J. Appl. Probab. 1–27 (2026)
*doi:*10.1017/jpr.2026.10119

ASYMPTOTICS OF PREDICTIVE DISTRIBUTIONS DRIVEN BY SAMPLE MEANS AND VARIANCES

SAMUELE GARELLI,[*] [**] *University of Bologna*
FABRIZIO LEISEN,[***] *King's College London*
LUCA PRATELLI,[****] *Accademia Navale Livorno*
PIETRO RIGO,[*] [*****] *University of Bologna*

Abstract

Let $\alpha_n(\cdot) = \mathbb{P}\big(X_{n+1} \in \cdot \mid X_1, \ldots, X_n\big)$ be the predictive distributions of a sequence $(X_1, X_2, \ldots)$ of p-dimensional random vectors. Suppose $\alpha_n = \mathcal{N}(M_n, Q_n)$, where $M_n = (1/n)\sum_{i=1}^n X_i$ and $Q_n = (1/n)\sum_{i=1}^n (X_i - M_n)(X_i - M_n)^\top$. Then there is a random probability measure α on the Borel subsets of $\mathbb{R}^p$ such that $\|\alpha_n - \alpha\| \xrightarrow{\text{a.s.}} 0$, where $\|\cdot\|$ is the total variation distance. An explicit expression for α is provided and the convergence rate of $\|\alpha_n - \alpha\|$ is shown to be arbitrarily close to $n^{-1/2}$. Moreover, it is still true that $\|\alpha_n - \alpha\| \xrightarrow{\text{a.s.}} 0$ even if $\alpha_n = \mathcal{L}(M_n, Q_n)$, where $\mathcal{L}$ belongs to a class of distributions much larger than the normal. The predictives α_n are useful in various frameworks, including Bayesian predictive inference and predictive resampling. Finally, the asymptotic behavior of copula-based predictive distributions is investigated and a numerical experiment is performed.

Keywords: Bayesian predictive inference; conditional identity in distribution; convergence of probability measures; predictive distribution; predictive resampling; total variation distance

2020 Mathematics Subject Classification: Primary 60B10; 60G57
Secondary 62F15; 62E20

1. Introduction

Let $X = (X_1, X_2, \ldots)$ be a sequence of random variables, with values in a standard Borel space $(S, \mathcal{B})$, and $X(n) = (X_1, \ldots, X_n)$. The *predictive distributions* of X are $\alpha_0(\cdot) = \mathbb{P}(X_1 \in \cdot)$ and $\alpha_n(\cdot) = \mathbb{P}\big[X_{n+1} \in \cdot \mid X(n)\big]$ for each $n \geq 1$. Thus, α_0 is the marginal distribution of X_1 and α_n the conditional distribution of X_{n+1} given $X(n)$. Note that α_0 is a fixed probability measure on $\mathcal{B}$ while α_n is a random probability measure (RPM) on $\mathcal{B}$ for $n \geq 1$.

Received 4 June 2025 UTC; accepted 12 June 2026 UTC.
[*] Postal address: Dipartimento di Scienze Statistiche "P. Fortunati", Università di Bologna, via delle Belle Arti 41, 40126 Bologna, Italy
[**] Email: samuele.garelli2@unibo.it
[***] Postal address: Department of Mathematics, King's College, Strand WC2R 2LS, London, UK. Email: fabrizio.leisen@gmail.com
[****] Postal address: viale Italia 72, 57100 Livorno, Italy. Email: luca_pratelli@marina.difesa.it
[*****] Email: pietro.rigo@unibo.it

By the Ionescu–Tulcea theorem, in order to assess the probability distribution of X, it suffices to specify the sequence (α_n); see, e.g., [2]. More precisely, suppose the X_n are defined on a set Ω and $X(\Omega) = S^\infty$. (The condition $X(\Omega) = S^\infty$ is just to avoid unnecessary complications; it is automatically true if $\Omega = S^\infty$ and the X_n are the canonical projections.) Select any sequence $(\alpha_n : n \geq 0)$ such that:

(i) $\alpha_0 \in \mathcal{P}$ and $\alpha_n : \Omega \to \mathcal{P}$ for $n \geq 1$, where $\mathcal{P}$ is the set of all probability measures on $\mathcal{B}$;

(ii) the map $\omega \in \Omega \mapsto \alpha_n(\omega, A)$ is measurable with respect to $\sigma[X(n)]$ for fixed $n \geq 1$ and $A \in \mathcal{B}$.

Then, there is a unique probability measure $\mathbb{P}$ on $\sigma(X)$ which admits the α_n as predictive distributions. Importantly, apart from conditions (i) and (ii), the α_n are *arbitrary*. Roughly speaking, to assign the distribution of X, it suffices to select (arbitrarily) 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. This fact has various consequences.

For instance, consider a Bayesian forecaster who aims to predict X_{n+1} based on $X(n)$ for every n. To this end, it is (necessary and) sufficient to select (α_n). This suggests a few remarks.

- To make Bayesian predictions, the data sequence X is not forced to have any specific distributional form (such as exchangeable, stationary, Markov, and so on).
- However, in the special case where X is required to be exchangeable, it is possible to make predictions without explicitly selecting a prior. In fact, in the exchangeable case, selecting a prior or selecting the predictives (α_n) are two equivalent strategies to determine the distribution of X. Sometimes, the second strategy (i.e. assigning (α_n) directly) is more convenient.
- By choosing (α_n), the forecaster is attaching probabilities to observable facts only. The value of X_{n+1} is actually observable, while the prior (being a probability on some random parameter) does not necessarily deal with observable facts.
- In some Bayesian problems, even if the primary goal is not prediction, it may be convenient to assign the distribution of X through the choice of (α_n). This happens mainly in Bayesian nonparametrics.

In a nutshell, these are the basic ideas underlying the predictive approach to Bayesian inference; see [1–3] and references therein.

Finally, in addition to Bayesian predictive inference, there are other frameworks where the choice of (α_n) is the main task. Without any claim of being exhaustive, we mention machine learning [6, 12, 13, 20], causal inference [23], species sampling sequences [25, 26], Dawid's prequential approach [8, 9], and martingale posterior distributions [14, 19].

1.1. Convergence of predictive distributions

After selecting the sequence (α_n) of predictives, a natural question is whether it converges, in some sense, as $n \to \infty$. To formalize, suppose X is defined on the probability space $(\Omega, \sigma(X), \mathbb{P})$. Recall that, if μ and ν are probability measures on $\mathcal{B}$, their *total variation distance* is $\|\mu - \nu\| = \sup_{A \in \mathcal{B}} |\mu(A) - \nu(A)|$. Note also that, if α and β are RPMs on $\mathcal{B}$ (defined on $(\Omega, \sigma(X), \mathbb{P})$), the map $\omega \mapsto \|\alpha(\omega, \cdot) - \beta(\omega, \cdot)\|$ is measurable. In fact, since α and β are RPMs, the map $\omega \mapsto |\alpha(\omega, A) - \beta(\omega, A)|$ is measurable for fixed $A \in \mathcal{B}$. Moreover, since $(S, \mathcal{B})$

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

is standard Borel, there is a countable field $\mathcal{B}_0$ such that $\mathcal{B} = \sigma(\mathcal{B}_0)$. Hence, measurability follows from $\mathcal{B}_0$ being countable, and

$$\|\alpha(\omega, \cdot) - \beta(\omega, \cdot)\| = \sup_{A \in \mathcal{B}} |\alpha(\omega, A) - \beta(\omega, A)| = \sup_{A \in \mathcal{B}_0} |\alpha(\omega, A) - \beta(\omega, A)|.$$

That said, let α be an RPM on $\mathcal{B}$. Then α_n *converges weakly to* α *a.s.* (almost surely) if

$$\alpha_n(\omega, \cdot) \xrightarrow{\text{weakly}} \alpha(\omega, \cdot) \quad \text{as } n \to \infty \text{ for } \mathbb{P}\text{-almost all } \omega \in \Omega.$$

Similarly, α_n *converges in total variation to* α *a.s.* if

$$\|\alpha_n(\omega, \cdot) - \alpha(\omega, \cdot)\| \longrightarrow 0 \quad \text{as } n \to \infty \text{ for } \mathbb{P}\text{-almost all } \omega \in \Omega.$$

Clearly, $\alpha_n \to \alpha$ in total variation a.s. implies $\alpha_n \to \alpha$ weakly a.s., but not conversely.

Usually, convergence of α_n is a desirable property. This claim can be supported by various examples and remarks. Here, we report three of them.

Example 1. Suppose $\alpha_n \to \alpha$ weakly a.s. for some RPM α on $\mathcal{B}$. Then $\mu_n \to \alpha$ weakly a.s., where $\mu_n = (1/n) \sum_{i=1}^{n} \delta_{X_i}$ is the empirical measure. Moreover, X is asymptotically exchangeable, in the sense that $(X_n, X_{n+1}, \dots) \xrightarrow{\text{dist}} Z$ as $n \to \infty$, for some exchangeable sequence $Z = (Z_1, Z_2, \dots)$. As an aside, exploiting the convergence of (α_n), we obtain the following characterization of exchangeability: X exchangeable $\iff$ X stationary and (α_n) converges weakly a.s. In fact, '$\Rightarrow$' is very well known; see, e.g., [5]. As to '$\Leftarrow$', just note that stationarity of X implies $X \sim (X_n, X_{n+1}, \dots)$ for each n. Since (α_n) converges weakly a.s., we also obtain $(X_n, X_{n+1}, \dots) \xrightarrow{\text{dist}} Z$ for some exchangeable Z. Hence, $X \sim Z$.

Example 2. Fix a σ-finite measure λ on $\mathcal{B}$ and suppose that $X(n)$ has a density with respect to λ^n for each n. Suppose also that X is exchangeable or, more generally, *conditionally identically distributed* (CID) in the sense of [4]. In this case, there is an RPM α on $\mathcal{B}$ such that $\alpha_n \to \alpha$ weakly a.s., and an obvious question is whether α still admits a density with respect to λ. Precisely, this means that

$$\alpha(\omega, \mathrm{d}x) = f(\omega, x)\, \lambda(\mathrm{d}x) \tag{1}$$

for some non-negative measurable function f and $\mathbb{P}$-almost all $\omega \in \Omega$. By [5, Theorem 1], the answer is: condition (1) holds $\iff$ $\alpha_n \to \alpha$ in total variation a.s.

Example 3. A new method for making Bayesian inference, referred to as *predictive resampling* (Pr), has been introduced in [14]. A brief description of this method is provided in Section 3.1. Here, we just note that, for PR to apply, we have to select a converging sequence (α_n) of predictive distributions.

A last remark is in order. Suppose the inferrer has a specific distribution in mind, that is, they believe that the distribution of X is a specific probability measure ν. In this case, the inferrer should take the α_n as the predictives induced by ν, and there is no reason for $\alpha_n \to \nu$ in some sense. Indeed, α_n could fail to converge. Or else, α_n could converge to a limit different from ν. For instance, if X is exchangeable but not i.i.d. then α_n converges weakly a.s., but the limit is not ν. A related (but different) issue is when X is required to satisfy some distributional assumptions, such as exchangeable, stationary, CID, Markov, and so on. In this case, the choice of (α_n) is not free, since (α_n) should guarantee that X satisfies the requested assumptions. This problem is briefly discussed in [2, Section 1.2].

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

1.2. Our contribution

This paper introduces and investigates a new sequence (α_n) of predictive distributions for the case $S = \mathbb{R}^p$.

It happens quite often that, at time $n = 0$, a dataset $x = (x_1, \ldots, x_s)$ is available (here, $n = 0$ is regarded as the starting time of the inferential procedure). For instance, x may come from previous investigations or any other reliable source. In this paper, *if an initial dataset x is available, such x is regarded as given and the predictives α_n are allowed to depend on it*. Moreover, M_n and Q_n denote the sample mean and covariance matrix based on x and $X(n)$, i.e.

$$M_0 = \frac{1}{s}\sum_{i=1}^{s} x_i, \qquad M_n = \frac{\sum_{i=1}^{s} x_i + \sum_{i=1}^{n} X_i}{n+s},$$

$$Q_0 = \frac{1}{s}\sum_{i=1}^{s} (x_i - M_0)(x_i - M_0)^\top,$$

$$Q_n = \frac{\sum_{i=1}^{s} (x_i - M_n)(x_i - M_n)^\top + \sum_{i=1}^{n} (X_i - M_n)(X_i - M_n)^\top}{n+s}.$$

To define the α_n, further notation is needed. Let $\mathcal{L}(a, B)$ denote the probability distribution of $a + B^{1/2}Z$, where $a \in \mathbb{R}^p$, B is a symmetric non-negative definite matrix of order $p \times p$, and Z is a p-dimensional random vector such that

$$\mathbb{E}(Z) = 0, \qquad \mathbb{E}(ZZ^\top) = I, \qquad \mathbb{E}\{(Z^\top z)^{u/2}\} < \infty \quad \text{for some } u > 2, \text{ and}$$
$$Z \text{ has a density with respect to Lebesgue measure on } \mathbb{R}^p. \tag{2}$$

After selecting one such $\mathcal{L}$, possibly depending on x, the predictives α_n are

$$\alpha_n = \mathcal{L}(M_n, Q_n) \quad \text{for } n \geq 0. \tag{3}$$

Thus, to use the α_n, we only have to select the kernel $\mathcal{L}$, which may or may not depend on x. For instance, independently of x, we can let $\mathcal{L} = \mathcal{N}$ where $\mathcal{N}$ is the Gaussian kernel. Even if obvious, this choice of $\mathcal{L}$ is useful in various cases, including prediction of future observations and parameter inference via PR. But there are many other meaningful choices of $\mathcal{L}$, possibly depending on x. An example is taking $\mathcal{L}$ as a suitable mixture kernel, where the mixture is based on x (see Section 3.3).

Once $\mathcal{L}$ has been selected, α_n depends only on M_n and Q_n. This simple structure reduces computational times and makes the use of α_n very friendly. Moreover, in spite of its simple form, α_n has a good behavior in parameter inference via PR. More precisely, based on numerical experiments, α_n behaves very well when the targets are the mean or the covariance matrix of the underlying population. This could be expected, in a sense, since α_n explicitly depends on M_n and Q_n. However, α_n performs well (and it is not inferior to some other predictive distributions) even for targets such as higher-order moments and quantiles.

Our main result is that α_n converges, in total variation a.s., to an RPM α. An explicit formula for α is provided as well. In addition, if $\mathcal{L} = \mathcal{N}$, the rate of convergence of α_n is shown to be arbitrarily close to $n^{-1/2}$.

If $S = \mathbb{R}$, a natural competitor of α_n are the copula-based predictive distributions introduced in [19] and denoted by β_n in this paper. Hence, the asymptotic behavior of β_n is investigated as well. In Section 2.3, it is shown that β_n converges weakly a.s. but not necessarily in total variation a.s. Moreover, conditions for β_n to converge in total variation a.s. are provided.

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

Finally, in Section 3 the empirical behavior of α_n is tested via numerical experiments. In particular, α_n is compared with β_n as regards the speed of convergence. Moreover, both α_n and β_n are used to implement PR for parameter estimation.

Two final remarks are in order.

Remark 1. To support the α_n, we focused mainly on their behaviour in parameter estimation via PR. However, the α_n's scope is much larger. As an obvious example, in Bayesian predictive inference, the α_n may be used to predict future observations. Or they can be used in all the other frameworks mentioned above.

Remark 2. If the initial dataset x is not available, the α_n can be defined as

$$\alpha_0 = \mathcal{L}(0, I), \qquad \alpha_1 = \mathcal{L}(X_1, I), \qquad \alpha_n = \mathcal{L}(M_n, Q_n) \quad \text{for } n \geq 2,$$

where now

$$M_n = \frac{\sum_{i=1}^n X_i}{n}, \qquad Q_n = \frac{\sum_{i=1}^n (X_i - M_n)(X_i - M_n)^\top}{n}.$$

All the results of this paper are still valid if the α_n are as above.

2. Results

This section includes our main results. Any consideration regarding their practical application is postponed to Section 3. Similarly, to make the paper more readable, all the proofs are deferred to Appendix A.

2.1. Preliminaries and notation

From now on, $S = \mathbb{R}^p$ for some $p \geq 1$ and *the predictives α_n are defined by* (3). Each point $x \in \mathbb{R}^p$ is regarded as a column vector and $x^{(i)}$ denotes the ith coordinate of x. Similarly, if B is any matrix, $B^{(i,j)}$ is the (i,j)th entry of B. We denote by $\mathcal{M}$ the collection of symmetric positive definite matrices of order $p \times p$. We also recall that $\mathcal{L}(a, B)$ denotes the probability distribution of $a + B^{1/2}Z$, where $a \in \mathbb{R}^p$, B is a covariance matrix, and Z is a p-dimensional random vector satisfying condition (2). We will write $\mathcal{M}_p$ or $\mathcal{L}_p(a, B)$ instead of $\mathcal{M}$ or $\mathcal{L}(a, B)$ when a mention of p is necessary.

Finally, given any sequence (γ_n) of predictive distributions, convergence is always meant with respect to the probability measure $\mathbb{P}$ under which X has predictives γ_n. As noted in Section 1, such a $\mathbb{P}$ exists and is unique by the Ionescu–Tulcea theorem.

2.2. Asymptotics of the new predictive distributions

The predictives α_n have a nice asymptotic behavior. This fact is formalized by the next two results, which are the main contributions of this paper.

Theorem 1. *If the α_n are defined by (3) then $M_n \xrightarrow{\text{a.s.}} M$ and $Q_n \xrightarrow{\text{a.s.}} Q$, where M is a p-dimensional random vector and Q a random matrix of order $p \times p$. In addition, $Q \in \mathcal{M}$ a.s., so that $\|\alpha_n - \alpha\| \xrightarrow{\text{a.s.}} 0$, where $\alpha = \mathcal{L}(M, Q)$.*

The informal idea of Theorem 1 is that, if we fix a suitable kernel $\mathcal{L}$ depending on some parameter θ, and plug in an estimate $\hat{\theta}$ in place of θ, then asymptotically things work out. This simple concept underlies various papers; see, e.g., [7], [15, Section 5.2], and [21, p. 8].

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

By Theorem 1, α_n converges in total variation a.s. to $\alpha = \mathcal{L}(M, Q)$, where M and Q are the almost-sure limits of the sample means and the sample covariance matrices, respectively. As noted in Section 1.1, this asymptotic behavior of α_n is useful in various frameworks. To make Theorem 1 more effective, however, it is desirable to grasp some information on the convergence rate of α_n to α. This is possible if $\mathcal{L}$ is the Gaussian kernel $\mathcal{N}$.

Theorem 2. *Let α_n, M, and Q be as in Theorem 1 and suppose $\mathcal{L} = \mathcal{N}$. Define $\alpha = \mathcal{N}(M, Q)$, and fix any sequence (d_n) of constants. Then $d_n \|\alpha_n - \alpha\| \xrightarrow{\mathbb{P}} 0$ whenever $d_n/\sqrt{n} \to 0$.*

To better appreciate Theorem 2, it is worth noting that $\sqrt{n}\|\alpha_n - \alpha\|$ fails to converge to 0 in probability. We prove this fact for $p = 1$. In this case, for each $x \in \mathbb{R}$, we obtain

$$\|\alpha_n - \alpha\| \geq \left|\alpha_n((-\infty, x]) - \alpha((-\infty, x])\right| = \left|\Phi\left(\frac{x - M_n}{\sqrt{Q_n}}\right) - \Phi\left(\frac{x - M}{\sqrt{Q}}\right)\right|,$$

where Φ is the standard normal distribution function. Letting $x = M_n$, it follows that

$$\begin{aligned}\sqrt{n}\|\alpha_n - \alpha\| &\geq \sqrt{n}\left|\Phi(0) - \Phi\left(\frac{M_n - M}{\sqrt{Q}}\right)\right| \\ &= \frac{\sqrt{n}}{\sqrt{2\pi}} \int_0^{|M_n - M|/\sqrt{Q}} \mathrm{e}^{-t^2/2}\, \mathrm{d}t \\ &\geq \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(M_n - M)^2}{2Q}\right) \frac{\sqrt{n}}{\sqrt{Q}} |M_n - M|.\end{aligned}$$

However, $(\sqrt{n}/\sqrt{Q})(M_n - M)$ converges in distribution to $\mathcal{N}(0, 1)$. Hence, the previous inequality implies that $\sqrt{n}\|\alpha_n - \alpha\|$ does not converge to 0 in probability.

2.3. Copula-based predictive distributions

A natural competitor of the α_n are the copula-based predictive distributions of [19].

A bivariate *copula* is a distribution function on $\mathbb{R}^2$ whose marginals are uniform on (0, 1). The density of a bivariate copula, provided it exists, is said to be a *copula density*. (Henceforth, all densities are meant with respect to Lebesgue measure.)

Let $S = \mathbb{R}$. Fix a density f_0, a sequence (c_n) of bivariate copula densities, and a sequence (r_n) of constants in (0, 1]. For the sake of simplicity, assume $f_0 > 0$ and $c_n > 0$ for all $n \geq 0$. Then, a sequence (β_n) of predictive distributions may be built as follows. Call F_0 the distribution function corresponding to f_0 and define

$$f_1(x \mid y) = (1 - r_0) f_0(x) + r_0 f_0(x) c_0(F_0(x), F_0(y)) \quad \text{for } x, y \in \mathbb{R}.$$

Note that $f_1 : \mathbb{R}^2 \to \mathbb{R}^+$ is a measurable function such that $f_1(\cdot \mid y)$ is a density for each fixed $y \in \mathbb{R}$. Hence, if $F_1(\cdot \mid y)$ denotes the distribution function corresponding to $f_1(\cdot \mid y)$, we can let

$$f_2(x \mid y, z) = (1 - r_1) f_1(x \mid y) + r_1 f_1(x \mid y) c_1(F_1(x \mid y), F_1(z \mid y))$$

for $x, y, z \in \mathbb{R}$. In general, given a measurable function $f_n : \mathbb{R}^{n+1} \to \mathbb{R}^+$ such that $f_n(\cdot \mid y)$ is a density for each fixed $y \in \mathbb{R}^n$, we can define

$$f_{n+1}(x \mid y, z) = (1 - r_n) f_n(x \mid y) + r_n f_n(x \mid y) c_n(F_n(x \mid y), F_n(z \mid y))$$

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

for all $x, z \in \mathbb{R}$ and $y \in \mathbb{R}^n$. In the above formula, $F_n(\cdot \mid y)$ denotes the distribution function corresponding to $f_n(\cdot \mid y)$. Proceeding in this way, we obtain a collection of densities $\{f_n(\cdot \mid y)\colon n \geq 1, y \in \mathbb{R}^n\}$ and the β_n can be defined as $\beta_0(\mathrm{d}x) = f_0(x)\,\mathrm{d}x$ and $\beta_n(\mathrm{d}x) = f_n(x \mid X(n))\,\mathrm{d}x$ for $n \geq 1$.

The predictives β_n were introduced in [19] and then used in [14]. The construction given here is from [2].

If the predictive distributions of X are the β_n, then X is CID. This is proved in [2] when $r_n = 1$, but that proof can be easily extended to any choice of r_n. Since X is CID, there is an RPM β such that $\beta_n \to \beta$ weakly a.s. A natural question is whether $\beta_n \to \beta$ in total variation a.s., or equivalently

$$\beta(\mathrm{d}x) = f(x)\,\mathrm{d}x \quad \text{for some random density } f; \tag{4}$$

see Example 2. Among other things, condition (4) is tacitly assumed in [14, 19]. We now prove that, if r_n and c_n are arbitrary, condition (4) may fail.

Example 4. A necessary condition for (4) is

$$\lim_n r_n \int_0^1 \int_0^1 |c_n(u, v) - 1|\,\mathrm{d}u\,\mathrm{d}v = 0. \tag{5}$$

Hence, trivially, condition (4) fails whenever $\limsup_n r_n > 0$ and $c_n = c$ for all $n \geq 0$, where c is any bivariate copula density such that $\int_0^1 \int_0^1 |c(u, v) - 1|\,\mathrm{d}u\,\mathrm{d}v > 0$.

Let us prove that (4) $\Rightarrow$ (5). Suppose the predictives of X are the β_n, condition (4) holds, and define

$$D_n = \int_{-\infty}^{\infty} \big|f_{n+1}(x \mid X(n+1)) - f_n(x \mid X(n))\big|\,\mathrm{d}x.$$

Since X is CID, condition (4) amounts to $\int_{-\infty}^{\infty} \big|f_n(x \mid X(n)) - f(x)\big|\,\mathrm{d}x \xrightarrow{\text{a.s.}} 0$ for some random density f. Hence, $D_n \xrightarrow{\text{a.s.}} 0$. Since $D_n \leq 2$ a.s., we also obtain $\mathbb{E}(D_n) \to 0$. Moreover,

$$\begin{aligned} D_n &= r_n \int_{-\infty}^{\infty} \big|c_n\big(F_n(x \mid X(n)),\, F_n(X_{n+1} \mid X(n))\big) - 1\big|\, f_n(x \mid X(n))\,\mathrm{d}x \\ &= r_n \int_0^1 \big|c_n\big(u,\, F_n(X_{n+1} \mid X(n))\big) - 1\big|\,\mathrm{d}u. \end{aligned}$$

Conditionally on $X(n)$, the random variable $F_n(X_{n+1} \mid X(n))$ is uniformly distributed on (0,1). Hence, D_n is distributed as $D_n^* = r_n \int_0^1 |c_n(u, V) - 1|\,\mathrm{d}u$, where V is any random variable with uniform distribution on (0,1). Therefore,

$$r_n \int_0^1 \int_0^1 |c_n(u,\ v) - 1|\,\mathrm{d}u\,\mathrm{d}v = \mathbb{E}(D_n^*) = \mathbb{E}(D_n) \longrightarrow 0.$$

Example 4 should be compared with Theorem 1, which states that $\alpha_n \to \alpha$ in total variation a.s., where $\alpha = \mathcal{L}(M, Q)$. Since $\mathcal{L}(0, I)$ admits a density, we obtain $\alpha(\mathrm{d}x) = f(x)\,\mathrm{d}x$ a.s. for some random density f.

Even if not generally true, condition (4) holds under some assumptions.

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

Theorem 3. *Condition (4) holds whenever* $\sum_n r_n < \infty$.

The condition $\sum_n r_n < \infty$ is actually strong. In a sense, when $\sum_n r_n < \infty$, the algorithm receives only finite learning. The next result requires a boundedness condition on c_n but replaces $\sum_n r_n < \infty$ with the weaker condition $\sum_n r_n^2 < \infty$.

Theorem 4. *Condition (4) holds whenever* $\int_{-\infty}^{\infty} f_0(x)^2 \, dx < \infty$, $\sum_n r_n^2 < \infty$, *and there is a constant* b *such that* $c_n(u, v) \le b$ *for all* $n \ge 0$ *and* $u, v \in [0, 1]$.

Bounded copula densities, such as the c_n in Theorem 4, play a role in applications. A meaningful example is Bernstein copulas, which provide a uniform approximation to any copula; see, e.g., [24, 27].

Finally, we note that a reasonable condition on the weights is

$$\sum_n r_n = \infty, \qquad \sum_n r_n^2 < \infty. \tag{6}$$

For instance, condition (6) is standard in stochastic approximation. Theorem 3 violates condition (6). Similarly, Example 4 conflicts with (6) (since $\limsup_n r_n > 0$). On the contrary, Theorem 4 is consistent with (6). An open question is whether $\sum_n r_n^2 < \infty$ suffices for condition (4) in the special case where $c_n = c$ for all n, where c is a fixed copula density. In some cases, the answer is yes; see, e.g., [14, Theorem 5].

3. Numerical illustrations

This section reports the results of an empirical comparison between the α_n and the copula-based predictive distributions β_n. The comparison is based on the speed of convergence and the performance in Bayesian parameter inference via PR.

We let $p = 1$. Moreover, we assume specific forms for α_n and β_n. As they depend on the type of experiment, they will be provided in Sections 3.2 and 3.3. For now, we begin with a brief recap on PR.

3.1. Predictive resampling

The PR method, introduced in [14], allows an inference to be made on a random parameter and/or to predict future observations. In a Bayesian framework, it is an alternative to the usual likelihood/prior scheme, and it has various similarities with the predictive approach to statistics [2]. In particular, the explicit assignment of a prior distribution is not required.

Suppose the object of inference is a function of the sequence X, say $\theta = f(X) = f(X_1, X_2, \dots)$, where f is a known measurable function. As an obvious example, think of θ as the mean of a population described by X, i.e.

$$\theta = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} X_i, \quad \text{where the limit is assumed to exist a.s.} \tag{7}$$

More generally, the informal idea is that the uncertainty about θ would disappear if the whole data sequence X could be observed. The goal is to estimate the distribution of θ based on the available data.

To achieve this target, as in Section 1.2, we assume that an initial dataset $x = (x_1, \dots, x_s)$ is available at time 0. Then, we select a sequence (γ_n) of predictive distributions depending on x (in the interests of readability, we avoid making such dependence explicit in the notation). Finally, through (γ_n) and x, we draw N more observations $(X_1, \dots, X_N)$ and use the

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

Algorithm 1: Predictive resampling.

for $i = 1, \ldots, B$ **do**
  for $j = 1, \ldots, N$ **do**
    Sample $X_j \sim \gamma_{j-1}$
    Update: $\{X_j, \gamma_{j-1}\} \mapsto \gamma_j$
  end
  Evaluate $\theta_i = \hat{\theta}(x, X_1, \ldots, X_N)$
end
Return $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_B)$.

data $(x, X_1, \ldots, X_N)$ to estimate θ. This procedure is repeated a number B of times in order to end up with a sample of estimates of θ. More precisely, let $\hat{\theta} = \hat{\theta}(x, X_1, \ldots, X_N)$ denote the adopted estimate of θ. For instance, if θ is given by (7), it is quite natural to let

$$\hat{\theta} = \frac{\sum_{i=1}^{s} x_i + \sum_{i=1}^{N} X_i}{N + s}.$$

Algorithm 1 provides pseudocode for PR.

The final output of PR is a random sample $(\theta_1, \ldots, \theta_B)$ from the probability distribution of $\hat{\theta}(x, X_1, \ldots, X_N)$. In particular, $\theta_1, \ldots, \theta_B$ are i.i.d. and their common distribution $\Pi_N(\cdot \mid x)$ can be written as

$$\Pi_N(A \mid x) = \int_{\mathbb{R}^N} \mathbf{1}_A\{\hat{\theta}(x, x_1, \ldots, x_N)\} \prod_{j=1}^{N} \gamma_{j-1}(x_1, \ldots, x_{j-1})(\mathrm{d}x_j).$$

$\Pi_N(\cdot \mid x)$ can be regarded as the PR analogue of the usual posterior distribution and it is used to approximate the distribution of θ given x, i.e. $\Pi_\infty(A \mid x) = \mathbb{P}(f \in A \mid x)$, where $\mathbb{P}(\cdot \mid x)$ denotes the probability law induced by the predictives γ_n according to the Ionescu–Tulcea theorem; see Section 1.

So far, (γ_n) is any sequence of predictives. However, for PR to make sense, (γ_n) should satisfy some conditions.

- First of all, γ_n should converge (in some sense) as $n \to \infty$. In addition, γ_0 should fit the observed data x and the sequence (γ_n) should have an updating mechanism that *remembers* x. These are a sort of consistency conditions. Indeed, PR is essentially a tool to make inference on some population which is only partially observed. Thus, a fictitious population (X_n) is created and is exploited alongside x to estimate features of the unknown population. Since x is all we know about the phenomenon/population of interest, the reconstructed population should not add any new information. Rather, it should be a version of x with similar shape and location but more uncertainty. This informal idea is realized precisely by regarding such a population as distributed according to γ, where γ is the limit of the γ_n, which preserves the information provided by x.
- An additional property of (γ_n), which is not mandatory but certainly desirable, is that X is CID under (γ_n). Equivalently, there is an RPM γ on $\mathcal{B}$ such that

$$\mathbb{E}[\gamma(A)] = \gamma_0(A) \quad \text{and} \quad \mathbb{E}[\gamma(A) \mid X(n)] = \gamma_n(A) \quad \text{a.s. for all } A \in \mathcal{B}. \tag{8}$$

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

Under (8), not only does $\gamma_n \to \gamma$ weakly a.s., but γ_n is already the posterior expectation of its limit γ. In a sense, (8) is a coherence condition which provides a meaningful Bayesian interpretation to γ_n. Note that condition (8) holds if $\gamma_n = \beta_n$ (for X is CID under (β_n)) but it fails if $\gamma_n = \alpha_n$.

A final conceptual remark concerns the interpretation of (γ_n).

- Is (γ_n) an actual probabilistic model for the data X, or is it merely a computational device for implementing PR? The goal of PR is generating pseudo-observations that resemble the true underlying population. This cannot be expected if (γ_n) is not a good model for X, or if it does not preserve at least those features of X which are objects of inference. Thus, overall, the role of (γ_n) is the former. Incidentally, this is why γ_0 is required to fit x and (γ_n) to preserve the information provided by the initial sample. Note that, here, we are referring to a generic sequence (γ_n), to be used in PR, while the predictives α_n of this paper have a potential scope much larger than PR; see Remark 1. They can be used in PR, but they are also suitable for modeling the data-generating mechanism in Bayesian predictive inference, as well as in the other frameworks mentioned in Section 1.

3.2. Speed of convergence

In this section, we contrast α_n and β_n on the basis of speed of convergence. Our aim is investigating how fast they converge and how far they get from the predictive at $n = 0$. Intuitively, the distance between α_n and α_0 should inform on how much uncertainty is created by PR, while speed of convergence should provide an idea of how many iterations are needed to exhaust all such uncertainty. The same applies to β_n. For the comparison to be fair, we apply PR without assuming an initial dataset x, so that we can easily set $\alpha_0 = \beta_0$. Indeed, since we want to assess how differently α_n and β_n evolve, it seems reasonable to make them start from the same point. In this way, we should more clearly isolate the impact of their updating mechanisms. Thus, we set $\mathcal{L} = \mathcal{N}$, so that $\alpha_0 = \mathcal{N}(0, 1)$, $\alpha_1 = \mathcal{N}(X_1, 1)$, and $\alpha_n = \mathcal{N}(M_n, Q_n)$ for $n > 1$ (see Remark 2). On the other hand, the β_n are built as in Section 2.3, letting $f_0 =$ standard normal density and $c_n = c_\rho$ for all $n \geq 0$, where c_ρ is the copula density corresponding to a bivariate normal distribution with mean 0, variance 1, and correlation $\rho \in (0, 1)$.

Let f_n^α and f_n^β denote the densities of α_n and β_n (with respect to Lebesgue measure). Following [14], to evaluate speed of convergence we compute (numerically) the L^1-distances $\|f_0 - f_n^\alpha\|_1$ and $\|f_0 - f_n^\beta\|_1$, where f_0 is the standard normal density. Roughly speaking, the informal idea is that, for any sequence (f_n) of random densities, the value of n at which $\|f_0 - f_n\|_1$ plateaus can be regarded as the index after which f_n is close to its limit (provided it exists).

Recall that, thanks to Theorem 1, $\|f_n^\alpha - f\|_1 \xrightarrow{\text{a.s.}} 0$ for some random density f. On the contrary, β_n may not always converge in total variation. Thus, we explore empirically the following scenarios:

(i) $r_n = r_n^{(a)} = \left(2 - \dfrac{1}{n+1}\right)\dfrac{1}{n+2}$ and $\rho = 0.9$.
This choice of r_n, inspired by the Dirichlet process mixture model, is proposed in [14]. However, Theorems 3 and 4 do not apply and the same happens for [14, Theorem 5] since $\rho > 1/\sqrt{3}$. Hence, whether or not β_n converges in total variation is unknown (to our knowledge).

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

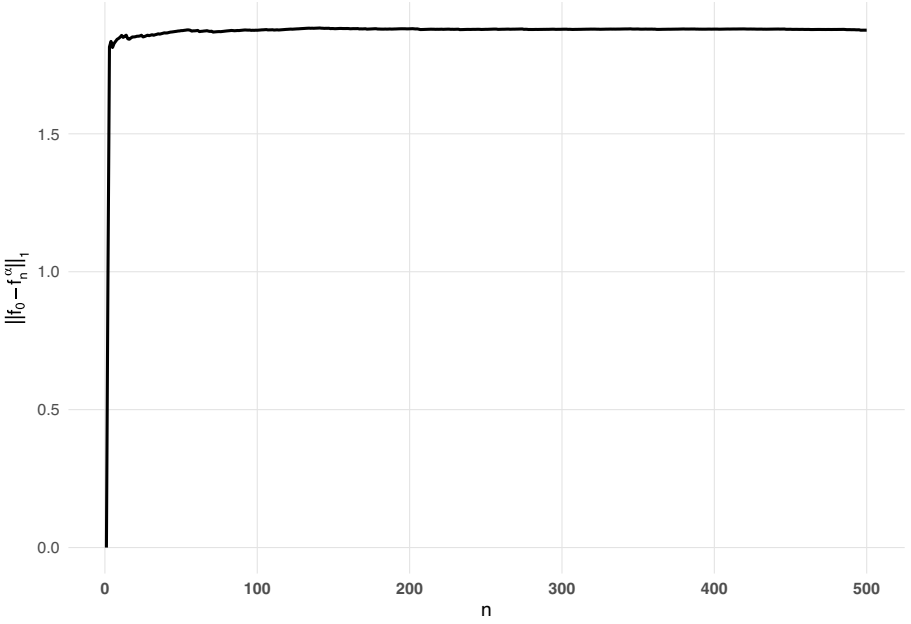


FIGURE 1. $||f_0 - f_n^{\alpha}||_1$.

(ii) $r_n = r_n^{(b)} = \left(2 - \frac{1}{n+2}\right) \frac{1}{(n+3)(\log(n+3))^2}$ and $\rho = 0.9$.
In this case, $\sum_n r_n < \infty$ so that β_n converges in total variation by Theorem 3.

(iii) $r_n = r_n^{(a)}$ and $\rho = 0.4$.
In this case, β_n converges in total variation by [14, Theorem 5].

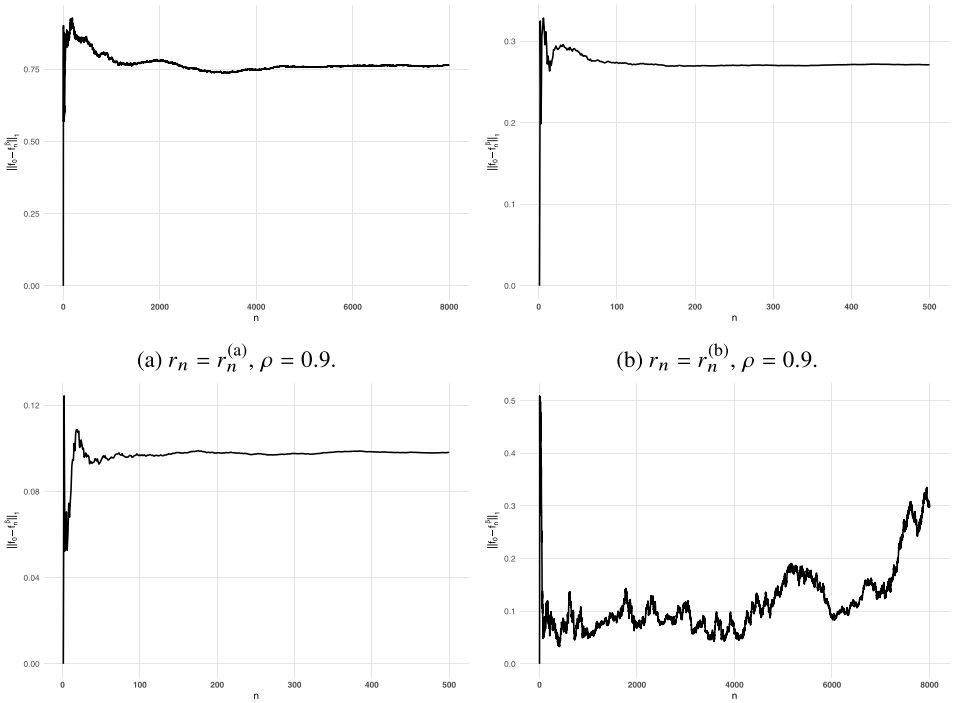
(iv) $r_n = r_n^{(c)} = \left(2 - \frac{1}{n+1}\right) \frac{1}{n+2} + 0.01$ and $\rho = 0.4$.
These r_n are similar to the $r_n^{(a)}$, where 0.01 is added just to verify the conditions of Example 4, which guarantee that β_n does not converge in total variation.

Figure 1 shows the path of $\|f_0 - f_n^{\alpha}\|_1$. It takes about 150 iterations of PR for the L^1-distance to stabilize.

Figures 2(a)– 2(d) show the path of $\|f_0 - f_n^{\beta}\|_1$ for scenarios (i)–(iv) respectively. In the first case, it takes about 4000 iterations of PR for the L^1-distance to stabilize. Thus, either β_n does not converge in total variation or it does so more slowly than in scenarios (ii) and (iii). Indeed, in Figures 2(b) and 2(c) the L_1-distance stabilizes quickly, which is consistent with the results that guarantee convergence in total variation in those cases (respectively, Theorem 3 and [14, Theorem 5]). Figure 2(d) suggests lack of convergence in total variation, which is in line with Example 4.

A couple of remarks are in order. Firstly, we have seen that the choice of r_n and ρ in the β_n significantly affects the reconstruction of uncertainty. This highlights what stated at the end of the previous section, i.e. that when making inference via PR we should also care about the updating mechanism of the predictives. For the β_n with $c_n = c_\rho$, some guidelines are provided

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press



(a) $r_n = r_n^{(a)}$, $\rho = 0.9$.

(b) $r_n = r_n^{(b)}$, $\rho = 0.9$.

(c) $r_n = r_n^{(a)}$, $\rho = 0.4$.

(d) $r_n = r_n^{(c)}$, $\rho = 0.4$.

FIGURE 2. $||f_0 - f_n^{\beta}||_1$.

in [14] for choosing r_n (as mentioned in this section) and ρ. Thus, we follow them in the next section.

Secondly, Figures 2(a)– 2(c) show that f_n^{α} moves further from f_0 than f_n^{β} does. This hints that α_n might reconstruct more uncertainty than β_n, which would be consistent with the posterior variance estimates obtained in the next section.

3.3. Parameter estimation

We now compare α_n and β_n when used to estimate the posterior distribution $\Pi_N(\cdot \mid x)$ of some parameter θ via PR; see Section 3.1. We assume an initial dataset $x = (x_1, \ldots, x_s)$ is available and we build the predictives to account for it. Indeed, since we are performing PR, α_0 should fit x. Similarly, the updating mechanism of the sequence (α_n) should preserve the information provided by x. The same remarks apply to (β_n).

The α_n are specified as follows. We partition x with a model-based clustering technique, available in the `mclust` R package [16–18], which finds the mixture of Gaussians that best fits x, accounting also for model complexity, and the partition of x which most likely was induced by such mixture. Thus, $x = (x_{1,1}, \ldots, x_{1,n_1}, \ldots, x_{k,1}, \ldots, x_{k,n_k})$, where k is the number of clusters, $x_{i,j}$ is the jth observation of cluster i, n_i is the size of cluster i, and $\sum_{i=1}^{k} n_i = s$. Then, for $i = 1, \ldots, k$, we let

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

$$a_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i,j}, \qquad b_i = \frac{1}{n_i}\sum_{j=1}^{n_i}(x_{i,j}-a_i)^2, \qquad w_i = \frac{n_i}{s},$$

and Y be a random variable with distribution $m = \sum_{i=1}^{k} w_i \mathcal{N}(a_i, b_i)$. Finally, the kernel $\mathcal{L}$ is chosen as the law of the random variable $(Y - \mathbb{E}(Y))/\sqrt{\mathrm{Var}(Y)}$ and the α_n are defined by (3).

Similarly to α_0, β_0 is constructed as an estimate of the density of the distribution that $x_1, \ldots, x_s$ are drawn from (x is assumed to be an i.i.d. sample; see below). Following [14], this is done via prequential log-likelihood optimization. In particular, let $f^{\beta}_{n,\rho}$ be the density of β_n for $\beta_0 = \mathcal{N}(0, 1)$, $c_n = c_\rho$, and $r_n = r_n^{(\mathrm{a})}$. Then, we evaluate $f^{\beta}_{i-1,\rho}(x_i \mid x_1, \ldots, x_{i-1})$ for $i = 1, \ldots, s$ and $\rho \in \{0.05, 0.1, \ldots, 0.95\}$, and let

$$\hat{\rho} = \arg\max_{\rho} \sum_{i=1}^{s} \log f^{\beta}_{i-1,\rho}(x_i \mid x_1, \ldots, x_{i-1}).$$

Finally, we set $\beta_0(\mathrm{d}y) = f^{\beta}_{s,\hat{\rho}}(y \mid x)\,\mathrm{d}y$. The rest of the sequence (β_n) follows from the construction given in Section 2.3, upon setting $c_n = c_{\hat{\rho}}$ and $r_n = r_n^{(\mathrm{a})}$.

The parameter θ is taken as the mean, variance, skewness, kurtosis, and median of a reference population. In what follows, $\hat{\theta}$ is short for $\hat{\theta}(x, X_1, \ldots, X_N)$, which represents the sample equivalent of the corresponding population quantity (e.g., if θ is the population mean, $\hat{\theta}$ is the sample mean). Moreover, when necessary we use the notations $\Pi^{\alpha}_N(\cdot \mid x)$ and $\Pi^{\beta}_N(\cdot \mid x)$ to stress that $\Pi_N(\cdot \mid x)$ is obtained using the α_n or the β_n.

The simulation study is organized as follows. We draw random samples of size $s = 50, 100, 500, 1000, 1500, 2000$ from a mixture of three shifted t-distributions with weights $w_1 = 0.3$, $w_2 = 0.1$, and $w_3 = 0.6$, shifting constants $\mu_1 = -5$, $\mu_2 = 0$, and $\mu_3 = 4$, and degrees of freedom $\nu = 4.5$. Then, we run PR with $N = 5000$ and $B = 1000$. $B = 1000$ allows us to visualize well posterior density and does not make PR too slow. Similarly, $N = 5000$ allows a satisfying compromise between computational speed and reconstruction of uncertainty. Once a sample $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_B)$ from $\Pi_N(\cdot \mid x)$ is obtained, we compute the root mean square error (RMSE) between $\boldsymbol{\theta}$ and the true parameter value (say θ_0) and the sample variance of $(\theta_1, \ldots, \theta_B)$ (which is an estimate of posterior variance). These indicators quantify how close $\Pi_N(\cdot \mid x)$ is to θ_0 and how much posterior uncertainty is generated by PR. We also keep track of computation times.

Tables 1 and 2 show the RMSE respectively for α_n and β_n, while Tables 3 and 4 show the posterior variance. Overall, estimates appear to be more precise and less dispersed as s increases. This seems reasonable, as a larger sample size typically implies more information and less uncertainty. Averaging the RMSE values in Tables 1 and 2 over all sample sizes and parameters yields a mean RMSE of 0.818 and 2.731 respectively. One might think this discrepancy is due to $\Pi^{\beta}_N(\cdot \mid x)$ being more dispersed than $\Pi^{\alpha}_N(\cdot \mid x)$, but Tables 3 and 4 provide opposing evidence. Indeed, averaging the values of those tables, we obtain a mean posterior variance of 0.206 for α_n and 0.019 for β_n. Therefore, such a difference might be due to $\Pi^{\alpha}_N(\cdot \mid x)$ being located on θ_0 more often than $\Pi^{\beta}_N(\cdot \mid x)$. Evidence in favour of this can be found by looking at 95% credible intervals, estimated by taking sample quantiles of $(\theta_1, \ldots, \theta_B)$. Indeed, β_n tends to give rise to shorter intervals which contain θ_0 less frequently. In particular, out of 30

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

TABLE 1. RMSE α_n.

θ	$s=50$	$s=100$	$s=500$	$s=1000$	$s=1500$	$s=2000$
Mean	2.055	0.381	0.341	0.063	0.009	0.011
Variance	7.936	3.149	1.662	0.646	0.230	0.103
Skewness	0.231	0.018	0.061	0.009	0.001	0.001
Kurtosis	0.439	0.178	0.245	0.002	0.017	0.037
Median	5.983	0.382	0.270	0.022	0.014	0.055

TABLE 2. RMSE β_n.

θ	$s=50$	$s=100$	$s=500$	$s=1000$	$s=1500$	$s=2000$
Mean	0.119	0.039	0.112	0.047	0.002	0.009
Variance	22.117	25.452	3.125	0.108	1.308	0.856
Skewness	0.141	0.210	0.275	0.173	0.155	0.095
Kurtosis	0.111	0.394	1.074	1.386	0.781	1.782
Median	5.255	3.240	5.216	2.488	2.912	2.950

TABLE 3. Posterior variance α_n.

θ	$s=50$	$s=100$	$s=500$	$s=1000$	$s=1500$	$s=2000$
Mean	0.4118	0.1744	0.0357	0.0154	0.0089	0.0067
Variance	2.5623	1.5149	0.3038	0.2419	0.1328	0.0913
Skewness	0.0006	0.0006	0.0006	0.0005	0.0004	0.0004
Kurtosis	0.0001	0.0008	0.0005	0.0015	0.0009	0.0009
Median	0.4591	0.1558	0.0495	0.0107	0.0077	0.0041

TABLE 4. Posterior variance β_n.

θ	$s=50$	$s=100$	$s=500$	$s=1000$	$s=1500$	$s=2000$
Mean	0.003	0.003	0.004	0.003	0.002	0.002
Variance	0.048	0.053	0.149	0.105	0.080	0.084
Skewness	0.0004	0.0004	0.001	0.001	0.001	0.001
Kurtosis	0.001	0.001	0.003	0.003	0.002	0.002
Median	0.005	0.005	0.006	0.005	0.004	0.006

different combinations of θ and s, the 95% credible intervals produced by α_n and β_n contain θ_0 respectively 15 and 3 times (see also Appendix B).

There might be a couple of ways to improve the performance of β_n. Firstly, we could take a finer grid of candidate values of ρ for the prequential log-likelihood optimisation. In this way, β_0 should fit better the observed data and thus make estimates more precise. Secondly, we

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

TABLE 5. Computation times in minutes.

Predictive	$s = 50$	$s = 100$	$s = 500$	$s = 1000$	$s = 1500$	$s = 2000$
α_n	0.012	0.014	0.017	0.021	0.022	0.026
β_n	5.10	5.87	7.12	8.26	8.28	9.98

could make the sampling mechanism more accurate. Indeed, $X_1, \ldots, X_N$ are generated via inverse sampling from a grid $\{y_1, \ldots, y_G\}$. Thus, once again, better posterior estimates could be obtained by refining such a grid. It is worth stressing that, if one wants only to update β_n into β_{n+1}, it is not necessary to sample $X_{n+1} \sim \beta_n$ and plug it in to β_{n+1}. As explained in [14, Section 4.2], it is enough to draw an observation from a uniform distribution on (0, 1). But clearly, we cannot use this shortcut for we need to know the actual value of the sampled variables. Finally, we note that these ameliorations come at the cost of longer computation times. We explore this further in the next paragraph.

Table 5 shows the time needed to perform PR for α_n, β_n, and each value of s. The PR algorithm is coded in `Rcpp`. Computations are carried out on a single-core version of R installed on a machine with an AMD Epyc 7763 2.45 Ghz processor with 2 TB RAM. The difference between the computation times of α_n and β_n, with the former being much faster than the latter, is quite evident. Reasons for this discrepancy have partially been pointed out already, so we summarize them and provide some more details. Sampling from α_n requires just computing a mean, a variance, drawing one of the components of the Gaussian mixture (with probability equal to its weight), and then sampling from it. This can be done efficiently in any statistical software. On the other hand, sampling from β_n entails estimating the hyperparameter ρ (at the beginning of PR) and doing inverse sampling. The time needed for these tasks depends on the size of the grid of values taken to perform them. Such a grid cannot be too coarse, otherwise the accuracy of the estimates is worsened. Therefore, these two steps are likely to be the main reason why α_n is faster than β_n. Furthermore, computational times for β_n increase significantly with sample size. This might be due to the estimation of ρ being affected by s. Indeed, as explained at the beginning of this section, optimizing the prequential log-likelihood requires a recursive evaluation of the predictive density from step 0 to s, for each candidate value of ρ. Clearly, this procedure is slower the larger the value of s. Even the performance of α_n is affected by s, which is likely due to model-based clustering being slower the larger the sample to partition. Yet, the impact is way less marked.

4. Discussion

Let $X = (X_1, X_2, \ldots)$ be a sequence of p-dimensional random vectors, and

$$\alpha_n(\cdot) = \mathbb{P}\big(X_{n+1} \in \cdot \mid X_1, \ldots, X_n\big)$$

the corresponding predictive distributions. In this paper, a new sequence (α_n) of predictives is introduced, namely $\alpha_n = \mathcal{L}(M_n, Q_n)$, where $\mathcal{L}$ is a suitable kernel, and M_n and Q_n are the sample mean and covariance matrix of $(X_1, \ldots, X_n)$. Convergence in total variation of α_n is proved, along with an explicit formula for the limit RPM. In the (meaningful) special case where $\mathcal{L} = \mathcal{N}$, the convergence rate is shown to be $n^{-1/2}$. A natural competitor of α_n is the copula-based predictive distribution β_n of [19]. Thus, β_n is investigated as well. It turns out that β_n converges weakly but not necessarily in total variation. Hence, sufficient conditions for

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

β_n to converge in total variation are provided. Eventually, an empirical comparison between α_n and β_n is performed, assessing their speed of convergence and their ability to estimate, via PR, the posterior distribution of population moments and quantiles.

When making parameter inference, $\mathcal{L}$ is chosen as a mixture of Gaussians retrieved via model-based clustering on the initial dataset x. This flexible structure allows us to capture well the shape and location of the underlying population, as demonstrated by the satisfying performance of α_n in PR.

We finally mention two issues for future research. The first is to give conditions on $\mathcal{L}$ under which the convergence rate of α_n is still $n^{-1/2}$ (as happens when $\mathcal{L} = \mathcal{N}$). The second is using the α_n in regression problems (as done in [14] with the β_n). In this case, the data are of the type $X_n = (Y_n, Z_n)$, where Y_n is a response variable and Z_n a vector of covariates. Thus, one could generate several multivariate populations via PR and then use standard frequentist estimators to obtain a posterior sample of regression functions.

Appendix A. Proofs

In this appendix, we prove the results of Section 2. To this end, we first recall two known facts and we make a preliminary remark.

Lemma 1. *If $x, y \in \mathbb{R}^p$ and B is a real non-singular matrix of order $p \times p$, then*

$$\det\big(B + xy^\top\big) = \det(B)\big(1 + x^\top b^{-1} y\big).$$

Lemma 2. *If $(V_n : n \geq 1)$ is an i.i.d. sequence of real random variables such that $\mathbb{E}(V_1) = 0$ and $\mathbb{E}\big(|V_1|^{1+\varepsilon}\big) < \infty$ for some $\varepsilon > 0$, then the series $\sum_{n=1}^{\infty} V_n/n$ converges a.s.*

Proof. Since $\sum_n \mathbb{P}(|V_n| > n) = \sum_n \mathbb{P}(|V_1| > n) \leq \mathbb{E}\big(|V_1|\big) < \infty$, the Borel–Cantelli lemma yields $\mathbb{P}(|V_n| > n \text{ for infinitely many } n) = 0$. Hence, it suffices to show that $\sum_n W_n/n$ converges a.s., where $W_n = V_n \mathbf{1}_{\{|V_n| \leq n\}}$. To this end, first note that $(W_n - \mathbb{E}(W_n))/n$ are independent and centered, and

$$\sum_n \mathbb{E}\left\{\left(\frac{W_n - \mathbb{E}(W_n)}{n}\right)^2\right\} \leq \sum_n \mathbb{E}\left\{\frac{V_1^2 \mathbf{1}_{\{|V_1| \leq n\}}}{n^2}\right\} \leq \sum_n \frac{\mathbb{E}\big(|V_1|^{1+\varepsilon}\big)}{n^{1+\varepsilon}} < \infty.$$

Hence, the series $\sum_n (W_n - \mathbb{E}(W_n))/n$ converges a.s. Finally, $\mathbb{E}(V_1) = 0$ implies

$$\mathbb{E}(W_n) = \mathbb{E}\big\{V_1 \mathbf{1}_{\{|V_1| \leq n\}}\big\} = -\mathbb{E}\big\{V_1 \mathbf{1}_{\{|V_1| > n\}}\big\}.$$

Therefore,

$$\sum_n \frac{|\mathbb{E}(W_n)|}{n} \leq \sum_n \frac{\mathbb{E}\big\{|V_1| \mathbf{1}_{\{|V_1| > n\}}\big\}}{n} \leq \sum_n \frac{\mathbb{E}\big(|V_1|^{1+\varepsilon}\big)}{n^{1+\varepsilon}} < \infty.$$

This concludes the proof. □

Both the previous lemmas are well known. A possible reference for Lemma 1 is [11], while Lemma 2 has been proved here as we know of no explicit reference. Incidentally, in Lemma 2, the condition $\mathbb{E}\big(|V_1|^{1+\varepsilon}\big) < \infty$ can possibly be weakened but not completely removed. In fact, as shown in the next example, the series $\sum_{n=1}^{\infty} V_n/n$ may fail to converge a.s. even if (V_n) is i.i.d. and $\mathbb{E}(V_1) = 0$.

Example 5. Let (V_n) be i.i.d. with $\mathbb{P}(V_1 = -c/\log 2) = \frac{1}{2}$ and

$$\mathbb{P}(2 < V_1 \leq x) = \frac{c}{2} \int_2^x \frac{\mathrm{d}t}{(t \log t)^2} \quad \text{for } x > 2,$$

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

where the constant c is given by $c = \left(\int_2^\infty \mathrm{d}t/(t\log t)^2\right)^{-1}$. Then,

$$\mathbb{E}(V_1) = -\left\{\frac{c}{2\log 2} + \frac{c}{2}\right\}\int_2^\infty \frac{t}{(t\log t)^2}\,\mathrm{d}t = 0.$$

Letting $W_n = (V_n/n)\mathbf{1}_{\{|V_n|\le n\}}$, a direct calculation shows that we have $\sum_n \mathbb{E}(W_n) = -\infty$ and $\sum_n \mathrm{Var}\,(W_n) < \infty$. Hence, $\sum_{n=1}^\infty W_n \overset{\text{a.s.}}{=} -\infty$. In turn, by the Borel–Cantelli lemma, this implies that $\sum_{n=1}^\infty V_n/n \overset{\text{a.s.}}{=} -\infty$.

Remark 3. In order to prove Theorems 1 and 2, without loss of generality, some simplifications are admissible. Firstly, the α_n can be assumed to be as in Remark 2, i.e.

$$\begin{aligned}&\alpha_0 = \mathcal{L}(0, I), \qquad \alpha_1 = \mathcal{L}(X_1, I), \qquad \alpha_n = \mathcal{L}(M_n, Q_n) \quad \text{for } n \ge 2,\\ &\text{where} \quad M_n = \frac{\sum_{i=1}^n X_i}{n}, \qquad Q_n = \frac{\sum_{i=1}^n (X_i - M_n)(X_i - M_n)^\top}{n}.\end{aligned} \tag{9}$$

Secondly, the X_n can be defined as follows. Fix an i.i.d. sequence $Z_1, Z_2, \ldots$ of p-dimensional random vectors such that $Z_1 \sim \mathcal{L}(0, I)$. Define a sequence $Y = (Y_1, Y_2, \ldots)$ as

$$\begin{aligned}&Y_1 = Z_1, \qquad Y_2 = Y_1 + Z_2, \qquad Y_{n+1} = N_n + R_n^{1/2} Z_{n+1} \quad \text{for } n \ge 2,\\ &\text{where} \quad N_n = \frac{1}{n}\sum_{i=1}^n Y_i, \qquad R_n = \frac{1}{n}\sum_{i=1}^n (Y_i - N_n)(Y_i - N_n)^\top.\end{aligned}$$

It is straightforward to see that $\mathbb{P}(Y_1 \in \cdot) = \mathcal{L}(0, I)$, $\mathbb{P}(Y_2 \in \cdot \mid Y_1) = \mathcal{L}(Y_1, I)$, and that $\mathbb{P}(Y_{n+1} \in \cdot \mid Y_1, \ldots, Y_n) = \mathcal{L}(N_n, R_n)$ for $n \ge 2$. Hence, $Y \sim X$ whenever the predictive distributions of X are given by (9). In particular, since $Y \sim X$, in the proofs of Theorems 1 and 2 we can suppose that $X_1 = Z_1$, $X_2 = X_1 + Z_2$, and $X_{n+1} = M_n + Q_n^{1/2} Z_{n+1}$ for $n \ge 2$, where $Z_1, Z_2, \ldots$ are i.i.d. and $Z_1 \sim \mathcal{L}(0, I)$.

We are now able to attack the results of Section 2.2. To this end, we let $\mathcal{F}_n = \sigma(Z_1, \ldots, Z_n)$, with $\mathcal{F}_0$ the trivial σ-field.

Proof of Theorem 1. Suppose $M_n \xrightarrow{\text{a.s.}} M$, $Q_n \xrightarrow{\text{a.s.}} Q$, and $Q \in \mathcal{M}$ a.s. Let ϕ be a density of $\mathcal{L}(0, I)$ with respect to p-dimensional Lebesgue measure. For each $n \ge 1$, under the predictives $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$, the probability distribution of $(X_1, \ldots, X_n)$ is absolutely continuous with respect to np-dimensional Lebesgue measure. Based on this fact, by standard arguments, it can be easily seen that $Q_n \in \mathcal{M}$ a.s. Therefore,

$$\|\alpha_n - \alpha\| = \|\mathcal{L}(M_n, Q_n) - \mathcal{L}(M, Q)\| \overset{\text{a.s.}}{=} \frac{1}{2}\int_{\mathbb{R}^p} \left| \frac{\phi\big(Q_n^{-1/2}(x - M_n)\big)}{\det(Q_n)^{1/2}} - \frac{\phi\big(Q^{-1/2}(x - M)\big)}{\det(Q)^{1/2}} \right| \mathrm{d}x.$$

Given $\varepsilon > 0$, take an integrable continuous function f on $\mathbb{R}^p$ such that $\int_{\mathbb{R}^p} |\phi(x) - f(x)|\,\mathrm{d}x < \varepsilon$. Define

$$I_n = \int_{\mathbb{R}^p} \left| \frac{f\big(Q_n^{-1/2}(x - M_n)\big)}{\det(Q_n)^{1/2}} - \frac{f\big(Q^{-1/2}(x - M)\big)}{\det(Q)^{1/2}} \right| \mathrm{d}x.$$

Since f is continuous, $I_n \xrightarrow{\text{a.s.}} 0$. Therefore, $\limsup_n \|\alpha_n - \alpha\| \le \varepsilon + \frac{1}{2}\limsup_n I_n \overset{a.s.}{=} \varepsilon$. This proves that $\|\alpha_n - \alpha\| \xrightarrow{\text{a.s.}} 0$ provided $M_n \xrightarrow{\text{a.s.}} M$, $Q_n \xrightarrow{\text{a.s.}} Q$, and $Q \in \mathcal{M}$ a.s.

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

After noting this fact, the rest of the proof is split into three steps. Recall that, by Remark 3, we can assume $X_1 = Z_1$, $X_2 = X_1 + Z_2$, and $X_{n+1} = M_n + Q_n^{1/2} Z_{n+1}$ for $n \geq 2$, where $Z_1, Z_2, \ldots$ are i.i.d. and $Z_1 \sim \mathcal{L}(0, I)$.

Step 1: Q_n converges a.s. Define $M_0 = 0$, $Q_0 = Q_1 = I$, and $L_n = Q_n^{1/2} Z_{n+1}$. Then

$$M_{n+1} - M_n = \frac{Q_n^{1/2} Z_{n+1}}{n+1} = \frac{L_n}{n+1}.$$

After some algebra, this implies that

$$Q_{n+1} = \frac{n}{n+1} Q_n + \frac{n}{(n+1)^2} L_n L_n^\top, \tag{10}$$

or, equivalently,

$$Q_{n+1}^{(i,j)} = \frac{n}{n+1} Q_n^{(i,j)} + \frac{n}{(n+1)^2} L_n^{(i)} L_n^{(j)} \quad \text{for all } i, j = 1, \ldots, p.$$

The conditional distribution of L_n given $\mathcal{F}_n$ is $\mathcal{L}(0, Q_n)$. Therefore, $\mathbb{E}\{L_n^{(i)} L_n^{(j)} \mid \mathcal{F}_n\} = Q_n^{(i,j)}$ and

$$\mathbb{E}\{Q_{n+1}^{(i,j)} \mid \mathcal{F}_n\} = \frac{n}{n+1} Q_n^{(i,j)} + \frac{n}{(n+1)^2} Q_n^{(i,j)} = Q_n^{(i,j)} \left(1 - \frac{1}{(n+1)^2}\right).$$

For $i = j$, it follows that $(Q_n^{(i,i)} : n \geq 2)$ is a non-negative supermartingale. Hence,

$$\mathbb{E}\{|Q_n^{(i,j)}|\} \leq \frac{\mathbb{E}\{Q_n^{(i,i)}\} + \mathbb{E}\{Q_n^{(j,j)}\}}{2} \leq \frac{\mathbb{E}\{Q_2^{(i,i)}\} + \mathbb{E}\{Q_2^{(j,j)}\}}{2}$$

and

$$\begin{aligned} \sum_n \mathbb{E}\left|\mathbb{E}\{Q_{n+1}^{(i,j)} \mid \mathcal{F}_n\} - Q_n^{(i,j)}\right| &= \sum_n \frac{\mathbb{E}\{|Q_n^{(i,j)}|\}}{(n+1)^2} \\ &\leq \frac{\mathbb{E}\{Q_2^{(i,i)}\} + \mathbb{E}\{Q_2^{(j,j)}\}}{2} \sum_n \frac{1}{(n+1)^2} < \infty. \end{aligned}$$

To sum up, for fixed i and j, the sequence $(Q_n^{(i,j)} : n \geq 2)$ is a quasi-martingale such that $\sup_n \mathbb{E}\{|Q_n^{(i,j)}|\} < \infty$, and this implies that $Q_n^{(i,j)} \xrightarrow{\text{a.s.}} Q^{(i,j)}$ for some random variable $Q^{(i,j)}$. Hence, $Q_n \xrightarrow{\text{a.s.}} Q$, where Q is the matrix $Q = \left(Q^{(i,j)} : 1 \leq i, j \leq p\right)$.

Step 2: M_n is a martingale and it is bounded in L_2. Just note that

$$\mathbb{E}(M_{n+1} - M_n \mid \mathcal{F}_n) = \mathbb{E}\left\{\frac{Q_n^{1/2} Z_{n+1}}{n+1} \mid \mathcal{F}_n\right\} = \frac{Q_n^{1/2} \mathbb{E}(Z_{n+1} \mid \mathcal{F}_n)}{n+1} = \frac{Q_n^{1/2} \mathbb{E}(Z_{n+1})}{n+1} = 0.$$

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

Hence, M_n is a martingale. In addition,

$$\begin{aligned}\mathbb{E}(M_n^\top M_n) &= \mathbb{E}\left\{\left(\sum_{i=1}^n \frac{Q_{i-1}^{1/2} Z_i}{i}\right)^\top \left(\sum_{j=1}^n \frac{Q_{j-1}^{1/2} Z_j}{j}\right)\right\} \\ &= \sum_{i=1}^n \frac{\mathbb{E}\big(Z_i^\top Q_{i-1} Z_i\big)}{i^2} \\ &= \sum_{i=1}^n \sum_{r,k=1}^p \frac{\mathbb{E}\big(Z_i^{(r)} Z_i^{(k)}\big)\mathbb{E}\big(Q_{i-1}^{(r,k)}\big)}{i^2} \\ &= \sum_{i=1}^n \sum_{k=1}^p \frac{\mathbb{E}\big(Q_{i-1}^{(k,k)}\big)}{i^2} \\ &\le \sum_{i=1}^\infty \sum_{k=1}^p \frac{\mathbb{E}\big(Q_{i-1}^{(k,k)}\big)}{i^2} \le p + \frac{p}{4} + \sum_{k=1}^p \mathbb{E}\big(Q_2^{(k,k)}\big) \sum_{i=3}^\infty \frac{1}{i^2},\end{aligned}$$

where the last inequality is because $(Q_n^{(k,k)} : n \ge 2)$ is a supermartingale for fixed $k = 1, \ldots, p$. Therefore, M_n is a martingale such that $\sup_n \mathbb{E}(M_n^\top M_n) < \infty$, and this implies that $M_n \xrightarrow{\text{a.s.}} M$ for some p-dimensional random vector M.

Step 3: $Q \in \mathcal{M}$ a.s. Since $Q \overset{\text{a.s.}}{=} \lim_n Q_n$, the matrix Q is symmetric and non-negative definite a.s. Hence, to prove $Q \in \mathcal{M}$ a.s., it suffices to show that $\det(Q) > 0$ a.s.

Exploiting (10) and Lemma 1, we obtain

$$\begin{aligned}\det(Q_{n+1}) &= \det\left(\frac{n}{n+1} Q_n^{1/2}\left(I + \frac{1}{(n+1)} Z_{n+1} Z_{n+1}^\top\right) Q_n^{1/2}\right) \\ &= \det(Q_n)\left(1 - \frac{1}{n+1}\right)^p \left(1 + \frac{1}{n+1} Z_{n+1}^\top Z_{n+1}\right).\end{aligned}$$

Letting $V_n = Z_n^\top Z_n - p$, the above equation can be written as

$$\begin{aligned}\det(Q_{n+1}) &= \det(Q_n) \sum_{j=0}^p \binom{p}{j} \left(\frac{-1}{n+1}\right)^j \left(1 + \frac{V_{n+1} + p}{n+1}\right) \\ &= \det(Q_n)\left\{1 + \frac{V_{n+1}}{n+1} - p\frac{V_{n+1} + p}{(n+1)^2} + R_{n+1}\right\},\end{aligned}$$

where

$$R_{n+1} = \sum_{j=2}^p \binom{p}{j} \left(\frac{-1}{n+1}\right)^j \left(1 + \frac{V_{n+1} + p}{n+1}\right).$$

Therefore,

$$\det(Q_n) = \det(Q_2) \prod_{j=3}^n \frac{\det(Q_j)}{\det(Q_{j-1})} = \det(Q_2) \prod_{j=3}^n \left\{1 + \frac{V_j}{j} - p\frac{V_j + p}{j^2} + R_j\right\}.$$

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

Next, define

$$U_j = 1 + \frac{V_j}{j} - p\frac{V_j + p}{j^2} + R_j.$$

If $\sum_{j=3}^{n} \log U_j \xrightarrow{\text{a.s.}} T$ as $n \to \infty$ for some real random variable T, then

$$\det(Q) = \lim_n \det(Q_n) = \det(Q_2) \lim_n \exp\left(\sum_{j=3}^{n} \log U_j\right) = \det(Q_2)\mathrm{e}^T > 0 \quad \text{a.s.}$$

Hence, it suffices to show that the series $\sum_j \log U_j$ converges a.s. In turn, this is true since each series involved in the definition of U_j converges a.s. Precisely, $\sum_j V_j/j$ converges a.s. because of Lemma 2. The series $\sum_j (V_j + p)/j^2$ converges a.s. since $V_j + p = Z_j^\top Z_j \geq 0$ and

$$\mathbb{E}\left\{\sum_j \frac{V_j + p}{j^2}\right\} = \sum_j \frac{\mathbb{E}(Z_j^\top Z_j)}{j^2} = \sum_j \frac{p}{j^2} < \infty.$$

Finally, $\sum_j R_j$ converges a.s. by exactly the same argument as used for $\sum_j (V_j + p)/j^2$. Therefore, $\sum_j \log U_j$ converges a.s., and this concludes the proof. □

Proof of Theorem 2. Let $\|\cdot\|_{\mathrm{E}}$ be the Euclidean norm on $\mathbb{R}^p$, and $\|\cdot\|_{\mathrm{F}}$ the Frobenius norm on the matrices of order $p \times p$. The latter is defined as $\|B\|_{\mathrm{F}} = \sqrt{\sum_{i,j} (B^{(i,j)})^2} = \sqrt{\operatorname{trace}(B^\top B)}$ for any $p \times p$ matrix B and has the nice property that $\|B_1 B_2\|_{\mathrm{F}} \leq \|B_1\|_{\mathrm{F}} \|B_2\|_{\mathrm{F}}$.

The total variation distance between two p-dimensional normal laws, say $\mathcal{N}(a, \Sigma_1)$ and $\mathcal{N}(b, \Sigma_2)$ with $a, b \in \mathbb{R}^p$ and $\Sigma_1, \Sigma_2 \in \mathcal{M}$, can be estimated as

$$\|\mathcal{N}(a, \Sigma_1) - \mathcal{N}(b, \Sigma_2)\| \leq \|\mathcal{N}(a, \Sigma_1) - \mathcal{N}(b, \Sigma_1)\| + \tfrac{3}{2}\left\|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - I\right\|_{\mathrm{F}};$$

see, e.g., [22, Theorem 2] and [10, (2)]. In addition,

$$\left\|\Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} - I\right\|_{\mathrm{F}} = \left\|\Sigma_1^{-1/2}(\Sigma_2 - \Sigma_1)\Sigma_1^{-1/2}\right\|_{\mathrm{F}} \leq \left\|\Sigma_1^{-1/2}\right\|_{\mathrm{F}}^2 \|\Sigma_2 - \Sigma_1\|_{\mathrm{F}}$$

and $\|\mathcal{N}(a, \Sigma_1) - \mathcal{N}(b, \Sigma_1)\| \leq \|b - a\|_{\mathrm{E}} \left\|\Sigma_1^{-1/2}\right\|_{\mathrm{F}}$; see [10, p. 5]. Collecting all these facts together, $\|\mathcal{N}(a, \Sigma_1) - \mathcal{N}(b, \Sigma_2)\| \leq C(\Sigma_1)\{\|b - a\|_{\mathrm{E}} + \|\Sigma_2 - \Sigma_1\|_{\mathrm{F}}\}$, where $C(\Sigma_1)$ is a constant that depends on Σ_1 only.

Applying the previous inequality to our framework, we obtain

$$\|\alpha_n - \alpha\| = \|\mathcal{N}(M_n, Q_n) - \mathcal{N}(M, Q)\| \leq C(Q)\{\|M_n - M\|_{\mathrm{E}} + \|Q_n - Q\|_{\mathrm{F}}\}.$$

Hence, it suffices to show that $b := \sup_n \sqrt{n}\mathbb{E}\{\|M_n - M\|_{\mathrm{E}} + \|Q_n - Q\|_{\mathrm{F}}\} < \infty$. In fact, if $b < \infty$ and d_n is any sequence of constants such that $d_n/\sqrt{n} \to 0$, then

$$d_n\mathbb{E}\{\|M_n - M\|_{\mathrm{E}} + \|Q_n - Q\|_{\mathrm{F}}\} \leq \frac{bd_n}{\sqrt{n}} \longrightarrow 0.$$

In particular, $d_n\{\|M_n - M\|_{\mathrm{E}} + \|Q_n - Q\|_{\mathrm{F}}\} \xrightarrow{\mathbb{P}} 0$, which in turn implies that

$$d_n\|\alpha_n - \alpha\| \leq d_n C(Q)\{\|M_n - M\|_{\mathrm{E}} + \|Q_n - Q\|_{\mathrm{F}}\} \xrightarrow{\mathbb{P}} 0.$$

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

To prove $b < \infty$, we first note that

$$M - M_n = \sum_{i=n}^{\infty} (M_{i+1} - M_i) = \sum_{i=n}^{\infty} \frac{Q_i^{1/2} Z_{i+1}}{i+1}.$$

Therefore, for each $n \geq 2$,

$$\begin{aligned}
\mathbb{E}\big\{\|M_n - M\|_{\mathrm{E}}\big\}^2 \leq \mathbb{E}\big\{\|M_n - M\|_{\mathrm{E}}^2\big\} &= \mathbb{E}\big\{(M - M_n)^{\top}(M - M_n)\big\} \\
&= \sum_{i=n}^{\infty} \frac{\mathbb{E}\big(Z_{i+1}^{\top} Q_i Z_{i+1}\big)}{(i+1)^2} \\
&= \sum_{i=n}^{\infty} \sum_{r,k=1}^{p} \frac{\mathbb{E}\big(Z_{i+1}^{(r)} Z_{i+1}^{(k)}\big)\mathbb{E}\big(Q_i^{(r,k)}\big)}{(i+1)^2} \\
&= \sum_{i=n}^{\infty} \sum_{k=1}^{p} \frac{\mathbb{E}\big(Q_i^{(k,k)}\big)}{(i+1)^2} \leq \sum_{k=1}^{p} \mathbb{E}(Q_2^{(k,k)}) \sum_{i=n}^{\infty} \frac{1}{(i+1)^2},
\end{aligned}$$

where the last inequality is because $\{Q_i^{(k,k)} : i \geq 2\}$ is a supermartingale for fixed k. On noting that $\sum_{i=n}^{\infty} 1/(i+1)^2 \leq 1/n$, we obtain

$$\sqrt{n}\mathbb{E}\big\{\|M_n - M\|_{\mathrm{E}}\big\} \leq \sqrt{\sum_{k=1}^{p} \mathbb{E}\big(Q_2^{(k,k)}\big)}.$$

Similarly, recalling that

$$Q_{n+1} = \frac{n}{n+1} Q_n^{1/2} \left(I + \frac{Z_{n+1} Z_{n+1}^{\top}}{n+1} \right) Q_n^{1/2},$$

we obtain

$$Q - Q_n = \sum_{i=n}^{\infty} (Q_{i+1} - Q_i) = \sum_{i=n}^{\infty} \left\{ \frac{Q_i^{1/2}\big(Z_{i+1} Z_{i+1}^{\top} - I\big) Q_i^{1/2}}{i+1} - \frac{Q_i^{1/2} Z_{i+1} Z_{i+1}^{\top} Q_i^{1/2}}{(i+1)^2} \right\}.$$

Define $q = \sup_i \mathbb{E}\big\{\|Q_i^{1/2}\|_{\mathrm{F}}^4\big\}$ and $K_i = Q_i^{1/2}\big(Z_{i+1} Z_{i+1}^{\top} - I\big) Q_i^{1/2}/(i+1)$. Since $\mathbb{E}\big\{\|Q_i^{1/2}\|_{\mathrm{F}}^2\big\} \leq \sqrt{q}$ for all i,

$$\begin{aligned}
\sum_{i=n}^{\infty} \frac{\mathbb{E}\big\{\|Q_i^{1/2} Z_{i+1} Z_{i+1}^{\top} Q_i^{1/2}\|_{\mathrm{F}}\big\}}{(i+1)^2} &\leq \sum_{i=n}^{\infty} \frac{\mathbb{E}\big\{\|Z_{i+1} Z_{i+1}^{\top}\|_{\mathrm{F}}\big\}\mathbb{E}\big\{\|Q_i^{1/2}\|_{\mathrm{F}}^2\big\}}{(i+1)^2} \\
&\leq \mathbb{E}\big\{\|Z_1 Z_1^{\top}\|_{\mathrm{F}}\big\} \sum_{i=n}^{\infty} \frac{\sqrt{q}}{(i+1)^2} \leq \frac{\sqrt{q}}{n} \mathbb{E}\big\{\|Z_1 Z_1^{\top}\|_{\mathrm{F}}\big\}.
\end{aligned}$$

Since K_i is symmetric and $\mathbb{E}(K_i K_j) = 0$ for $i \neq j$,

$$\mathbb{E}\left\{ \left\| \sum_{i=n}^{\infty} K_i \right\|_{\mathrm{F}}^2 \right\} = \mathbb{E}\left\{ \operatorname{trace}\left(\left(\sum_{i=n}^{\infty} K_i \right) \left(\sum_{j=n}^{\infty} K_j \right) \right) \right\}$$

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

$$
\begin{aligned}
&= \operatorname{trace}\left(\sum_{i,j=n}^{\infty} \mathbb{E}(K_i K_j)\right) \\
&= \operatorname{trace}\left(\sum_{i=n}^{\infty} \mathbb{E}(K_i K_i)\right) \\
&= \sum_{i=n}^{\infty} \mathbb{E}\left\{\frac{\|Q_i^{1/2}(Z_{i+1}Z_{i+1}^\top - I)Q_i^{1/2}\|_{\mathrm{F}}^2}{(i+1)^2}\right\} \\
&\le \mathbb{E}\{\|Z_1 Z_1^\top - I\|_{\mathrm{F}}^2\} \sum_{i=n}^{\infty} \frac{\mathbb{E}\{\|Q_i^{1/2}\|_{\mathrm{F}}^4\}}{(i+1)^2} \le \mathbb{E}\{\|Z_1 Z_1^t - I\|_{\mathrm{F}}^2\}\frac{q}{n}.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\sqrt{n}\mathbb{E}\{\|Q_n - Q\|_{\mathrm{F}}\} &\le \sqrt{n}\sum_{i=n}^{\infty} \frac{\mathbb{E}\{\|Q_i^{1/2} Z_{i+1} Z_{i+1}^\top Q_i^{1/2}\|_{\mathrm{F}}\}}{(i+1)^2} + \sqrt{n}\mathbb{E}\left\{\left\|\sum_{i=n}^{\infty} K_i\right\|_{\mathrm{F}}\right\} \\
&\le \frac{\sqrt{q}}{\sqrt{n}}\mathbb{E}\{\|Z_1 Z_1^\top\|_{\mathrm{F}}\} + \sqrt{q\mathbb{E}\{\|Z_1 Z_1^\top - I\|_{\mathrm{F}}^2\}}.
\end{aligned}
$$

Hence, $b < \infty$ provided $q < \infty$.

Finally, we prove $q < \infty$. Fix $i \ge 2$, $1 \le k \le p$, and define $W_i = \big(Q_i^{(k,k)}\big)^2$. As noted in the proof of Theorem 1, the conditional distribution of $L_j = Q_j^{1/2} Z_{j+1}$ given $\mathcal{F}_j$ is $\mathcal{N}(0, Q_j)$. Hence, $\mathbb{E}\{\big(L_j^{(k)}\big)^2 \mid \mathcal{F}_j\} = Q_j^{(k,k)}$ and $\mathbb{E}\{\big(L_j^{(k)}\big)^4 \mid \mathcal{F}_j\} = 3\big(Q_j^{(k,k)}\big)^2 = 3W_j$. Using (10), it follows that

$$
\mathbb{E}(W_{j+1} \mid \mathcal{F}_j) = W_j\left\{\frac{j^2}{(j+1)^2} + \frac{3j^2}{(j+1)^4} + \frac{2j^2}{(j+1)^3}\right\} \le W_j.
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}(W_i) = \mathbb{E}\left\{W_2 \prod_{j=2}^{i-1} \frac{W_{j+1}}{W_j}\right\} &= \mathbb{E}(W_2)\prod_{j=2}^{i-1} \mathbb{E}\left\{\frac{W_{j+1}}{W_j}\right\} \\
&= \mathbb{E}(W_2)\prod_{j=2}^{i-1} \mathbb{E}\left\{\frac{\mathbb{E}(W_{j+1} \mid \mathcal{F}_j)}{W_j}\right\} \le \mathbb{E}(W_2).
\end{aligned}
$$

To sum up, $\mathbb{E}\{\big(Q_i^{(k,k)}\big)^2\} \le \mathbb{E}\{\big(Q_2^{(k,k)}\big)^2\}$ for all $i \ge 2$ and $1 \le k \le p$. Hence,

$$
\begin{aligned}
\mathbb{E}\{\|Q_i^{1/2}\|_{\mathrm{F}}^4\} = \mathbb{E}\{\operatorname{trace}(Q_i)^2\} &= \sum_{r,k=1}^{p} \mathbb{E}\{Q_i^{(r,r)} Q_i^{(k,k)}\} \\
&\le \sum_{r,k=1}^{p} \sqrt{\mathbb{E}\{\big(Q_i^{(r,r)}\big)^2\}\mathbb{E}\{\big(Q_i^{(k,k)}\big)^2\}} \\
&\le p^2 \max_{1\le k\le p} \mathbb{E}\{\big(Q_i^{(k,k)}\big)^2\} \le p^2 \max_{1\le k\le p} \mathbb{E}\{\big(Q_2^{(k,k)}\big)^2\}.
\end{aligned}
$$

This proves that $q < \infty$ and concludes the proof of the theorem. □

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

Let us turn to the results of Section 2.3. In such results, the predictives of X are the copula-based predictive distributions β_n. Hence, as noted in Section 2.3, X is CID.

Proof of Theorem 3. As in Example 4, define $D_n = \int_{-\infty}^{\infty} |f_{n+1}(x \mid X(n+1)) - f_n(x \mid X(n))| \, \mathrm{d}x$. For $n < m$,

$$\int_{-\infty}^{\infty} \left| f_n(x \mid X(n)) - f_m(x \mid X(m)) \right| \mathrm{d}x = \int_{-\infty}^{\infty} \left| \sum_{i=n}^{m-1} \left\{ f_{i+1}(x \mid X(i+1)) - f_i(x \mid X(i)) \right\} \right| \mathrm{d}x \leq \sum_{i=n}^{m-1} D_i.$$

If $\sum_n D_n < \infty$ a.s., the sequence $f_n(\cdot \mid X(n))$ is Cauchy in the L^1-norm a.s., so that it converges in the L^1-norm a.s. Hence, it suffices to show that $\sum_n D_n < \infty$ a.s. To this end, recall that D_n is distributed as $r_n \int_0^1 |c_n(u, V) - 1| \, \mathrm{d}u$, where V is any random variable with uniform distribution on (0,1); see Example 4. It follows that

$$\begin{aligned} \mathbb{E}(D_n) &= r_n \mathbb{E}\left\{ \int_0^1 |c_n(u, V) - 1| \, \mathrm{d}u \right\} \\ &= r_n \int_0^1 \mathbb{E}\{|c_n(u, V) - 1|\} \, \mathrm{d}u = r_n \int_0^1 \int_0^1 |c_n(u, v) - 1| \, \mathrm{d}v \, \mathrm{d}u \leq 2 r_n. \end{aligned}$$

Therefore, $\sum_n \mathbb{E}(D_n) \leq 2 \sum_n r_n < \infty$, which in turn implies that $\sum_n D_n < \infty$ a.s. □

Proof of Theorem 4. Since X is CID, by [5, Theorem 4] it suffices to show that

$$\sup_n \mathbb{E}\left\{ \int_{-\infty}^{\infty} f_n(x \mid X(n))^2 \, \mathrm{d}x \right\} < \infty. \tag{11}$$

To prove (11), we note that, since X is CID, $\mathbb{E}\{f_{k+1}(x \mid X(k+1)) \mid X(k)\} = f_k(x \mid X(k))$ for almost every $x \in \mathbb{R}$ (with respect to Lebesgue measure). This implies that

$$\mathbb{E}\{c_k(F_k(x \mid X(k)), F_k(X_{k+1} \mid X(k))) \mid X(k)\} = 1$$

for almost every $x \in \mathbb{R}$. Since $c_k(u, v) \leq b$ for all $u, v \in [0, 1]$, it follows that

$$\mathbb{E}\left\{ \left(\frac{f_{k+1}(x \mid X(k+1))}{f_k(x \mid X(k))} \right)^2 \mid X(k) \right\} \leq (1 - r_k)^2 + r_k^2 b^2 + 2(1 - r_k) r_k = 1 + r_k^2 (b^2 - 1).$$

Finally, the previous inequality implies that

$$\begin{aligned} \mathbb{E}\{f_n(x \mid X(n))^2\} &= \mathbb{E}\left\{ \left(f_0(x) \prod_{k=0}^{n-1} \frac{f_{k+1}(x \mid X(k+1))}{f_k(x \mid X(k))} \right)^2 \right\} \\ &\leq f_0(x)^2 \prod_{k=0}^{n-1} \{1 + r_k^2 (b^2 - 1)\} \leq f_0(x)^2 \exp\left\{ (b^2 - 1) \sum_{k=0}^{n-1} r_k^2 \right\}. \end{aligned}$$

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

Thus, to check condition (11), it suffices to note that

$$\mathbb{E}\left\{\int_{-\infty}^{\infty} f_n(x \mid X(n))^2 \, \mathrm{d}x\right\} = \int_{-\infty}^{\infty} \mathbb{E}\{f_n(x \mid X(n))^2\} \, \mathrm{d}x$$
$$\leq \exp\left\{(b^2 - 1)\sum_{k=0}^{\infty} r_k^2\right\} \int_{-\infty}^{\infty} f_0(x)^2 \, \mathrm{d}x. \qquad \square$$

Appendix B. Additional numerical illustrations

This appendix shows the plots of the densities of the posterior distributions $\Pi_N^{\alpha}(\cdot \mid x)$ (filled) and $\Pi_N^{\beta}(\cdot \mid x)$ (open) obtained via PR, for all estimated population parameters, showing the density obtained with both α_n and β_n. Figures 3, 4, 5, 6, and 7 refer respectively to the mean, variance, skewness, kurtosis, and median. The dashed horizontal lines represent the true parameter value θ_0. To interpret these plots, one caveat is in order. Some figures show that θ_0 is caught in the main body of the posterior density, while others do not. Thus, one might think that in those cases better parameter estimates are obtained (i.e. smaller RMSE). However, this is not necessarily true. For instance, Figures 3 and 6 appear to suggest that α_n estimates population mean better than skewness. Actually, it is exactly the opposite (see Table 1). Indeed, PR with α_n produces skewness estimates which have such low variability that, despite looking further from θ_0, they are actually closer on average. Thus, these plots should always be compared with RMSE (Tables 1 and 2) and posterior variance (Tables 3 and 4).

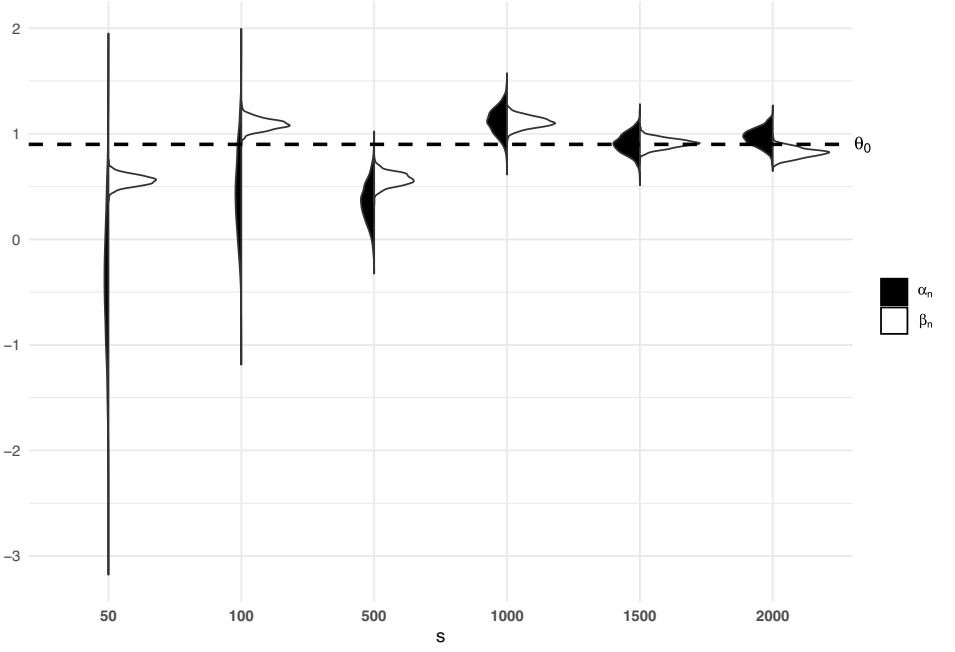


FIGURE 3. Mean.

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

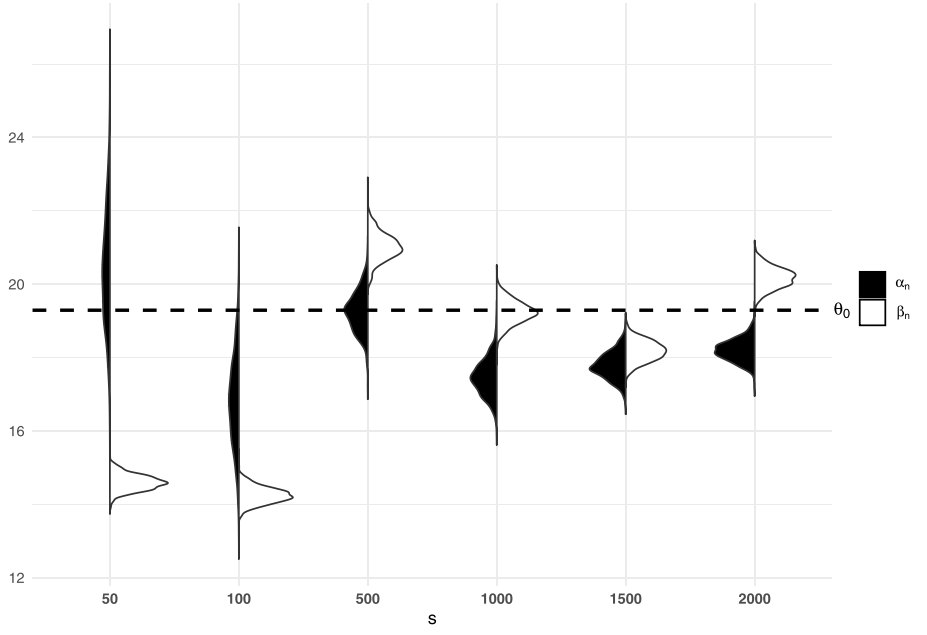


FIGURE 4. Variance.

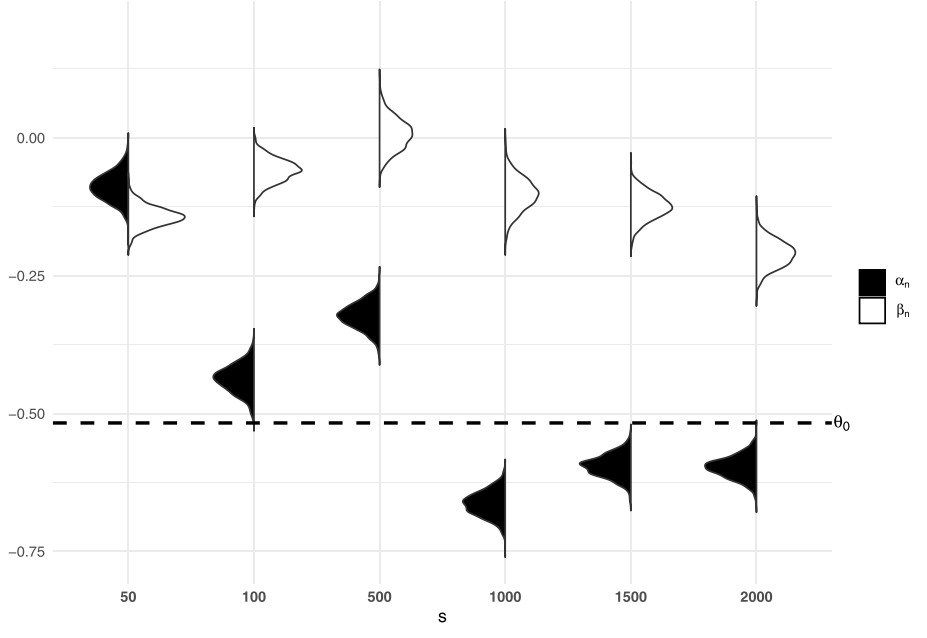


FIGURE 5. Skewness.

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

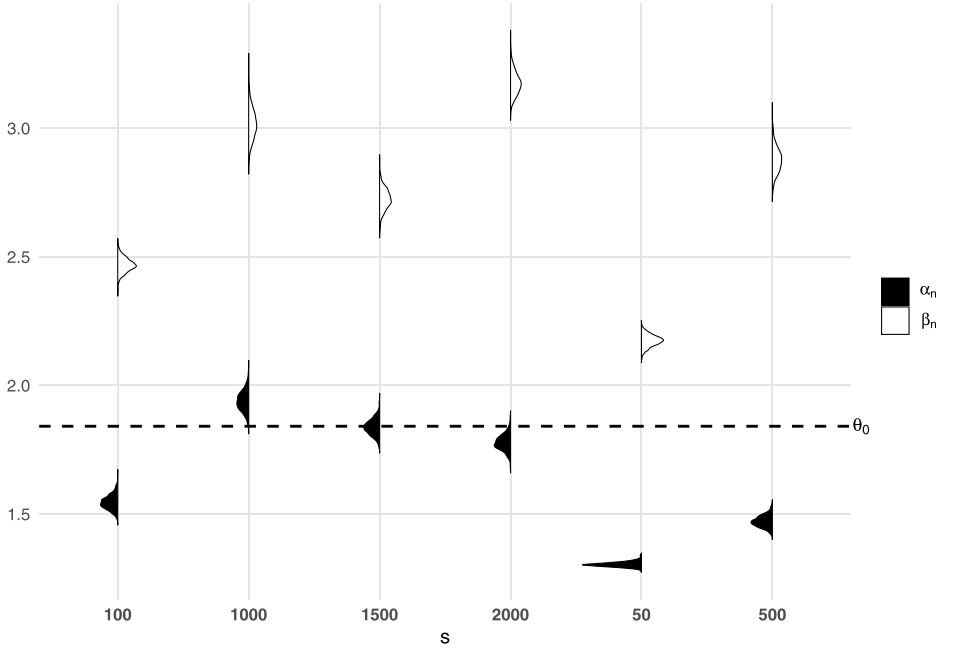


FIGURE 6. Kurtosis.

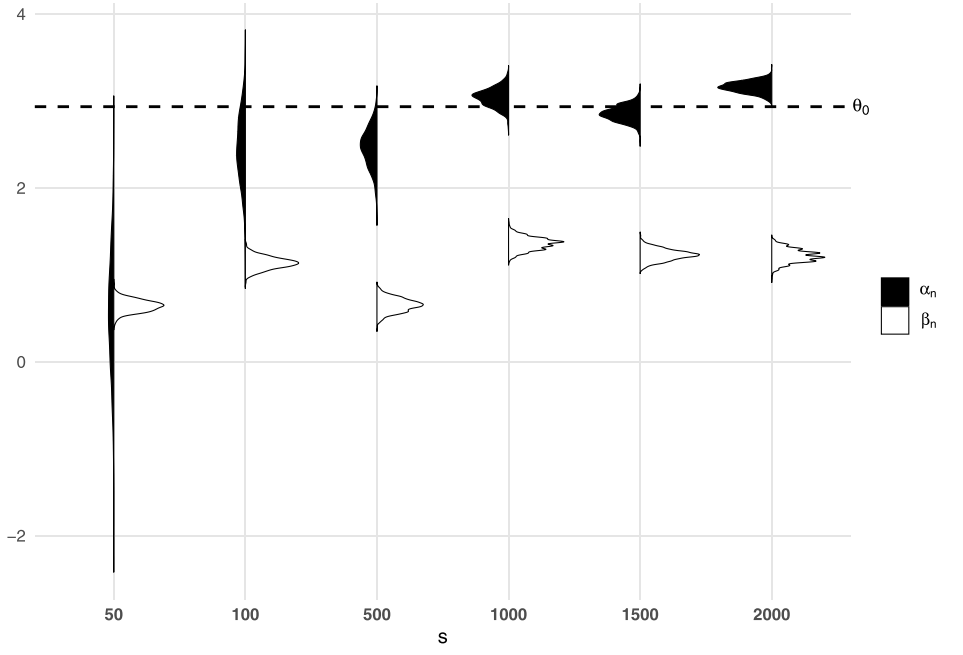


FIGURE 7. Median.

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press

References

[1] Berti, P., Dreassi, E., Leisen, F., Pratelli, L. and Rigo, P. (2023). Bayesian predictive inference without a prior. *Statistica Sinica* **33**, 2405–2429.

[2] Berti, P., Dreassi, E., Leisen, F., Pratelli, L. and Rigo, P. (2025). A probabilistic view on predictive constructions for Bayesian learning. *Statist. Sci.* **40**, 25–39.

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

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

[5] Berti, P., Pratelli, L. and Rigo, P. (2013). Exchangeable sequences driven by an absolutely continuous random measure. *Ann. Prob.* **41**, 2090–2102.

[6] Clarke, B., Fokoue, E. and Zhang, H. H. (2009). *Principles and Theory for Data Mining and Machine Learning*. Springer, New York.

[7] Cui, F. and Walker, S. G. (2024). A Bayesian bootstrap for mixture models. *Bayesian Anal.*, to appear.

[8] Dawid, A. P. (1984). Present position and potential developments: Some personal views – statistical theory, the prequential approach. *J. R. Statist. Soc. A* **147**, 278–292.

[9] Dawid, A. P. and Vovk, V. G. (1999). Prequential probability: Principles and properties. *Bernoulli* **5**, 125–162.

[10] Devroye, L., Mehrabian, A. and Reddad, T. (2023). The total variation distance between high-dimensional Gaussians with the same mean. Preprint, arXiv:1810.08693v7.

[11] Ding, J. and Zhou, A. (2007). Eigenvalues of rank-one updated matrices with some applications. *Appl. Math. Lett.* **20**, 1223–1226.

[12] Dutordoir, V., Saul, A., Ghahramani, Z. and Simpson, F. (2023). Neural diffusion processes. *Proc. Mach. Learn. Res*. **202**, 8990–9012.

[13] Efron, B. (2020). Prediction, estimation, and attribution. *J. Amer. Statist. Assoc.* **115**, 636–655.

[14] Fong, E., Holmes, C., and Walker, S. G. (2023). Martingale posterior distributions (with discussion). *J. R. Statist. Soc. B* **85**, 1357–1391.

[15] Fortini, S. and Petrone S. (2025). Exchangeability, prediction and predictive modeling in Bayesian statistics. *Statist. Sci.* **40**, 40–67.

[16] Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis and density estimation. *J. Amer. Statist. Assoc.* **97**, 611–631.

[17] Fraley, C. and Raftery, A. (2007). Bayesian regularization for normal mixture estimation and model-based clustering, *J. Classification* **24**, 155–181.

[18] Fraley, C., Raftery, A., Murphy, T., and Scrucca, L. (2012). Mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report No. 597, Department of Statistics, University of Washington.

[19] Hahn, P. R., Martin, R. and Walker, S. G. (2018). On recursive Bayesian predictive distributions. *J. Amer. Statist. Assoc.* **113**, 1085–1093.

[20] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

[21] Holmes, C. and Walker, S. G. (2023). Statistical inference with exchangeability and martingales. *Phil. Trans. R. Soc. A* **381**, 20220143.

[22] Kelbert, M. (2023). Survey of distances between the most popular distributions. *Analytics* **2**, 225–245.

[23] Li, F., Ding, P. and Mealli, F. (2023). Bayesian causal inference: A critical review. *Phil. Trans. R. Soc. A*, **381**, 20220153.

[24] Lu, L. and Ghosh, S. (2023). Nonparametric estimation of multivariate copula using empirical Bayes methods. *Mathematics* **11**, 4383.

[25] Pitman, J. (1996). Some developments of the Blackwell–MacQueen urn scheme. In *Statistics, Probability and Game Theory* (IMS Lect. Notes Mon. Ser. **30**), eds T. S. Ferguson, L. S. Shapley and J. B. MacQueen. Institute of Mathematical Studies, Hayward, CA, pp. 245–267.

[26] Pitman, J. and Yor, M. (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. *Ann. Prob.* **25**, 855–900.

[27] Sancetta, A. and Satchell, S. (2004). The Bernstein copula and its applications to modeling and approximations of multivariate distributions. *Econometric Theory* **20**, 535–562.

https://doi.org/10.1017/jpr.2026.10119 Published online by Cambridge University Press