1/\gamma} Z \stackrel{d}{=} \mu \quad \text{and} \quad \nu^* = \mathcal{S}(0,r),$$

one obtains

$$\begin{aligned}
\int \sigma_1(x,A)\nu(dx) &= \int P\big(f(x) + a + b^{1/\gamma} Z \in A\big)\nu(dx) \\
&= \int P\big(h(x) + b^{1/\gamma} Z \in A\big)\nu(dx) \\
&= \int P\big(x + b^{1/\gamma} Z \in A\big)\nu^*(dx) \\
&= \int \mathcal{S}(x,b;A)\mathcal{S}(0,r;dx) \\
&= \mathcal{S}(0,b+r;A) \\
&= \mathcal{S}\left(0, \frac{b}{1-|c|^\gamma}; A\right) = \nu(A).
\end{aligned}$$

Therefore, equation (9) holds. □

ACKNOWLEDGMENTS

We are grateful to Federico Bassetti and Paola Bortot for very useful conversations.

REFERENCES

[1] AIROLDI, E. M., COSTA, T., BASSETTI, F., LEISEN, F. and GUINDANI, M. (2014). Generalized species sampling priors with latent beta reinforcements. *J. Amer. Statist. Assoc.* **109** 1466–1480. MR3293604 https://doi.org/10.1080/01621459.2014.950735

[2] BASSETTI, F., CRIMALDI, I. and LEISEN, F. (2010). Conditionally identically distributed species sampling sequences. *Adv. in Appl. Probab.* **42** 433–459. MR2675111 https://doi.org/10.1239/aap/1275055237

[3] BASSETTI, F. and LADELLI, L. (2020). Asymptotic number of clusters for species sampling sequences with non-diffuse base measure. *Statist. Probab. Lett.* **162** 108749, 7. MR4079614 https://doi.org/10.1016/j.spl.2020.108749

[4] BERTI, P., DREASSI, E., LEISEN, F., PRATELLI, L. and RIGO, P. (2023). Kernel based Dirichlet sequences. *Bernoulli* **29** 1321–1342. MR4550225 https://doi.org/10.3150/22-BEJ1500

[5] BERTI, P., DREASSI, E., LEISEN, F., PRATELLI, L. and RIGO, P. (2023). Bayesian predictive inference without a prior. *Statist. Sinica* **33**. https://doi.org/10.5705/ss.202021.0238

[6] BERTI, P., DREASSI, E., PRATELLI, L. and RIGO, P. (2021). A class of models for Bayesian predictive inference. *Bernoulli* **27** 702–726. MR4177386 https://doi.org/10.3150/20-BEJ1255

[7] BERTI, P., PRATELLI, L. and RIGO, P. (2004). Limit theorems for a class of identically distributed random variables. *Ann. Probab.* **32** 2029–2052. MR2073184 https://doi.org/10.1214/009117904000000676

[8] BERTI, P., PRATELLI, L. and RIGO, P. (2012). Limit theorems for empirical processes based on dependent data. *Electron. J. Probab.* **17** no. 9, 18. MR2878788 https://doi.org/10.1214/EJP.v17-1765

[9] BERTI, P., PRATELLI, L. and RIGO, P. (2013). Exchangeable sequences driven by an absolutely continuous random measure. *Ann. Probab.* **41** 2090–2102. MR3098068 https://doi.org/10.1214/12-AOP786

[10] BERTI, P., REGAZZINI, E. and RIGO, P. (1997). Well-calibrated, coherent forecasting systems. *Theory Probab. Appl.* **42** 82–102.

[11] BLACKWELL, D. and MACQUEEN, J. B. (1973). Ferguson distributions via Pólya urn schemes. *Ann. Statist.* **1** 353–355. MR0362614

[12] BLADT, M. and MCNEIL, A. J. (2022). Time series with infinite-order partial copula dependence. *Depend. Model.* **10** 87–107. MR4426880 https://doi.org/10.1515/demo-2022-0105

[13] CANALE, A., LIJOI, A., NIPOTI, B. and PRÜNSTER, I. (2017). On the Pitman–Yor process with spike and slab base measure. *Biometrika* **104** 681–697. MR3694590 https://doi.org/10.1093/biomet/asx041

[14] CASSESE, A., ZHU, W., GUINDANI, M. and VANNUCCI, M. (2019). A Bayesian nonparametric spiked process prior for dynamic model selection. *Bayesian Anal.* **14** 553–572. MR3934097 https://doi.org/10.1214/18-BA1116

[15] CHEN, K., SHEN, W. and ZHU, W. (2023). Covariate dependent Beta-GOS process. *Comput. Statist. Data Anal.* **180** Paper No. 107662, 13. MR4515770 https://doi.org/10.1016/j.csda.2022.107662

[16] CIFARELLI, D. M. and REGAZZINI, E. (1996). De Finetti's contribution to probability and statistics. *Statist. Sci.* **11** 253–282. MR1445983 https://doi.org/10.1214/ss/1032280303

[17] CLARKE, B., FOKOUE, E. and ZHANG, H. H. (2009). *Principles and Theory for Data Mining and Machine Learning*. Springer, New York.

[18] CLARKE, B. S. and CLARKE, J. L. (2018). *Predictive Statistics*: *Analysis and Inference Beyond Models*. *Cambridge Series in Statistical and Probabilistic Mathematics* **46**. Cambridge Univ. Press, Cambridge. MR3791464 https://doi.org/10.1017/9781139236003

[19] DAWID, A. P. (1984). Statistical theory. The prequential approach. *J. Roy. Statist. Soc. Ser. A* **147** 278–292. MR0763811 https://doi.org/10.2307/2981683

[20] DAWID, A. P. (1992). Prequential data analysis. In *Current Issues in Statistical Inference*: *Essays in Honor of D. Basu* (M. Ghosh and P. K. Pathak, eds.). *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **17** 113–126. IMS, Hayward. MR1194413 https://doi.org/10.1214/lnms/1215458842

[21] DAWID, A. P. and VOVK, V. G. (1999). Prequential probability: Principles and properties. *Bernoulli* **5** 125–162. MR1673572 https://doi.org/10.2307/3318616

[22] DE FINETTI, B. (1931). Sul significato soggettivo della probabilità. *Fund. Math.* **17** 298–329.

[23] DE FINETTI, B. (1937). La prévision : Ses lois logiques, ses sources subjectives. *Ann. Inst. Henri Poincaré* **7** 1–68. MR1508036

[24] DIACONIS, P. and FREEDMAN, D. A. (1990). Cauchy's equation and de Finetti's theorem. *Scand. J. Stat.* **17** 235–249. MR1092946

[25] DIACONIS, P. and YLVISAKER, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7** 269–281. MR0520238

[26] DUBINS, L. E. and SAVAGE, L. J. (1965). *How to Gamble If You Must. Inequalities for Stochastic Processes*. McGraw-Hill, New York. MR0236983

[27] EFRON, B. (2020). Prediction, estimation, and attribution. *J. Amer. Statist. Assoc.* **115** 636–655. MR4107663 https://doi.org/10.1080/01621459.2020.1762613

[28] FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

[29] FONG, E., HOLMES, C. and WALKER, S. G. (2023). Martingale posterior distributions (with discussion). *J. Roy. Statist. Soc. Ser. B*. To appear.

[30] FONG, E. and LEHMANN, B. (2022). A predictive approach to Bayesian nonparametric survival analysis. Available at arXiv:2202.10361v1 [stat.ME].

[31] FORTINI, S., LADELLI, L. and REGAZZINI, E. (2000). Exchangeability, predictive distributions and parametric models. *Sankhyā Ser. A* **62** 86–109. MR1769738

[32] FORTINI, S. and PETRONE, S. (2012). Predictive construction of priors in Bayesian nonparametrics. *Braz. J. Probab. Stat.* **26** 423–449. MR2949087 https://doi.org/10.1214/11-BJPS176

[33] FORTINI, S. and PETRONE, S. (2017). Predictive characterization of mixtures of Markov chains. *Bernoulli* **23** 1538–1565. MR3624870 https://doi.org/10.3150/15-BEJ787

[34] 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. MR4136504

[35] FORTINI, S., PETRONE, S. and SPORYSHEVA, P. (2018). On a notion of partially conditionally identically distributed sequences. *Stochastic Process. Appl.* **128** 819–846. MR3758339 https://doi.org/10.1016/j.spa.2017.06.008

[36] GEISSER, S. (1993). *Predictive Inference*: *An Introduction*. *Monographs on Statistics and Applied Probability* **55**. CRC Press, New York. MR1252174 https://doi.org/10.1007/978-1-4899-4467-2

[37] GHOSAL, S. and VAN DER VAART, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. *Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge Univ. Press, Cambridge. MR3587782 https://doi.org/10.1017/9781139029834

[38] GNEDIN, A. (2010). A species sampling model with finitely many types. *Electron. Commun. Probab.* **15** 79–88. MR2606505 https://doi.org/10.1214/ECP.v15-1532

[39] GNEDIN, A. and PITMAN, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci.* **138** 5674–5685.

[40] HAHN, P. R. (2017). Predictivist Bayes density estimation. Unpublished technical report. Available at https://math.la.asu.edu/prhahn/pred-bayes.pdf.

[41] HAHN, P. R., MARTIN, R. and WALKER, S. G. (2018). On recursive Bayesian predictive distributions. *J. Amer. Statist. Assoc.* **113** 1085–1093. MR3862341 https://doi.org/10.1080/01621459.2017.1304219

[42] HANSEN, B. and PITMAN, J. (2000). Prediction rules for exchangeable sequences related to species sampling. *Statist. Probab. Lett.* **46** 251–256. MR1745692 https://doi.org/10.1016/S0167-7152(99)00109-1

[43] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7

[44] HILL, B. M. (1993). Parametric models for A_n: Splitting processes and mixtures. *J. Roy. Statist. Soc. Ser. B* **55** 423–433. MR1224406

[45] HJORT, N. L., HOLMES, C., MÜLLER, P. and WALKER, S. G., eds. (2010). *Bayesian Nonparametrics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **28**. Cambridge Univ. Press, Cambridge. MR2722987 https://doi.org/10.1017/CBO9780511802478

[46] HOFFMANN-JØRGENSEN, J. (1994). *Probability with a View Toward Statistics*, *Vol. II*. Chapman & Hall, New York.

[47] KALLENBERG, O. (1988). Spreading and predictable sampling in exchangeable sequences and processes. *Ann. Probab.* **16** 508–534. MR0929061

[48] LEE, J., QUINTANA, F. A., MÜLLER, P. and TRIPPA, L. (2013). Defining predictive probability functions for species sampling models. *Statist. Sci.* **28** 209–222. MR3112406 https://doi.org/10.1214/12-sts407

[49] LIJOI, A., PRÜNSTER, I. and WALKER, S. G. (2008). Bayesian nonparametric estimators derived from conditional Gibbs structures. *Ann. Appl. Probab.* **18** 1519–1547. MR2434179 https://doi.org/10.1214/07-AAP495

[50] MORVAI, G. and WEISS, B. (2021). On universal algorithms for classifying and predicting stationary processes. *Probab. Surv.* **18** 77–131. MR4255241 https://doi.org/10.1214/20-ps345

[51] NEWTON, M. A. (2002). On a nonparametric recursive estimator of the mixing distribution **64** 306–322. Sankhyā Ser. A, 2, Selected articles from San Antonio Conference in honour of C. R. Rao (San Antonio, TX, 2000). MR1981761

[52] NEWTON, M. A. and ZHANG, Y. (1999). A recursive algorithm for nonparametric analysis with missing data. *Biometrika* **86** 15–26. MR1688068 https://doi.org/10.1093/biomet/86.1.15

[53] PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probab. Theory Related Fields* **102** 145–158. MR1337249 https://doi.org/10.1007/BF01213386

[54] PITMAN, J. (1996). Some developments of the Blackwell–MacQueen urn scheme. In *Statistics*, *Probability and Game Theory*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **30** 245–267. IMS, Hayward. MR1481784 https://doi.org/10.1214/lnms/1215453576

[55] PITMAN, J. (2006). *Combinatorial Stochastic Processes*. *Lecture Notes in Math.* **1875**. Springer, Berlin. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard. MR2245368

[56] PITMAN, J. and YOR, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25** 855–900. MR1434129 https://doi.org/10.1214/aop/1024404422

[57] SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

[58] SHMUELI, G. (2010). To explain or to predict? *Statist. Sci.* **25** 289–310. MR2791669 https://doi.org/10.1214/10-STS330

[59] SMITH, A. F. M. and MAKOV, U. E. (1978). A quasi-Bayes sequential procedure for mixtures. *J. Roy. Statist. Soc. Ser. B* **40** 106–112. MR0512148