Statistica Sinica Preprint No: SS-2021-0238

Title	Bayesian Predictive Inference Without a Prior
Manuscript ID	SS-2021-0238
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0238
Complete List of Authors	Patrizia Berti, Emanuela Dreassi, Fabrizio Leisen, Luca Pratelli and Pietro Rigo
Corresponding Author	Pietro Rigo
E-mail	pietro.rigo@unibo.it
Notice: Accepted version subject to English editing.	

Statistica Sinica

BAYESIAN PREDICTIVE INFERENCE WITHOUT A PRIOR

Patrizia Berti[1], Emanuela Dreassi[2], Fabrizio Leisen[3],

Luca Pratelli[4] and Pietro Rigo[5]

[1] *Università di Modena e Reggio-Emilia,* [2] *Università di Firenze,*

[3] *University of Nottingham,* [4] *Accademia Navale di Livorno and*

[5] *Università di Bologna*

Abstract: Let $(X_n : n \geq 1)$ be a sequence of random observations. Let $\sigma_n(\cdot) = P\big(X_{n+1} \in \cdot \mid X_1, \ldots, X_n\big)$ be the n-th predictive distribution and $\sigma_0(\cdot)=P(X_1 \in \cdot)$ the marginal distribution of X_1. To make predictions on (X_n), a Bayesian forecaster only needs the collection $\sigma = (\sigma_n : n \geq 0)$. 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 has to be selected. This point of view is adopted in this paper. The choice of σ is only subjected to two requirements: (i) The resulting sequence (X_n) is conditionally identically distributed, in the sense of [4]; (ii) Each σ_{n+1} is a simple recursive update of σ_n. Various new σ satisfying (i)-(ii) are introduced and investigated. For such σ, the asymptotics of σ_n, as $n \to \infty$, is determined. In some cases, the probability distribution of (X_n) is also evaluated.

Key words and phrases: Asymptotics, Bayesian nonparametrics, Conditional identity in distribution, Exchangeability, Predictive distribution, Sequential prediction, Total variation distance.

MSC 2020 subject classifications: 62F15, 62M20, 60G25, 60G09.

1. Introduction

Consider a Bayesian forecaster who makes predictions on a sequence $(X_n : n \geq 1)$ of random observations. At each time n, she aims to predict X_{n+1} based on $(X_1, \ldots, X_n)$. To this end, she needs to assign the conditional distribution of X_{n+1} given $(X_1, \ldots, X_n)$, usually called the n-th *predictive distribution*.

To formalize this problem, we fix a measurable space $(S, \mathcal{B})$ and we take X_n to be the n-th coordinate random variable on S^∞, namely

$$X_n(s_1, \ldots, s_n, \ldots) = s_n \qquad \text{for all } n \geq 1 \text{ and } (s_1, \ldots, s_n, \ldots) \in S^\infty.$$

To avoid technicalities, we assume that S is a Borel subset of a Polish space and $\mathcal{B}$ is the Borel σ-field on S. Moreover, following Dubins and Savage [14], we introduce the notion of strategy.

Let $\mathcal{P}$ denote the collection of all probability measures on $\mathcal{B}$. A *strategy* is a sequence $\sigma = (\sigma_0, \sigma_1, \ldots)$ such that

- $\sigma_0 \in \mathcal{P}$ and $\sigma_n = \{\sigma_n(x) : x \in S^n\}$ is a collection of elements of $\mathcal{P}$;
- 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 probability attached to the event A by the probability measure $\sigma_n(x)$.

An important special case is when the strategy σ is dominated by a fixed measure λ on $(S, \mathcal{B})$. This means that $\sigma_n(x)$ has a density with respect to λ, say $f_n(\cdot \mid x)$, for all n and x. Hence, $\sigma_n(x)$ can be written as

$$\sigma_n(x)(A) = \int_A f_n(z \mid x)\, \lambda(dz) \qquad \text{for all } A \in \mathcal{B}.$$

For instance, if S is countable, any strategy σ is dominated by $\lambda =$ counting measure. Or else, if $S = \mathbb{R}$, some meaningful strategies are dominated by $\lambda =$ Lebesgue measure. Clearly, in the dominated case, the strategy σ can be identified with the sequence $(f_0, f_1, \ldots)$ of predictive densities. However, in this paper, we deal with general strategies and dominated strategies are just a (remarkable) special case.

For any strategy σ (dominated or not), there is a unique probability measure P_σ on $(S^\infty, \mathcal{B}^\infty)$ such that

$$P_\sigma(X_1 \in \cdot) = \sigma_0 \quad \text{and} \quad P_\sigma\big(X_{n+1} \in \cdot \mid (X_1, \ldots, X_n) = x\big) = \sigma_n(x)$$
$$\text{for all } n \geq 1 \text{ and } P_\sigma\text{-almost all } x \in S^n.$$

The above result, due to Ionescu-Tulcea, provides the theoretical foundations of Bayesian predictive inference. To make predictions on (X_n), one needs precisely a strategy σ. The Ionescu-Tulcea theorem guarantees that, for *any* σ, the predictions based on σ are consistent with a unique probability distribution P_σ for the data sequence (X_n).

However, (X_n) is usually required some distributional properties suggested by the specific problem under consideration. For instance, (X_n) is asked to be exchangeable, or stationary, or Markov, and so on. In these cases, the strategy σ can not be arbitrary, for P_σ must belong to some given class of probability measures on $(S^\infty, \mathcal{B}^\infty)$.

1.1 Motivations

In a Bayesian framework, (X_n) is typically assumed to be *exchangeable*. In that case, there are essentially two approaches for selecting a strategy σ. For definiteness, as in [8], we call them the *standard approach* (SA) and the *non-standard approach* (NSA). Both are admissible, from the Bayesian point of view, and both lead to a full specification of the probability distribution of (X_n).

According to SA, to obtain σ, one should:

- Select a prior π, namely, a probability measure on $\mathcal{P}$;

- Calculate the posterior of π given that $(X_1, \ldots, X_n) = x$, say $\pi_n(x)$;

- Evaluate σ as

  $$\sigma_n(x)(A) = \int_{\mathcal{P}} p(A)\,\pi_n(x)(dp) \quad \text{for all } A \in \mathcal{B},$$

  where $\pi_0(x)$ is meant as $\pi_0(x) = \pi$.

Instead, according to NSA, the strategy σ can be assigned directly, without passing through the above prior/posterior scheme. Rather than choosing π and evaluating π_n and σ_n, the forecaster merely selects her predictive σ_n. This procedure makes sense because of the Ionescu-Tulcea theorem. See e.g. [3], [5], [8], [9], [12], [15], [16], [17], [19], [20], [21], [23], [25], [26].

The merits and drawbacks of SA and NSA are discussed in [8]. In short, SA is a cornerstone of Bayesian inference but is not motivated by prediction alone. Its main scope is to make inference on other features of the data distribution, such as a random parameter (possibly, infinite dimensional). However, when prediction is the main target, SA is clearly involved. In turn, NSA has essentially four merits.

- NSA requires the assignment of probabilities on *observable facts* only. The next observation X_{n+1} is actually observable, while π and π_n (being probabilities on $\mathcal{P}$) do not deal with observable facts.

- The data sequence (X_n) is not forced to satisfy any distributional assumption. In particular, (X_n) may fail to be exchangeable.

- 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. According to us, this is consistent with the Bayesian view, where the observed data are fixed and one should condition on them. A similar point of view is highlighted in [15].

- NSA is more straightforward than SA when prediction is the main goal. In this case, why select the prior π explicitly ? Rather than wondering about π, it seems reasonable to reflect on how X_{n+1} is affected by $(X_1, \ldots, X_n)$.

The above remarks refer to any (Bayesian) prediction problem, both parametric and nonparametric. However, NSA is especially appealing in the *nonparametric* case, where selecting a prior with large support is usually hard. For instance, NSA is quite natural when dealing with species sampling sequences. Indeed, this paper has been written having the nonparametric framework in mind.

If (X_n) is assumed to be exchangeable, however, NSA has a drawback. To apply NSA with exchangeable data, one should first characterize those strategies σ which make (X_n) exchangeable under P_σ. A nice characterization is [16, Th. 3.1]. However, the conditions on σ for making (X_n)

exchangeable are quite hard to check in real problems.

To bypass this drawback, the exchangeability assumption could be weakened. One option is to assume (X_n) *conditionally identically distributed* (c.i.d.). We refer to Subsection 2.2 for c.i.d. sequences. Here, we just mention a few reasons for taking c.i.d. data into account.

- Essentially, (X_n) is c.i.d. if, at each time n, the future observations $(X_k : k > n)$ are identically distributed given the past $(X_1, \ldots, X_n)$. This assumption is quite natural in several prediction problems.

- The asymptotic behavior of c.i.d. sequences is very similar to that of exchangeable ones.

- A meaningful part of the usual Bayesian machinery can be developed under the sole assumption that (X_n) is c.i.d.; see [15].

- A number of interesting strategies cannot be used if (X_n) is asked to be exchangeable, but are available if (X_n) is only required to be c.i.d.; see e.g. [8]. Furthermore, conditional identity in distribution is more reasonable than exchangeability in a few real problems. Examples occur in various fields, including clinical trials, generalized Polya urns, species sampling models and disease surveillance; see [1], [2], [4], [11].

- It is not hard to characterize the strategies σ which make (X_n) c.i.d. under P_σ; see Theorem 1. Therefore, unlike the exchangeable case, NSA can be easily implemented.

1.2 Our contribution

This paper aims to develop NSA for c.i.d. data. It is the natural follow up of [8] but all results and examples (with the only exception of Example 3) are new. Our main goal is to introduce and investigate new strategies σ having the following two properties:

(i) The sequence (X_n) is c.i.d. under P_σ;

(ii) σ_{n+1} is a simple recursive update of σ_n for each $n \geq 0$.

Condition (i) has been already discussed. Condition (ii) is to obtain a fast online Bayesian prediction, in the spirit of [20]. Ideally, condition (ii) should imply that each predictive can be evaluated through a simple recursion on the previous one.

To make some examples, for all $x = (x_1, \ldots, x_n) \in S^n$ and $y \in S$, write

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

In this notation, (x, y) is a point of S^{n+1}, x is the sub-vector containing the first n coordinates and y is the $(n+1)$-th coordinate. Then, for instance, condition (ii) holds if σ satisfies the recursive equations

$$\sigma_0 = \alpha_0 \quad \text{and} \quad \sigma_{n+1}(x, y) = q_n(x)\, \sigma_n(x) + (1 - q_n(x))\, \alpha_{n+1}(x, y) \qquad (1.1)$$

for all $n \geq 0$, $x \in S^n$ and $y \in S$, where $q_n : S^n \to [0,1]$ is any measurable function and $\alpha = (\alpha_0, \alpha_1, \ldots)$ is a given strategy.

According to (1.1), the predictive $\sigma_{n+1}(x,y)$ is a convex combination of the previous predictive $\sigma_n(x)$ and the new contribution $\alpha_{n+1}(x,y)$, with a weight $q_n(x)$ not depending on the last observation y. A possible interpretation is that, at time $n+1$, after observing (x,y), the next observation is drawn from $\sigma_n(x)$ with probability $q_n(x)$ or from $\alpha_{n+1}(x,y)$ with probability $1 - q_n(x)$. Even if simple, this updating rule is able to model various real situations; see Examples 1−9. Moreover, no prior probability is required. The forecaster has only to choose the weights $q_0, q_1, \ldots$ and the strategy α.

An obvious criticism to (1.1) is that, to calculate σ, the forecaster should first select another strategy α (in addition to the weights $q_0, q_1, \ldots$). And, in general, choosing α is as difficult as choosing σ. This is only partially true, for the choice of α is not so hard in several real situations. Exploiting an idea from [20], for instance, α can be obtained via copulas; see Example 1. Or else, α can be built by iterated conditioning; see Example 2. More importantly, the choice of α is simpler in the Markovian case. In this paper, a strategy α is said to be *Markovian* if

$$\alpha_n(x,y) = \alpha_n^*(y) \qquad \text{for all } n \geq 2,\ x \in S^{n-1} \text{ and } y \in S$$

where $\alpha_n^* : S \to \mathcal{P}$ is any measurable map. With a slight abuse of notation, when α is Markovian, we will write $\alpha_n(y)$ instead of $\alpha_n^*(y)$.

In addition to (ii), σ is required to satisfy condition (i). Our first result is that, if σ satisfies (1.1), then (X_n) is c.i.d. under P_σ provided

$$\sigma_n(x)(A) = \int \alpha_{n+1}(x,y)(A)\,\sigma_n(x)(dy) \qquad \text{for all } n \geq 0,\ x \in S^n \text{ and } A \in \mathcal{B}.$$

Such a condition becomes simpler if α is Markovian. Suppose indeed α is Markovian and recall that a filtration on $(S, \mathcal{B})$ is an increasing sequence $\mathcal{G}_0 \subset \mathcal{G}_1 \subset \ldots \subset \mathcal{B}$ of sub-σ-fields of $\mathcal{B}$. Then, (X_n) is c.i.d. under P_σ if α_{n+1} is the conditional distribution of α_0 given $\mathcal{G}_n$ for all n and some filtration $(\mathcal{G}_n)$. Formally,

$$\alpha_{n+1}(\cdot)(A) = E_{\alpha_0}\big(1_A \mid \mathcal{G}_n\big), \qquad \text{a.s. with respect to } \alpha_0, \tag{1.2}$$

for all $n \geq 0$, all $A \in \mathcal{B}$ and some filtration $(\mathcal{G}_n)$.

For instance, if $\mathcal{G}_n = \mathcal{B}$ for all n, condition (1.2) yields $\alpha_{n+1}(y) = \delta_y$ for all $y \in S$ where δ_y denotes the unit mass at the point y. Indeed, some popular strategies admit representation (1.1) with $\alpha_{n+1}(y) = \delta_y$. Well known examples are Dirichlet sequences, Beta-GOS sequences, exponential smoothing and generalized Polya urns; see [1], [2] and [8, Sect. 4]. In all these cases, (X_n) is c.i.d. under P_σ. At the opposite extreme, if $\mathcal{G}_n$ is the trivial σ-field for all n, condition (1.2) implies $\alpha_{n+1}(y) = \alpha_0$ for all $y \in S$. In this case, under P_σ, (X_n) is i.i.d. with common distribution α_0.

More interestingly, take $\mathcal{G}_n$ to be the σ-field generated by a countable partition $\mathcal{H}_n$ of S, where $H \in \mathcal{B}$ and $\alpha_0(H) > 0$ for all $H \in \mathcal{H}_n$. In this case, condition (1.2) implies

$$\alpha_{n+1}(y) = \sum_{H \in \mathcal{H}_n} 1_H(y)\, \alpha_0(\cdot \mid H) = \alpha_0\big(\cdot \mid H_y^n\big)$$

where H_y^n is the only $H \in \mathcal{H}_n$ such that $y \in H$. Moreover, $\mathcal{G}_n \subset \mathcal{G}_{n+1}$ if the partition $\mathcal{H}_{n+1}$ is finer than $\mathcal{H}_n$. With this choice of α, several meaningful strategies satisfying (i)-(ii) can be obtained. For instance, if $q_n = (n+c)/(n+1+c)$ for some constant $c > 0$, one obtains

$$\sigma_n(x) = \frac{c\,\alpha_0 + \sum_{i=1}^n \alpha_0\big(\cdot \mid H_{x_i}^{i-1}\big)}{n+c}.$$

The above strategy σ is analogous to that of a Dirichlet sequence, i.e.,

$$\beta_n(x) = \frac{c\,\alpha_0 + \sum_{i=1}^n \delta_{x_i}}{n+c}.$$

However, σ and β give rise to completely different behaviors for (X_n). Firstly, (X_n) is exchangeable under P_β and only c.i.d. under P_σ. Secondly, if $G = \big\{X_i = X_j \text{ for some } i \neq j\big\}$, one obtains $P_\sigma(G) = 0$ and $P_\beta(G) = 1$ provided α_0 is non-atomic. We also note that attaching probability zero to G is useful in various applications.

This is just an example. Various other strategies come to the fore with a suitable choice of $\mathcal{H}_n$ and q_n; see Section 3.

In addition to (1.1), a second class of strategies is introduced and investigated in this paper. Let $S = \mathbb{R}$ and u_n a sequence of real numbers such that $0 = u_0 < u_1 < u_2 < \ldots < 1$. Define $f_0(x) = 0$ and

$$f_{n+1}(x, y) = \sqrt{\frac{u_{n+1} - u_n}{1 - u_n}}\; y + \Big(1 - \sqrt{\frac{u_{n+1} - u_n}{1 - u_n}}\Big)\, f_n(x)$$

for all $n \geq 0$, $x \in S^n$ and $y \in S$. Define also a strategy σ as

$$\sigma_n(x) = \mathcal{N}\Big(f_n(x),\, 1 - u_n\Big) \qquad \text{for all } n \geq 0 \text{ and } x \in S^n.$$

Such a σ satisfies condition (ii) since $\sigma_{n+1}(x, y)$ depends only on the last observation y and the mean of $\sigma_n(x)$. As shown in Section 4, σ satisfies condition (i) as well. Moreover, under P_σ, the sequence (X_n) is Gaussian with mean 0, variance 1, and a known covariance structure.

Due to its simple form, the above σ is potentially useful in applications. In addition, σ is just a special case of a larger class of strategies satisfying (i)-(ii). In fact, the normal distribution can be replaced by any symmetric stable law. For instance, the normal could be replaced by the Cauchy if heavier tails are regarded more suitable for prediction.

The last part of the paper is devoted to the asymptotics of σ_n as $n \to \infty$. In fact, because of condition (i), one obtains

$$P_\sigma\big(\sigma_n \to \mu \text{ weakly}\big) = 1$$

for some random probability measure μ on $(S, \mathcal{B})$; see Subsection 2.2. Hence, it is quite natural to investigate μ, and this is exactly the scope of Section 5. We give conditions for $\mu \ll \sigma_0$ a.s., for μ to be degenerate a.s., and for $\|\sigma_n - \mu\| \xrightarrow{a.s.} 0$ where $\|\cdot\|$ is total variation norm.

Finally, some applications are discussed in Section 6.

To make the paper more readable, all the proofs are gathered in the "supplementary material".

2. Preliminaries

2.1 Some further notation

Let λ, $\nu \in \mathcal{P}$. We write $\lambda \ll \nu$ to mean that λ is *absolutely continuous* with respect to ν, namely, $\lambda(A) = 0$ whenever $A \in \mathcal{B}$ and $\nu(A) = 0$. Moreover, λ and ν are *singular* if $\lambda(A) = \nu(A^c) = 0$ for some $A \in \mathcal{B}$.

We denote by x a point of S^n where n is an integer or $n = \infty$. In both cases, x_i is the i-th coordinate of x. If $n = 0$ and σ is a strategy, $\sigma_0(x)$ is meant as $\sigma_0(x) = \sigma_0$. Moreover, if $x \in S^\infty$ and f is any map on S^n, we write $f(x)$ to denote $f(x) = f(x_1, \ldots, x_n)$. In particular,

$$\sigma_n(x) := \sigma_n(x_1, \ldots, x_n) \qquad \text{for all } x \in S^\infty.$$

2.2 Conditional identity in distribution

C.i.d. sequences have been introduced in [4] and [22] and then investigated in various papers; see e.g. [1], [2], [5], [6], [7], [8], [9], [11], [15], [18], [19].

Let P be a probability measure on $(S^\infty, \mathcal{B}^\infty)$. Say that (X_n) is c.i.d.

(or that P is c.i.d.) if $X_2 \sim X_1$ and

$$P\big(X_k \in \cdot \mid X_1, \ldots, X_n\big) = P\big(X_{n+1} \in \cdot \mid X_1, \ldots, X_n\big) \quad \text{a.s. for all } k > n \geq 1.$$

Thus, at each time n, the future observations $(X_k : k > n)$ are identically distributed given the past. This is actually weaker than exchangeability. Indeed, (X_n) is exchangeable if and only if it is stationary and c.i.d.

The asymptotics of c.i.d. sequences is similar to that of exchangeable ones. To see this, suppose P is c.i.d. and define the empirical measures

$$\mu_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \qquad \text{for all } n \geq 1 \text{ and } x \in S^\infty.$$

Define also

$$\mu(x) = \lim_n \mu_n(x) \;\text{ if the limit exists and }\; \mu(x) = \delta_{x_1} \;\text{ otherwise,}$$

where $x \in S^\infty$ and the limit is meant as a weak limit of probability measures.

The random probability measure μ is a meaningful parameter of P (even if not as crucial as in the exchangeable case; see [8, Ex. 17]). In fact,

$$\mu_n(A) \xrightarrow{a.s.} \mu(A) \qquad \text{for each } A \in \mathcal{B}.$$

Moreover, for fixed $n \geq 0$ and $A \in \mathcal{B}$, one obtains

$$E_P\big\{\mu(A) \mid X_1, \ldots, X_n\big\} = P\big(X_{n+1} \in A \mid X_1, \ldots, X_n\big) \text{ a.s.}$$

By martingale convergence, this equality implies

$$P\big(X_{n+1} \in A \mid X_1, \ldots, X_n\big) \xrightarrow{a.s.} \mu(A) \quad \text{for each } A \in \mathcal{B}.$$

We also note that (X_n) is asymptotically exchangeable, in the sense that the probability distribution of the shifted sequence $(X_n, X_{n+1}, \ldots)$ converges weakly to an exchangeable probability measure Q on $(S^\infty, \mathcal{B}^\infty)$. Furthermore, $Q = P$ on the σ-field generated by μ.

Finally, we report from [6] a characterization of c.i.d. sequences in terms of strategies. The next result is fundamental for this paper.

Theorem 1. ([6, Th. 3.1]). *For any strategy σ, (X_n) is c.i.d. under P_σ if and only if*

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

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

Henceforth, we say "P_σ is c.i.d." to mean that "(X_n) is c.i.d. under P_σ".

3. Convex combinations of random probability measures

Let $\alpha = (\alpha_0, \alpha_1, \ldots)$ be a strategy and $q_n : S^n \to [0, 1]$ a sequence of measurable functions, where $n \geq 0$ and q_0 is constant. For easy of notation, we write ν instead of α_0, namely, we fix $\nu \in \mathcal{P}$ and we let $\alpha_0 = \nu$. We also recall that α is *Markovian* if $\alpha_n(x) = \alpha_n^*(x_n)$ for all $n \geq 2$ and $x \in S^n$, where $\alpha_n^* : S \to \mathcal{P}$ is any measurable map. In this case, with a slight abuse of notation, we write $\alpha_n(x_n)$ instead of $\alpha_n^*(x_n)$.

In this section, the strategy σ satisfies equation (1.1), namely

$$\sigma_0 = \nu \quad \text{and} \quad \sigma_{n+1}(x,y) = q_n(x)\,\sigma_n(x) + (1 - q_n(x))\,\alpha_{n+1}(x,y)$$

for all $n \geq 0$, $x \in S^n$ and $y \in S$. Arguing by induction, it follows that

$$\sigma_n(x) = \nu \prod_{j=0}^{n-1} q_j + \sum_{i=1}^{n} \alpha_i(x_1, \ldots, x_i)\,(1 - q_{i-1}) \prod_{j=i}^{n-1} q_j \tag{3.3}$$

for all $n \geq 1$ and $x = (x_1, \ldots, x_n) \in S^n$. In formula (3.3), $\prod_{j=i}^{n-1} q_j$ is meant as 1 when $i = n$, and q_j is a shorthand notation to denote

$$q_j = q_j(x_1, \ldots, x_j).$$

Our first goal is to give conditions for P_σ to be c.i.d.

Theorem 2. *P_σ is c.i.d. provided*

$$\sigma_n(x)(A) = \int \alpha_{n+1}(x,y)(A)\,\sigma_n(x)(dy) \tag{3.4}$$

for all $n \geq 0$, all $A \in \mathcal{B}$ and P_σ-almost all $x \in S^n$. Moreover, if α is Markovian, condition (3.4) follows from

$$\alpha_n(x)(A) = \int \alpha_{n+1}(y)(A)\,\alpha_n(x)(dy) \tag{3.5}$$

for all $n \geq 0$, all $A \in \mathcal{B}$ and ν-almost all $x \in S$.

In the Markovian case, Theorem 2 applies if α_{n+1} is a conditional distribution of ν given $\mathcal{G}_n$ for all n, where $(\mathcal{G}_n)$ is any filtration on $(S, \mathcal{B})$.

Corollary 1. *Let $\mathcal{G}_0 \subset \mathcal{G}_1 \subset \mathcal{G}_2 \subset \ldots \subset \mathcal{B}$ be an increasing sequence of sub-σ-fields of $\mathcal{B}$. If α is Markovian, then P_σ is c.i.d. whenever*

$$\alpha_{n+1}(\cdot)(A) = E_\nu\big(1_A \mid \mathcal{G}_n\big), \qquad \nu\text{-a.s., for all } n \geq 0 \text{ and } A \in \mathcal{B}.$$

We are now able to provide examples of strategies which satisfy equation (1.1) and make (X_n) c.i.d.

Example 1. (Copulas). A simple way to get the strategy α is to exploit an idea by Hahn, Martin and Walker [20]. To this end, we write "density" to mean "density with respect to Lebesgue measure". We also recall that, if C is a bivariate copula and F_1, F_2 are distribution functions on $\mathbb{R}$, then $F(x,y) = C\big\{F_1(x), F_2(y)\big\}$ is a distribution function on $\mathbb{R}^2$. In addition, if C, F_1 and F_2 have densities, then

$$f(x,y) = c\big\{F_1(x), F_2(y)\big\}\, f_1(x)\, f_2(y), \qquad (x,y) \in \mathbb{R}^2,$$

is a density of F, where c, f_1, f_2 are densities of C, F_1, F_2, respectively.

Having noted this fact, let $S = \mathbb{R}$ and suppose that ν has a density f_0. Moreover, fix a sequence $C_1, C_2, \ldots$ of bivariate copulas with densities $c_1, c_2, \ldots$ For the sake of simplicity, assume $f_0 > 0$ and $c_n > 0$ for all $n \geq 1$. Define $\sigma_0 = \nu$ and denote by F_0 the distribution function corresponding to σ_0. Next, for each $x \in \mathbb{R}$, let

$$\alpha_1(x)(dz) = f_1(z \mid x)\, dz \quad \text{where} \quad f_1(z \mid x) = c_1\big\{F_0(z),\, F_0(x)\big\}\, f_0(z).$$

Then, define $\sigma_1(x) = q_0\,\sigma_0 + (1-q_0)\,\alpha_1(x)$ and call $F_1(\cdot \mid x)$ the distribution function corresponding to $\sigma_1(x)$. Next, for each $(x,y) \in \mathbb{R}^2$, let $\alpha_2(x,y)(dz) = f_2(z \mid x,y)\,dz$ where

$$f_2(z \mid x,y) = c_2\big\{F_1(z \mid x),\, F_1(y \mid x)\big\}\, f_1(z \mid x).$$

Then, define $\sigma_2(x,y) = q_1(x)\,\sigma_1(x)+(1-q_1(x))\,\alpha_2(x,y)$. In general, suppose $\sigma_n(x)$ has been defined for all $x \in \mathbb{R}^n$ and denote by $f_n(\cdot \mid x)$ and $F_n(\cdot \mid x)$ the density and the distribution function of $\sigma_n(x)$. Then, it suffices to let

$$\alpha_{n+1}(x,y)(dz) = f_{n+1}(z \mid x,y)\,dz \quad \text{for all } x \in \mathbb{R}^n \text{ and } y \in \mathbb{R}$$
$$\text{where} \quad f_{n+1}(z \mid x,y) = c_{n+1}\big\{F_n(z \mid x),\, F_n(y \mid x)\big\}\, f_n(z \mid x).$$

Since $f_{n+1}(\cdot \mid x,y)$ is a density, α is a strategy dominated by Lebesgue measure. In addition, Fubini's theorem yields

$$\begin{aligned}\int \alpha_{n+1}(x,y)(A)\,\sigma_n(x)(dy) &= \int\int_A f_{n+1}(z \mid x,y)\,dz\, f_n(y \mid x)\,dy \\ &= \int_A\int c_{n+1}\big\{F_n(z \mid x),\, F_n(y \mid x)\big\}\, f_n(z \mid x)\, f_n(y \mid x)\,dy\,dz \\ &= \int_A f_n(z \mid x)\,dz = \sigma_n(x)(A) \qquad \text{for all } A \in \mathcal{B}.\end{aligned}$$

Hence, P_σ is c.i.d. because of Theorem 2.

To implement Example 1, one only needs f_0 and the copula densities c_n. Some useful choices of c_n are suggested in [20]. In particular, one can let $c_n = c_1$ for all n. Furthermore, one can use conditional copulas instead of plain copulas, in the sense that c_{n+1} is allowed to depend on the observed data $x \in \mathbb{R}^n$. We also note that, letting $q_n = 0$ for all n, the strategies obtained in [20] are a special case of Example 1.

In Example 1, the idea for building α is borrowed from [20]. A different idea is sketched in the next example.

Example 2. (Iterated conditioning). For each $\tau \in \mathcal{P}$ and each sub-σ-field $\mathcal{G} \subset \mathcal{B}$, let $\tau(\cdot \mid \mathcal{G}) = \{\tau(\cdot \mid \mathcal{G})(x) : x \in S\}$ denote a (regular) version of the conditional distribution of τ given $\mathcal{G}$. This means that $\tau(\cdot \mid \mathcal{G})(x)$ is a probability measure on $\mathcal{B}$, for fixed $x \in S$, and

$$\tau(A \mid \mathcal{G})(\cdot) = E_\tau(1_A \mid \mathcal{G}), \qquad \tau\text{-a.s., for all } A \in \mathcal{B}.$$

For each $n \geq 0$, take a sub-σ-field $\mathcal{G}_n \subset \mathcal{B}$. Define $\sigma_0 = \nu$ and

$$\alpha_1(x) = \nu\big(\cdot \mid \mathcal{G}_0\big)(x) \quad \text{for all } x \in S.$$

To realize equation (1.1), define also $\sigma_1(x) = q_0\,\sigma_0 + (1 - q_0)\,\alpha_1(x)$. Next, for each $(x, y) \in S^2$, define

$$\alpha_2(x, y) = \sigma_1(x)\big(\cdot \mid \mathcal{G}_1\big)(y) \quad \text{and} \quad \sigma_2(x, y) = q_1(x)\,\sigma_1(x) + (1 - q_1(x))\,\alpha_2(x, y).$$

In general, after $\sigma_n(x)$ has been defined for all $x \in S^n$, it suffices to let

$$\alpha_{n+1}(x, y) = \sigma_n(x)\big(\cdot \mid \mathcal{G}_n\big)(y) \qquad \text{for all } x \in S^n \text{ and } y \in S.$$

By construction, this strategy α satisfies condition (3.4). Hence, P_σ is c.i.d. because of Theorem 2.

As an example, take $q_n = \frac{n+c}{n+1+c}$ and $\mathcal{G}_n = \mathcal{G}$ for all $n \geq 0$, where $c > 0$ is a constant and $\mathcal{G} \subset \mathcal{B}$ a sub-σ-field. Then, formula (3.3) yields

$$\sigma_n(x) = \frac{c\,\nu + \sum_{i=1}^{n} \alpha_i(x_1, \ldots, x_i)}{n+c} = \frac{c\,\nu + \sum_{i=0}^{n-1} \sigma_i(x_1, \ldots, x_i)\big(\cdot \mid \mathcal{G}\big)(x_{i+1})}{n+c}.$$

A special case of the latter strategy is discussed in Examples 4-5.

If compared with Example 1, Example 2 replaces the choice of the copula densities c_n with that of the sub-σ-fields $\mathcal{G}_n$. In principle, since the $\mathcal{G}_n$ are completely arbitrary, this provides some more degrees of freedom for modeling real situations. Obviously, however, the practical calculation of $\sigma_n(x)\big(\cdot \mid \mathcal{G}_n\big)(y)$ may be very hard.

We next turn to the Markovian case. *In the rest of this Section, α is Markovian* (i.e., $\alpha_n(x) = \alpha_n(x_n)$ for all $n \geq 2$ and $x \in S^n$).

Example 3. (Example 13 of [8]). 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 $H \in \mathcal{H}_n$. Define

$$\alpha_{n+1}(y) = \sum_{H \in \mathcal{H}_n} 1_H(y)\,\nu(\cdot \mid H) = \nu\big(\cdot \mid H_y^n\big) \qquad \text{for all } y \in S,$$

where H_y^n denotes the only $H \in \mathcal{H}_n$ such that $y \in H$. If $\mathcal{G}_n$ is the σ-field generated by $\mathcal{H}_n$, one obtains $\alpha_{n+1}(\cdot)(A) = E_\nu\big(1_A \mid \mathcal{G}_n\big)$ for all $A \in \mathcal{B}$. Moreover, $\mathcal{G}_n \subset \mathcal{G}_{n+1}$ provided $\mathcal{H}_{n+1}$ is finer than $\mathcal{H}_n$ for all $n \geq 0$ (as we assume). Therefore, P_σ is c.i.d. because of Corollary 1.

Example 3 can be developed in various ways. For any partition $\mathcal{H}$ of

S, let

$$\mathcal{U}(\mathcal{H}) = \sup_{H\in\mathcal{H}} \sup_{y,z\in H} d(y,z) \qquad \text{where } d \text{ is the distance on } S.$$

Example 4. (Dirichlet-like sequences). Fix a constant $c > 0$ and define

$$q_n = \frac{n+c}{n+1+c}, \quad \alpha_{n+1}(y) = \nu\big(\cdot \mid H_y^n\big), \quad \nu_n(x) = \frac{\sum_{i=1}^n \nu\big(\cdot \mid H_{x_i}^{i-1}\big)}{n}.$$

Then, formula (3.3) yields

$$\sigma_n(x) = \frac{c\,\nu + \sum_{i=1}^n \nu\big(\cdot \mid H_{x_i}^{i-1}\big)}{n+c} = \frac{c}{n+c}\,\nu \,+\, \frac{n}{n+c}\,\nu_n(x).$$

In turn, the predictives of a Dirichlet sequence are

$$\beta_n(x) = \frac{c}{n+c}\,\nu \,+\, \frac{n}{n+c}\,\mu_n(x)$$

where $\mu_n(x) = (1/n)\sum_{i=1}^n \delta_{x_i}$ is the empirical measure. The strategies σ and β have a similar structure. Moreover, $\sigma_n(x)$ and $\beta_n(x)$ are usually close for large n. In fact, for various distances D on $\mathcal{P}$, one obtains

$$\lim_n D\big[\sigma_n(x),\, \beta_n(x)\big] = 0 \qquad \text{for each } x \in S^\infty \tag{3.6}$$

provided $\lim_n \mathcal{U}(\mathcal{H}_n) = 0$. For instance, relation (3.6) holds if D is the bounded Lipschitz metric; see Theorem 4. Despite (3.6), however, σ and β conflict under a fundamental aspect. Indeed, $\sigma_n(x) \ll \nu$ for all $n \geq 0$ and $x \in S^n$ while this is not true for $\beta_n(x)$. As a consequence, P_σ and P_β are even singular when ν is non-atomic; see Theorem 4 again.

Example 5. (Example 4 continued). The situation of Example 4 may appear strange. Suppose ν is non-atomic and $\lim_n \mathcal{U}(\mathcal{H}_n) = 0$. On one hand, since P_σ and P_β are singular, σ and β induce completely different distributions on the data. On the other hand, because of (3.6), σ and β provide similar predictions for large n.

Such a situation mostly depends on the distance D. In fact, $\sigma_n(x)$ and $\beta_n(x)$ are no longer close if D is replaced by some stronger distance on $\mathcal{P}$, such as the total variation distance.

More precisely, suppose the target is to predict $f(X_{n+1})$ based on $(X_1, \ldots, X_n)$, where $f : S \to \mathbb{R}$ is a bounded measurable function. Then, σ and β actually yield similar predictions for large n. As an example, if f is Lipschitz and D is the bounded Lipschitz metric, one obtains

$$\Big| E_\sigma \Big\{ f(X_{n+1}) \mid (X_1, \ldots, X_n) = x \Big\} - E_\beta \Big\{ f(X_{n+1}) \mid (X_1, \ldots, X_n) = x \Big\} \Big| \\ = \Big| \int f(t)\, \sigma_n(x)(dt) - \int f(t)\, \beta_n(x)(dt) \Big| \leq k\, D\big[\sigma_n(x),\, \beta_n(x)\big]$$

for some constant k depending only on f.

However, σ and β give conflicting predictions in more elaborated problems. For instance, suppose one aims to predict whether the next observation is new. Letting $G_n = \{X_{n+1} = X_i \text{ for some } i \leq n\}$, one obtains

$$P_\sigma\big(G_n \mid (X_1, \ldots, X_n) = x\big) = \sigma_n(x)(\{x_1, \ldots, x_n\}) = 0 \quad \text{while}$$

$$P_\beta\big(G_n \mid (X_1, \ldots, X_n) = x\big) = \beta_n(x)(\{x_1, \ldots, x_n\}) = n/(n+c).$$

Example 6. (Exponential smoothing-like sequences). Let

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

where $q \in [0,1]$ is any constant. Making predictions through β 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 i-th observation x_i should be an increasing function of i; see [2] and [8]. Now, if $q_n = q$ and $\alpha_{n+1}(y) = \nu\big(\cdot \mid H_y^n\big)$, formula (3.3) reduces to

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

Essentially the same remarks of Examples 4-5, about the connections between σ and β, can be repeated in this example.

Exploiting countable partitions is a flexible idea which may be realized in various ways. We support this claim with two further examples.

Example 7. (Mixed strategies). Let $\mathcal{H} \subset \mathcal{B}$ be a fixed countable partition of S and $A_0 \subset A_1 \subset A_2 \subset \ldots$ an increasing sequence of elements of $\mathcal{B}$. Assume $\nu(A_n^c \cap H) > 0$ whenever $A_n^c \cap H \neq \emptyset$ and define

$$\alpha_{n+1}(y) = 1_{A_n}(y)\,\delta_y + 1_{A_n^c}(y)\,\nu\big(\cdot \mid A_n^c \cap H_y\big)$$

where H_y is the only $H \in \mathcal{H}$ such that $y \in H$. Then, α_{n+1} satisfies Corollary 1 with $\mathcal{G}_n$ the σ-field generated by the sets $A \cap A_n$ and $H \cap A_n^c$ for all $A \in \mathcal{B}$ and $H \in \mathcal{H}$. Since $\mathcal{G}_n \subset \mathcal{G}_{n+1}$ for all n, it follows that P_σ is c.i.d.

For instance, take

$$S=\mathbb{R}, \quad \mathcal{H}=\big\{(-\infty,0),\,\{0\},\,(0,\infty)\big\}, \quad A_n=[-u_n,u_n]$$

where $0<u_0<u_1<u_2<\ldots$ are any constants. Suppose also that, at each time n, an observation y is informative about the future observations whenever $|y|\le u_n$. Otherwise, if $|y|>u_n$, the only relevant information provided by y is its sign. Then, choosing α_{n+1} as above may be reasonable. Finally, taking q_n as in Example 4, one obtains

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

$$=\frac{c\,\nu\,+\,\sum_{i=0}^{n-1}\Big\{1_{A_i}(x_{i+1})\delta_{x_{i+1}}\,+\,1_{B_i}(x_{i+1})\,\nu(\cdot\mid B_i)\,+\,1_{C_i}(x_{i+1})\,\nu(\cdot\mid C_i)\Big\}}{n+c}$$

where $B_i=(-\infty,-u_i)$ and $C_i=(u_i,\infty)$.

Example 8. (Occupancy models). At each time n, a random integer r_n is selected and r_n particles are randomly placed into p boxes. The i-th observation is $x_i=(j_1(i),\ldots,j_p(i))$ where $j_k(i)$ is the number of particles in box k at time i. To model this situation, set $S=\{0,1,2,\ldots\}^p$ and take $\mathcal{H}_n=\mathcal{F}$ for all n, where $\mathcal{F}$ is the partition of S with elements

$$F_r=\big\{(j_1,\ldots,j_p)\in S:\sum_{k=1}^p j_k=r\big\}\qquad\text{for } r=0,1,2,\ldots$$

Moreover, take ν to be the probability distribution of $(Y_1,\ldots,Y_p)$ where $Y_1,\ldots,Y_p$ are i.i.d. Poisson random variables. The conditional distribution of $(Y_1, \ldots, Y_p)$ given $\sum_{k=1}^p Y_k = r$ is multinomial with index r and equal cell probabilities $1/p$. Therefore,

$$\nu\Big(\{(j_1, \ldots, j_p)\} \mid F_r\Big) = \frac{1}{p^r} \frac{r!}{j_1! \ldots j_p!} \quad \text{for all } (j_1, \ldots, j_p) \in F_r.$$

Hence, denoting by $x_i^* = \sum_{k=1}^p j_k(i)$ the sum of the coordinates of x_i, the strategy σ is

$$\sigma_n(x) = \nu \prod_{j=0}^{n-1} q_j + \sum_{i=1}^{n} \nu\left(\cdot \mid F_{x_i^*}\right) (1 - q_{i-1}) \prod_{j=i}^{n-1} q_j.$$

The choice of q_j depends on the specific problem at issue. For instance, q_j could be as in Examples 4, 6 or 9. We just note that exchangeability is useful in the framework of occupancy models, and P_σ is actually exchangeable (and not only c.i.d.) if $q_j = (j + c)/(j + 1 + c)$; see [10] and [13].

Our last example deals with a more elaborate choice of q_n.

Example 9. (Reinforcements). For each $n \geq 1$, fix a set $C_n \in \mathcal{B}^n$, two constants $0 < a_n < 1/2 < b_n < 1$, and define

$$q_n(x) = b_n \, 1_{C_n}(x) + a_n \, (1 - 1_{C_n}(x)) \qquad \text{for all } x \in S^n.$$

Roughly speaking, the underlying idea is that $\sigma_n(x)$ exhibits good predictive performances whenever $x \in C_n$. Therefore, if $(X_1, \ldots, X_{n+1}) = (x, y)$ and $x \in C_n$, to predict X_{n+2} the forecaster is inclined to reinforce $\sigma_n(x)$ with respect to $\alpha_{n+1}(y)$. (Recall that $a_n < 1/2 < b_n$).

As a concrete example, let $S = [0,1]$ and $q_0 = 1/2$. For all $n \geq 1$ and $x \in S^n$, let $\overline{x}_n = (1/n)\sum_{i=1}^n x_i$ be the sample mean of x and $m_n(x)$ any (measurable) predictor of X_{n+1} based on $\sigma_n(x)$. For definiteness,

$$m_n(x) = \int t\,\sigma_n(x)(dt).$$

If $m_n(x)$ is regarded as a predictor of the past observations x_i, $i \leq n$, then

$$\overline{x}_n - m_n(x) = (1/n)\sum_{i=1}^n \{x_i - m_n(x)\}$$

is the arithmetic mean of the prediction errors. In a sense, $\sigma_n(x)$ works nicely whenever $\overline{x}_n - m_n(x)$ is small. Hence, given $\epsilon > 0$, one could let

$$C_n = \{x \in S^n : |\overline{x}_n - m_n(x)| < \epsilon\}.$$

To close this section, we note that the strategies obtained so far have applications beyond the predictive framework of this paper. In fact, various species sampling sequences correspond to strategies of the form (3.3). And, in Bayesian nonparametrics, species sampling sequences may be used to define priors; see [1].

4. Predictions via stable laws

In this section, we let $S = \mathbb{R}$, we fix a constant $\gamma \in (0,2]$, and we introduce a further class of strategies. Such strategies need not satisfy equation (1.1) (unless $q_n = 0$ for all n). However, they meet conditions (i)-(ii) and the probability measure $\sigma_n(x)$ is γ-stable for all $n \geq 0$ and $x \in S^n$. (The exponent γ of a stable law is usually denoted by α, but in this paper α denotes a strategy).

Let Z be a real random variable with characteristic function

$$E\big\{\exp(i\,t\,Z)\big\} = \exp\Big(-\frac{|t|^\gamma}{2}\Big) \qquad \text{for all } t \in \mathbb{R}.$$

For $a \in \mathbb{R}$ and $b > 0$, denote by $\mathcal{S}(a,b)$ the probability distribution of $a + b^{1/\gamma}Z$, namely

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

Note that $\mathcal{S}(a,b) = \mathcal{N}(a,b)$ if $\gamma = 2$, where $\mathcal{N}(a,b)$ is the Gaussian law on $\mathcal{B}$ with mean a and variance b. Similarly, $\mathcal{S}(a,b) = \mathcal{C}(a,b)$ if $\gamma = 1$, where $\mathcal{C}(a,b)$ is the probability measure on $\mathcal{B}$ with density $f(x) = \frac{2\,b}{\pi}\,\frac{1}{b^2+4\,(x-a)^2}$. (Incidentally, in this parametrization, the standard Cauchy distribution is $\mathcal{C}(0,2)$ and not $\mathcal{C}(0,1)$).

Next, fix the real numbers

$$0 = u_0 < u_1 < u_2 < \ldots < u,$$

and define $f_0 = 0$ and

$$f_{n+1}(x,y) = f_n(x)\,\left(1 - \left(\frac{u_{n+1}-u_n}{u-u_n}\right)^{1/\gamma}\right) + y\,\left(\frac{u_{n+1}-u_n}{u-u_n}\right)^{1/\gamma}$$

for all $n \geq 0$, $x \in S^n$ and $y \in S$.

In this section, we focus on the strategy

$$\sigma_n(x) = \mathcal{S}\Big(f_n(x),\, u - u_n\Big) \qquad \text{for all } n \geq 0 \text{ and } x \in S^n. \tag{4.7}$$

It is worth noting that $\sigma_0 = \mathcal{S}(0, u)$ and $\sigma_{n+1}(x, y)$ can be easily evaluated based on y and the median of $\sigma_n(x)$. Hence, condition (ii) holds. We now turn to condition (i).

Theorem 3. *If σ is given by* (4.7)*, then P_σ is c.i.d.*

In the rest of this section, σ denotes the strategy (4.7).

An useful feature of σ is its asymptotic behavior. Define

$$L = \big\{x \in S^\infty : \lim_n f_n(x) \text{ exists and is finite}\big\}$$

and $f(x) = \lim_n f_n(x)$ for each $x \in L$. Since P_σ is c.i.d., it follows that $P_\sigma(L) = 1$. Moreover, for each $x \in L$, one obtains

$$\sigma_n(x) \longrightarrow \delta_{f(x)} \text{ weakly if } \sup_n u_n = u \text{ and}$$

$$\sigma_n(x) \longrightarrow \mathcal{S}\Big(f(x),\, u - \sup_n u_n\Big) \text{ in total variation if } \sup_n u_n < u.$$

We refer to the proof of Theorem 6 for more details. Here, we turn to examples.

Example 10. (Cauchy and Normal distributions). The most popular cases are $\gamma = 1$ and $\gamma = 2$. Indeed,

$$\sigma_n(x) = \mathcal{C}\Big(f_n(x),\, u - u_n\Big) \quad \text{or} \quad \sigma_n(x) = \mathcal{N}\Big(f_n(x),\, u - u_n\Big)$$

according to whether $\gamma = 1$ or $\gamma = 2$. Both strategies can be useful in real problems. Note also that $f_n(x)$ is just a weighted average of the first n observations $x_1, \ldots, x_n$ and, in the normal case, the weights are connected to the conditional variances.

The next example provides further information on the sequence (X_n).

Example 11. (Finite dimensional distributions). Let

$$Y_{n+1} = \sum_{i=1}^{n} (u_i - u_{i-1})^{1/\gamma}\, Z_i + (u - u_n)^{1/\gamma}\, Z_{n+1} \quad \text{for all } n \geq 0,$$

where $Z_1, Z_2, \ldots$ is an i.i.d. sequence with $Z_1 \sim \mathcal{S}(0,1)$. Then, $Y_1 \sim \mathcal{S}(0,u)$. Furthermore,

$$(Y_1, \ldots, Y_n) = g_n(Z_1, \ldots, Z_n) \quad \text{and} \quad \sum_{i=1}^{n} (u_i - u_{i-1})^{1/\gamma}\, Z_i = f_n(Y_1, \ldots, Y_n)$$

where g_n is an invertible linear transformation. Therefore,

$$\begin{aligned} P\big(Y_{n+1} \in \cdot \mid Y_1, \ldots, Y_n\big) &= P\big(Y_{n+1} \in \cdot \mid Z_1, \ldots, Z_n\big) \\ &= P\,\big(\, f_n(Y_1, \ldots, Y_n) + (u - u_n)^{1/\gamma}\, Z_{n+1} \in \cdot \mid Z_1, \ldots, Z_n\big) \\ &= \mathcal{S}\Big(f_n(Y_1, \ldots, Y_n),\, u - u_n\Big) = \sigma_n(Y_1, \ldots, Y_n) \qquad \text{a.s.} \end{aligned}$$

In other terms, the predictive distributions of the sequence (Y_n) agree with those of σ, and this implies

$$P_\sigma(B) = P\big((Y_1, Y_2, \ldots) \in B\big) \qquad \text{for all } B \in \mathcal{B}^\infty.$$

This equation allows to determine the finite dimensional distributions of (X_n) under P_σ. Here, we just highlight two facts. Firstly,

$$f_n(Y_1,\ldots,Y_n)=\sum_{i=1}^n (u_i-u_{i-1})^{1/\gamma}\,Z_i\sim u_n^{1/\gamma}\,Z_1\sim\mathcal{S}(0,u_n).$$

Thus, $f_n\sim\mathcal{S}(0,u_n)$ under P_σ, namely, $P_\sigma(f_n\in A)=\mathcal{S}(0,u_n)(A)$ for all $A\in\mathcal{B}$. Secondly, since g_n is linear, the finite dimensional distributions of (X_n) under P_σ are Gaussian when $\gamma=2$. In this case, since (Y_n) is c.i.d., the moments are

$$E_{P_\sigma}(X_n)=0,\quad E_{P_\sigma}(X_n^2)=u\qquad\text{and}$$

$$E_{P_\sigma}(X_nX_m)=E(Y_nY_m)=E\big[Y_n\,E(Y_m\mid Y_1,\ldots,Y_n)\big]$$

$$=E(Y_n\,Y_{n+1})=u_{n-1}+\sqrt{(u_n-u_{n-1})(u-u_{n-1})}\qquad\text{for all } 1\le n<m.$$

The last example collects some miscellaneous remarks.

Example 12. (Choice of γ, u and u_n). To work with σ, one has only to select γ and $u,u_1,u_2,\ldots$ Obviously, the choice of γ depends on the specific problem at hand. We just note that, in applications, $\gamma\in\{1,2\}$ is not the unique meaningful choice. For instance, $\gamma\notin\{1,2\}$ is quite common when modeling financial data; see e.g. [24, Chap. 13]. The numbers u and u_n are scale parameters which control the dispersion structure of (X_n). If $\gamma=2$, for instance, u and u_n determine the variances and covariances of the Gaussian sequence (X_n); see Example 11. An important distinction is $\sup_n u_n=u$ or

$\sup_n u_n < u$, as the limiting distribution of σ_n is degenerate in the former case while it is not in the latter. Finally, we mention a practically useful choice of u_n. Fix $u > 0$ and $q \in (0, 1)$ and define

$$u_n = u\,(1 - q^n) \qquad \text{for all } n \geq 0.$$

Then, $u_{n+1} - u_n = (u - u_n)(1 - q)$ and the updating rule for f_n reduces to

$$f_{n+1}(x, y) = (1 - b)\, f_n(x) + b\, y \qquad \text{where } b = (1 - q)^{1/\gamma}.$$

Equivalently, $f_n(x) = b \sum_{j=1}^{n} (1 - b)^{n-j} x_j$ for each $x \in S^n$.

5. Asymptotics

We first recall two popular distances on $\mathcal{P}$. Let λ_1, $\lambda_2 \in \mathcal{P}$ and let F be the set of all functions $f : S \to [-1, 1]$ such that $|f(y) - f(z)| \leq d(y, z)$ for all y, $z \in S$, where d is the distance on S. The *bounded Lipschitz metric* and the *total variation* distance are, respectively,

$$D(\lambda_1, \lambda_2) = \sup_{f \in F} \left| \int f\, d\lambda_1 - \int f\, d\lambda_2 \right| \quad \text{and} \quad \|\lambda_1 - \lambda_2\| = \sup_{A \in \mathcal{B}} |\lambda_1(A) - \lambda_2(A)|.$$

It is not hard to see that $D \leq 2\,\|\cdot\|$. Moreover, D metrizes weak convergence of probability measures, in the sense that, for all λ_n, $\lambda \in \mathcal{P}$,

$$\lambda_n \to \lambda \text{ weakly} \quad \Leftrightarrow \quad \lim_n D(\lambda_n, \lambda) = 0.$$

We next make precise some claims made in Example 4.

Theorem 4. *Let σ and β be as in Example 4. If $\lim_n \mathcal{U}(\mathcal{H}_n) = 0$, then*

$$\lim_n D\big[\sigma_n(x),\, \beta_n(x)\big] = 0 \qquad \text{for each } x \in S^\infty.$$

Moreover, P_σ and P_β are singular if ν is non-atomic.

Next, for each $x \in S^\infty$, define

$$\mu(x) = \lim_n \mu_n(x) \text{ if the limit exists and } \mu(x) = \delta_{x_1} \text{ otherwise,}$$

where $\mu_n(x) = (1/n)\sum_{i=1}^n \delta_{x_i}$ is the empirical measure and the limit is meant as a weak limit of probability measures. The random probability measure μ is a meaningful object. In fact,

$$P_\sigma\Big\{x \in S^\infty : \sigma_n(x) \to \mu(x) \text{ weakly}\Big\} = 1$$

for any strategy σ such that P_σ is c.i.d.; see Subsection 2.2. In the sequel, we investigate μ when σ comes from Sections 3-4.

For each $\tau \in \mathcal{P}$, say that τ is degenerate if $\tau = \delta_z$ for some $z \in S$. The abbreviation "a.s." stands for "P_σ-a.s." For instance, $\mu \ll \tau$ a.s. means $\mu(x) \ll \tau$ for P_σ-almost all $x \in S^\infty$. Recall also that $q_n(x) = q_n(x_1, \ldots, x_n)$ for all $x \in S^\infty$.

Theorem 5. *If the strategy σ satisfies equation* (1.1)*, then $\sigma_n(x)$ converges in total variation distance for each $x \in S^\infty$ such that $\sum_n (1 - q_n(x)) < \infty$. Moreover, if σ is as in Example 3, then:*

- *$\mu \ll \nu$ a.s. and $\lim_n \|\sigma_n - \mu\| = 0$ a.s. provided $\sum_n (1-q_n) < \infty$ a.s.;*
- *μ is degenerate a.s. provided $\lim_n \mathcal{U}(\mathcal{H}_n) = 0$ and there are constants $a > 0$ and $c_n \geq 0$ such that*

$$\sum_n c_n^2 = \infty \quad \textit{and} \quad a \leq q_n \leq 1 - c_n \ \textit{a.s. for all}\ n \geq 0. \tag{5.8}$$

Theorem 5 can be applied to the examples of Section 3. Suppose in fact $\lim_n \mathcal{U}(\mathcal{H}_n) = 0$. Then, in Example 6, μ is degenerate a.s. In Example 9, $\mu \ll \nu$ a.s. if $\sum_n (1-b_n) < \infty$ and μ is degenerate a.s. if $\sum_n (1-b_n)^2 = \infty$ and $\inf_n a_n > 0$. However, Theorem 5 does not work in Example 4, for in that case $1 - q_n(x) = 1/(n+1+c)$ for all $x \in S^\infty$. Indeed, the behavior of μ in Example 4 is an open problem.

Finally, we turn to the strategies of Section 4.

Theorem 6. *In the notation of Section 4, let*

$$L = \{x \in S^\infty : \lim_n f_n(x) \ \textit{exists and is finite}\},$$

$$f(x) = \lim_n f_n(x) \ \textit{for each}\ x \in L \quad \textit{and} \quad u^* = \sup_n u_n.$$

If σ is the strategy (4.7) then, for each $x \in L$,

$$\sigma_n(x) \longrightarrow \delta_{f(x)} \ \textit{weakly if}\ u^* = u \ \textit{and}$$

$$\sigma_n(x) \longrightarrow \mathcal{S}\big(f(x),\, u - u^*\big) \ \textit{in total variation if}\ u^* < u.$$

Moreover, $P_\sigma(L) = 1$ and $f \sim \mathcal{S}(0, u^*)$ under P_σ, namely

$$P_\sigma(f \in A) = \mathcal{S}(0, u^*)(A) \qquad \textit{for all}\ A \in \mathcal{B}.$$

6. Applications

We finally discuss some applications of the strategies obtained via NSA. We let $S = \mathbb{R}$ and we denote by $x \in \mathbb{R}^n$ the observed data.

Roughly speaking, NSA replaces the choice of the prior with that of a strategy σ; see Section 1. Thus, in general, NSA applies to *any* Bayesian prediction problem. Practically, for any time series (X_n), the forecaster has only to choose her strategy σ. In making this choice, she has no constraints other than her feelings and the specific features of (X_n). Once σ is selected, its possible uses are the usual ones. For instance, the forecaster can build a pointwise predictor for X_{n+1}, such as the mean or the median of $\sigma_n(x)$. Or else, given $\gamma \in (0,1)$, she can build a prediction interval for X_{n+1}, namely an interval $I_n(x)$ such that

$$P_\sigma\Big(X_{n+1} \in I_n(x) \mid (X_1, \ldots, X_n) = x\Big) = \sigma_n(x)\big[I_n(x)\big] \geq 1 - \gamma.$$

The previous remarks, while reasonable, may look generic. Thus, we mention a more concrete application based on *martingale posterior distributions* (m.p.d.'s) as defined in [15]. A m.p.d. is the conditional distribution of θ given the observed data, where $\theta = \theta(X_1, X_2, \ldots)$ is any (measurable) function of the whole data sequence $(X_1, X_2, \ldots)$. Note that θ would be known if we knew $(X_1, X_2, \ldots)$. Hence, the only source of uncertainty is the ignorance about $(X_{n+1}, X_{n+2}, \ldots)$. Quoting from [15, Abst.], a m.p.d.

"returns Bayesian uncertainty directly on any statistic of interest without the need for the likelihood and prior".

In applications, m.p.d.'s can be sampled through a computational scheme, called *predictive resampling* and displayed in Algorithm 1 below. Based on predictive resampling, in [15], several applications to real datasets are provided, including the galaxy and air quality datasets which are classic benchmarks to test new procedures.

Algorithm 1 A practical algorithm for predictive resampling

```
Assign $\sigma_n(x)$ based on the observed data $x = (x_1, \dots, x_n)$
Set $\sigma_n^*(x) = \sigma_n(x)$
$M$ and $N > n$ are integers with $N$ large
for $j \leftarrow 1$ to $M$ do
    for $i \leftarrow n+1$ to $N$ do
        Sample $Y_i \sim \sigma_{i-1}^*$ where $\sigma_{i-1}^* = \sigma_{i-1}^*(x, Y_{n+1}, \dots, Y_{i-1})$
        Update $\sigma_i^* \leftarrow \{\sigma_{i-1}^*, Y_i\}$
    end for
    Compute the empirical measure $\mu_N = \frac{1}{N}\left(\sum_{i=1}^n \delta_{x_i} + \sum_{i=n+1}^N \delta_{Y_i}\right)$
    Compute $\theta_N^{(j)}$ according to $\mu_N$.
end for
Return $\theta_N^{(1)}, \dots, \theta_N^{(M)}$ where the $\theta_N^{(j)}$ are estimates of $\theta$ based on $\mu_N$
```

M.p.d.'s are introduced in the framework of NSA. As in this paper, the predictives are assigned directly and (X_n) is required to be c.i.d. Condition (ii) is very useful as well. Therefore, each of the strategies of Sections 3-4 can be exploited to obtain m.p.d.'s. To implement Algorithm 1, in fact, one needs to sample from a given predictive distribution. In turn, sampling from the strategies of Sections 3-4 is straightforward. In this sense, using such strategies in predictive resampling is computationally efficient.

Acknowledgments: We are grateful to the editors and an anonymous referee for their useful remarks and suggestions.

References

[1] Airoldi E.M., Costa T., Bassetti F., Leisen F., Guindani M. (2014) Generalized species sampling priors with latent beta reinforcements, *J.A.S.A.*, 109, 1466-1480.

[2] Bassetti F., Crimaldi I., Leisen F. (2010) Conditionally identically distributed species sampling sequences, *Adv. Appl. Probab.*, 42, 433-459.

[3] Berti P., Regazzini E., Rigo P. (1997) Well-calibrated, coherent forecasting systems, *Theory Probab. Appl.*, 42, 82-102.

[4] Berti P., Pratelli L., Rigo P. (2004) Limit theorems for a class of identically distributed random variables, *Ann. Probab.*, 32, 2029-2052.

[5] Berti P., Crimaldi I., Pratelli L., Rigo P. (2009) Rate of convergence of predictive distributions for dependent data, *Bernoulli*, 15, 1351-1367.

[6] Berti P., Pratelli L., Rigo P. (2012) Limit theorems for empirical processes based on dependent data, *Electronic J. Probab.*, 17, 1-18.

[7] Berti P., Pratelli L., Rigo P. (2013) Exchangeable sequences driven by an absolutely continuous random measure, *Ann. Probab.*, 41, 2090-2102.

[8] Berti P., Dreassi E., Pratelli L., Rigo P. (2021) A class of models for Bayesian predictive inference, *Bernoulli*, 27, 702-726.

[9] Berti P., Dreassi E., Pratelli L., Rigo P. (2021) Asymptotics of certain conditionally identically distributed sequences, *Statist. Prob. Lett.*, 168, 1-10.

[10] Berti P., Dreassi E., Leisen F., Pratelli L., Rigo P. (2021) Kernel based Dirichlet sequences, *arXiv:2106.00114 [math.PR]*.

[11] Cassese A., Zhu W., Guindani M., Vannucci M. (2019) A Bayesian nonparametric spiked process prior for dynamic model selection, *Bayesian Analysis*, 14, 553-572.

[12] Cifarelli D.M., Regazzini E. (1996) De Finetti's contribution to probability and statistics, *Statist. Science*, 11, 253-282.

[13] Collet F., Leisen F., Spizzichino F., Suter F. (2013) Exchangeable occupancy models and discrete processes with the generalized uniform order statistics property, *Probability in the Engineering and Informational Sciences*, 27, 533-552.

[14] Dubins L.E., Savage L.J. (1965) *How to gamble if you must: Inequalities for stochastic processes*, McGraw Hill.

[15] Fong E., Holmes C., Walker S.G. (2021) Martingale posterior distributions, *arXiv:2103.15671v1*

[16] Fortini S., Ladelli L., Regazzini E. (2000) Exchangeability, predictive distributions and parametric models, *Sankhya* A, 62, 86-109.

[17] Fortini S., Petrone S. (2012) Predictive construction of priors in Bayesian nonparametrics, *Brazilian J. Probab. Statist.*, 26, 423-449.

[18] Fortini S., Petrone S., Sporysheva P. (2018) On a notion of partially conditionally identically distributed sequences, *Stoch. Proc. Appl.*, 128, 819-846.

[19] Fortini S., Petrone S. (2020) Quasi-Bayes properties of a procedure for sequential learning in mixture models, *J. Royal Stat. Soc.* B, 82, 1087-1114.

[20] Hahn P.R., Martin R., Walker S.G. (2018) On recursive Bayesian predictive distributions, *J.A.S.A.*, 113, 1085-1093.

[21] Hill B.M. (1993) Parametric models for A_n: splitting processes and mixtures, *J. Royal Stat. Soc.* B, 55, 423-433.

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

[23] Lee J., Quintana F.A., Muller P., Trippa L. (2013) Defining predictive probability functions for species sampling models, *Statist. Science*, 28, 209-222.

[24] Mc Culloch J.H. (1996) Financial applications of stable distributions, in: *Statistical methods in finance* (Maddala G.S. and Rao C.R. Eds.), North Holland, Elsevier Science BV, Amsterdam.

[25] Pitman J. (1996) Some developments of the Blackwell-MacQueen urn scheme, *Statistics, Probability and Game Theory, IMS Lect. Notes Mon. Series*, 30, 245-267.

[26] Pitman J. (2006) Combinatorial stochastic processes, *Lectures from the XXXII Summer School in Saint-Flour*, 2002, Springer, Berlin.

[1] 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

[2] Dipartimento di Statistica, Informatica, Applicazioni, Università di Firenze, viale Morgagni 59, 50134 Firenze, Italy

E-mail: emanuela.dreassi@unifi.it

[3] School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK

E-mail: fabrizio.leisen@gmail.com

[4] Accademia Navale, viale Italia 72, 57100 Livorno, Italy

E-mail: pratel@mail.dm.unipi.it

[5] Dipartimento di Scienze Statistiche "P. Fortunati", Università di Bologna, via delle Belle Arti 41, 40126 Bologna, Italy

E-mail: pietro.rigo@unibo.it