Journal of the Royal Statistical Society Series B: Statistical Methodology, 2023, **85**, 1357–1391
https://doi.org/10.1093/jrsssb/qkad005

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology

Discussion Paper

Martingale posterior distributions

Edwin Fong[1,2], Chris Holmes[1,2] and Stephen G. Walker[3]

[1]The Alan Turing Institute, London, UK
[2]Department of Statistics, University of Oxford, Oxford, UK
[3]Department of Statistics and Data Sciences, University of Texas, Austin, United States

Address for correspondence: Chris Holmes, Department of Statistics, University of Oxford, 24-29 St Giles', Oxford OX2 3LB, UK. Email: cholmes@stats.ox.ac.uk

Read before The Royal Statistical Society at a meeting organized by the Research Section on Monday, 12 December 2022, Dr Maria De Lorio in the Chair.

Abstract

The prior distribution is the usual starting point for Bayesian uncertainty. In this paper, we present a different perspective that focuses on missing observations as the source of statistical uncertainty, with the parameter of interest being known precisely given the entire population. We argue that the foundation of Bayesian inference is to assign a distribution on missing observations conditional on what has been observed. In the i.i.d. setting with an observed sample of size n, the Bayesian would thus assign a predictive distribution on the missing $Y_{n+1:\infty}$ conditional on $Y_{1:n}$, which then induces a distribution on the parameter. We utilize Doob's theorem, which relies on martingales, to show that choosing the Bayesian predictive distribution returns the conventional posterior as the distribution of the parameter. Taking this as our cue, we relax the predictive machine, avoiding the need for the predictive to be derived solely from the usual prior to posterior to predictive density formula. We introduce the *martingale posterior distribution*, which returns Bayesian uncertainty on any statistic via the direct specification of the joint predictive. To that end, we introduce new predictive methodologies for multivariate density estimation, regression and classification that build upon recent work on bivariate copulas.

Keywords: Bayesian uncertainty, copula, martingale, predictive inference

1 Introduction

Statistical uncertainty in a parameter of interest arises due to missing observations. If a complete population is observed, then the parameter of interest can be assumed to be known precisely. In this paper, we argue that the Bayesian accounts for this uncertainty by constructing a distribution on the missing observations conditional on what has been observed. This, in turn, induces a distribution on the parameter given the observed data, which we will see is the posterior distribution. In this work, we will describe and generalize this framework in detail for the case where the observations are independent and identically distributed (i.i.d.), and we will also briefly consider other data structures.

In the i.i.d. case, given $Y_{1:n} \overset{\text{iid}}{\sim} F_0$, where F_0 is the unknown true sampling distribution, the missing observations are the remaining $Y_{n+1:\infty}$, and as such we focus our modelling efforts directly on the predictive density

$$p(y_{n+1:\infty} \mid y_{1:n}). \tag{1.1}$$

Here, the construction of the predictive density is for parameter inference and not for forecasting future observations as is more usual. For inference, we assume that the object of interest is fully defined once all the observations have been viewed, which we write as $\theta_\infty = \theta(Y_{1:\infty})$. It is clear then that (1.1) induces a distribution on θ_∞, and we call this scheme of imputing $Y_{n+1:\infty}$ and

Received: March 29, 2021. **Revised:** February 14, 2022. **Accepted:** February 18, 2022

computing θ_∞ as *predictive resampling*. A key observation is that $Y_{1:\infty}$ will always contain the observed $Y_{1:n} = y_{1:n}$ as the predictive Bayesian considers the observed sample to be fixed, in contrast to the frequentist consideration of other possible values of $Y_{1:n}$.

For i.i.d. observations, the traditional Bayesian approach is to elicit a prior density $\pi(\theta)$ and sampling density $f_\theta(y)$, derive the posterior $\pi(\theta \mid y_{1:n})$, then compute the predictive density through

$$p(y \mid y_{1:n}) = \int f_\theta(y)\pi(\theta \mid y_{1:n})\,\mathrm{d}\theta. \tag{1.2}$$

A concise summary of our approach is the following: while de Finetti (1937) provided a representation of Bayesian inference, which relies on exchangeability and the prior distribution, we will introduce a framework based on the results of Doob (1949), which relies solely, in the i.i.d. case, on the predictive distribution. We will see that this framework based on Doob's results is more flexible and the mathematical requirement amounts to the construction of a martingale—it is this flexibility we exploit in this paper. In fact, through Doob's theorem, we will see that predictive resampling, as described above, is identical to posterior sampling when using (1.2) as the predictive and θ indexes the sampling density, in which case $\theta_\infty \sim \pi(\theta \mid y_{1:n})$. Denoting by $p(y)$ the prior predictive, this connection is illustrated below for the traditional Bayesian case:

$$f_\theta(y),\, \pi(\theta) \underset{\text{Bayes' rule}}{\longrightarrow} \pi(\theta \mid y_{1:n}) \xrightarrow[\int f_\theta(y)\pi(\theta \mid y_{1:n})\,\mathrm{d}\theta]{\text{posterior predictive}} p(y \mid y_{1:n})$$

$$\pi(\theta \mid y_{1:n}) \xleftarrow[Y_{n+1:\infty} \sim p(\cdot \mid y_{1:n})]{\text{Doob's theorem}} p(y \mid y_{1:n}) \underset{\text{predictive update}}{\longleftarrow} p(y)$$

However, the traditional Bayesian focus on the prior on θ makes no appeal to the underlying cause of the uncertainty, that is, the unobserved part of the study population $Y_{n+1:\infty}$. Furthermore, the traditional prior-to-posterior computation is becoming increasingly strained as model complexity and data sizes grow. In our work, we advocate the predictive resampling strategy—given $y_{1:n}$, our starting point is directly the predictive model (1.1) and the target statistic of interest θ_∞, noting now that θ_∞ is no longer restricted to indexing the sampling density. We relax de Finetti's assumption of exchangeability, but we must now take care to construct (1.1) so that θ_N is indeed convergent to some θ_∞, where $\theta_N = \theta(Y_{1:N})$ can be viewed as an estimator. We highlight here that we use n and N for the size of the observed dataset and the imputed population, respectively. In the spirit of Doob, we rely heavily on martingales, which also aid in ensuring that expectations of limits coincide with fixed quantities seen at the sample of size n. This can be regarded as a predictive coherency condition, and we designate the distribution of θ_∞ as the *martingale posterior*. Our choice of (1.1) will be density estimators based on recent ideas in the literature, specifically the *conditionally identically distributed* (c.i.d.) sequence of Berti et al. (2004) and bivariate copula update of Hahn et al. (2018).

We now discuss why one would want to go through the route of obtaining the martingale posterior via the induced distribution of θ_∞ from (1.1) rather than the traditional likelihood-prior construction. Firstly, predictive models are probabilistic statements on observables, which removes the need to elicit subjective probability distributions on parameters that may have no real-world interpretations and only index the sampling density. Secondly, the martingale posterior establishes a direct connection between prediction and statistical inference, opening up the possibility of using modern probabilistic predictive methods for inference (Breiman, 2001) and transparently acknowledges the source of statistical uncertainty as the missing $Y_{n+1:\infty}$. Thirdly, working directly with predictive distributions is highly practical. For an elicited 1-step ahead predictive, we can predictively resample by carrying out the recursive update

$$\{p(y \mid y_{1:N-1}),\, y_N\} \mapsto p(y \mid y_{1:N})$$

to sample $Y_{n+1:N}$ for a large enough N such that θ_N has effectively converged to a sample from the martingale posterior, or N matches a known finite study population size. In complex scenarios such as multivariate density estimation and regression, we introduce new copula-based methodologies where our computations remain exact, GPU-friendly and parallelizable, returning us Bayesian uncertainty without any reliance on Markov chain Monte Carlo (MCMC). Finally, a predictive approach more clearly delineates the core similarities and differences between Bayesian and frequentist uncertainty.

We will focus on the i.i.d. data setting in this work, which corresponds to exchangeable traditional Bayesian models. In this setting, the martingale posterior can indeed be regarded as a generalization of the traditional Bayesian model, as the class of c.i.d. models is more general and contains the class of exchangeable models which we will see in Section 3.2. In more complex data structures beyond i.i.d. data, such as those encountered in hierarchical modelling or time series, our framework would still apply. In this case, the missing observations we require may no longer be $Y_{n+1:\infty}$, and model elicitation would no longer only involve a sequence of predictive distributions. For example, a simple hierarchical setting is the observation process $Y_i \sim p(y_i \mid \theta_i)$, where θ_i is itself drawn from an unknown G_0 and we may be interested in some functional $\gamma(G_0)$. Here, we only observe $Y_{1:n} = y_{1:n}$, so the missing observations of interest are now the unobserved random effects $\theta_{1:\infty}$. We can thus seek to impute $\theta_{1:n} \sim p(\theta_{1:n} \mid y_{1:n})$ from the data, followed by the missing remainder $\theta_{n+1:\infty} \sim p(\theta_{n+1:\infty} \mid \theta_{1:n})$. Computing $\gamma(\theta_{1:\infty})$ would then return us a posterior sample. For the remainder of the paper, we will focus only on the i.i.d. case and leave the details of non-i.i.d. settings for future work.

In Section 2, we formally investigate the connection between predictive and posterior inference and introduce a predictive framework for inference and the resulting martingale posterior. We then utilize the bootstrap as a canonical example to distinctly compare Bayesian and frequentist uncertainty. We postpone the discussion of related work until Section 2.5 in order to provide context beforehand. In Section 3, we discuss predictive coherence conditions for martingale posteriors, utilizing c.i.d. sequences. In Section 4, we revisit the bivariate copula methodology of Hahn et al. (2018) for univariate density estimation and extend it to obtain the martingale posterior. We then generalize this copula-based method to multivariate density estimation, regression and classification. Section 5 then provides a thorough demonstration of the above methods through examples. In Section 6, we discuss some theoretical properties of the martingale posterior with the copula-based methodology. Finally, we discuss our results in Section 7.

2 A predictive framework for inference

2.1 Doob's theorem and Bayesian uncertainty

Uncertainty quantification lies at the core of statistical inference, and Bayesian inference is one framework for handling uncertainty in a formal manner. The Bayesian begins with the random variables $(\Theta, Y_1, Y_2, \ldots)$, where $(Y_1, Y_2, \ldots)$ are the observables of interest, and Θ is the parameter which indexes the sampling density $f_\theta(y)$. We assume throughout that the appropriate densities exist. For i.i.d. data, the Bayesian elicits a joint probability model for the observables and parameter with joint density

$$p(\theta, y_{1:N}) = \pi(\theta) \prod_{i=1}^{N} f_\theta(y_i) \tag{2.1}$$

for each N. Here, the density $\pi(\theta)$ represents prior knowledge about the parameter which generates the observations, and under a subjectivist point of view, $\Pi(A) = \int_A \pi(\theta)\, d\theta$ represents the subjective probability that the generating parameter value Θ lies in the set A. Marginalizing out Θ gives the joint density of the observables,

$$p(y_{1:N}) = \int \prod_{i=1}^{N} f_\theta(y_i)\, d\Pi(\theta). \tag{2.2}$$

De Finetti, however, argued that the direct likelihood-prior interpretation of the Bayesian model was insufficient, as Θ is of a 'metaphysical' nature and probability statements should only be on observables (Bernardo & Smith, 2009). This then motivated the notion of exchangeability of the infinite sequence $(Y_1, Y_2, \ldots)$, where the joint probability P of the finite sequence of observables $Y_{1:N} = (Y_1, \ldots, Y_N)$ is invariant to the ordering of Y_i for all N. Through de Finetti's representation theorem (de Finetti, 1937) and extensions thereof (e.g., Hewitt & Savage, 1955), the assumption

of exchangeability induces the likelihood-prior form of the joint density in (2.2) (where Π may not have a density), which motivates such a specification of the Bayesian model. The representation theorem, however, is only part of the story. As alluded to in Section 1, the source of statistical uncertainty is the lack of the infinite dataset $Y_{n+1:\infty}$ with which we could pin down any quantity of interest precisely. Bayesian uncertainty through the lens of the prior is still opaque in this regard, even with the aforementioned representation theorem.

The key to understanding the source of uncertainty lies in the predictive imputation of observables, for which we require a result from Doob. Doob (1949) established consistency of the Bayesian method when the observations are distributed according to (2.2). For this result, we require that the model is identifiable, that is $F_\theta \neq F_{\theta'}$ whenever $\theta \neq \theta'$, where F_θ is the cumulative distribution function of f_θ. Let us assume that data has yet to be observed, so the missing observations are $Y_{1:\infty}$. Following the discussion in Section 1, one can regard (2.2) as the joint predictive density on the missing population and can estimate the parameter indexing the sampling density as a function of the imputed $Y_{1:N}$. An appropriate and intuitive point estimate for the Bayesian is the posterior mean, which we write as

$$\bar{\theta}_N = E[\Theta \mid Y_{1:N}].$$

We now use a secondary result of Doob (1949) to confirm that the prior uncertainty in Θ arises from the predictive uncertainty in $Y_{1:\infty}$.

Theorem 1 (Doob, 1949). Assume Θ is in a linear space with $E[|\Theta|] < \infty$, and $(\Theta, Y_1, Y_2, \ldots)$ is distributed according to (2.1), so $\Theta \sim \Pi$. Under identifiability and measurability conditions on F_θ, we have

$$\bar{\theta}_N \to \Theta \quad \text{a.s.}$$

For the above result, the key is to rely on $\bar{\theta}_N$ being a martingale, that is

$$E[\bar{\theta}_N \mid Y_{1:N-1}] = \bar{\theta}_{N-1}$$

almost surely. Doob's martingale convergence theorem then ensures that $\bar{\theta}_N$ converges to a limit almost surely. The identifiability condition ensures that the parameter is recoverable from the infinite sample so that the limit of $\bar{\theta}_N$ is indeed Θ. For Θ in more general metric spaces, consistency results with general notions of posterior expectations are provided in Ghosal and van der Vaart (2017, Theorem 6.8). As an aside, we highlight that Doob (1949) provided a more general result: the Bayesian posterior distribution converges weakly to the Dirac measure δ_Θ almost surely for Π-almost every Θ as $N \to \infty$. The technical details of a more general version of this result can be found in Ghosal and van der Vaart (2017, Theorem 6.9). In the Bayesian nonparametric (BNP) case where Θ is a probability density function, we have a nonparametric extension of the above results (Lijoi et al., 2004).

Returning to the task at hand, we can summarize the above by considering two distinct methods of sampling Θ from the prior Π before seeing any data. The first is to draw $\Theta \sim \Pi$ directly, which is the opaque view of the inherently random parameter that we are trying to shed light on. The second, which inspires the remainder of our paper, begins with sequentially imputing the unseen observables $Y_1, Y_2, Y_3 \ldots$ from the sequence of predictive densities

$$Y_1 \sim p(\cdot), \quad Y_2 \sim p(\cdot \mid y_1), \quad Y_3 \sim p(\cdot \mid y_2, y_1), \ \ldots$$

until we have the complete information $Y_{1:\infty}$ in the limit. Given this random infinite dataset, the limiting point estimate $\bar{\theta}_\infty = \lim_{N\to\infty} \bar{\theta}_N$, that is the posterior mean computed on the entire dataset, is in fact distributed according to Π. This equivalence highlights the fact that *a priori* uncertainty in Θ is a consequence of the uncertainty in $Y_{1:\infty}$, and the function $\bar{\theta}$ provides a means to precisely recover our quantity of interest when all information is made available to us.

Of course, such an interpretation is equally valid *a posteriori*, that is after we have observed $Y_{1:n} = y_{1:n}$. Here, sampling $\Theta \sim \Pi(\cdot \mid y_{1:n})$ is equivalent to sampling $Y_{n+1:\infty}$ conditional on $y_{1:n}$ and computing $\bar{\theta}_\infty$ as if we have observed the infinite dataset, noting that $Y_{1:n} = y_{1:n}$ is now fixed. This can be seen by simply substituting the prior π in (2.1), (2.2) and Theorem 1 with the posterior $\pi(\cdot \mid y_{1:n})$. In conclusion, Doob's result highlights that the Bayesian seeks to simulate what is needed to pin down the parameter but is missing from reality, that is $Y_{n+1:\infty}$ in the i.i.d. case, and we find this to be a compelling justification for the Bayesian approach.

We now conclude this section with a concrete demonstration of the equivalence between posterior sampling and the forward sampling of $Y_{n+1:\infty}$ through a simple normal model with unknown mean based on an example from Hahn (2015).

Example 1 Let $f_\theta(y) = \mathcal{N}(y \mid \theta, 1)$, with $\pi(\theta) = \mathcal{N}(\theta \mid 0, 1)$. Given an observed dataset $y_{1:n}$, the tractable posterior density takes on the form $\pi(\theta \mid y_{1:n}) = \mathcal{N}(\theta \mid \bar{\theta}_n, \bar{\sigma}_n^2)$ where

$$\bar{\theta}_n = \frac{\sum_{i=1}^n y_i}{n+1}, \quad \bar{\sigma}_n^2 = \frac{1}{n+1}.$$

The posterior predictive density then takes on the form $p(y \mid y_{1:n}) = \mathcal{N}(y \mid \bar{\theta}_n, 1 + \bar{\sigma}_n^2)$. For observed data, we generated $y_{1:n} \overset{\text{iid}}{\sim} f_\theta(y)$ for $n = 10$ with $\theta = 2$, giving $\bar{\theta}_n = 1.84$.

We can plot the independent sample paths for the posterior mean, $\bar{\theta}_{n+1:N}$, as we recursively forward sample $Y_{n+1:N}$, where $N = n + 1000$ in this example. In Figure 1, we see that the sample paths of $\bar{\theta}_{n+i}$ each converge to a random Θ as i increases, with the density of $\bar{\theta}_N$ very close to the analytic posterior. From Doob's consistency theorem, we know this is exact for $N \to \infty$.

2.2 The methodological approach

Through Doob's result in Theorem 1, we have demonstrated the predictive view of Bayesian inference as a means to understand how the posterior uncertainty in Θ arises from the missing information $Y_{n+1:\infty}$. The predictive view of Bayesian inference partitions posterior sampling into two distinct tasks. The first is the simulation of $Y_{n+1:\infty}$ through the sequence of 1-step ahead predictive distributions to assess the uncertainty that arises from the missing observables. The second is the recovery of the parameter of interest Θ from the simulated complete information, which is facilitated by the limiting posterior mean point estimate $\bar{\theta}_\infty$. The uncertainty in Θ then flows from the uncertainty in $Y_{n+1:\infty}$. Inspired by this, we will now demonstrate the practical importance of this interpretation by introducing a predictive framework for inference built exactly on these two tasks. This framework eliminates the need for the usual likelihood-prior construction of the

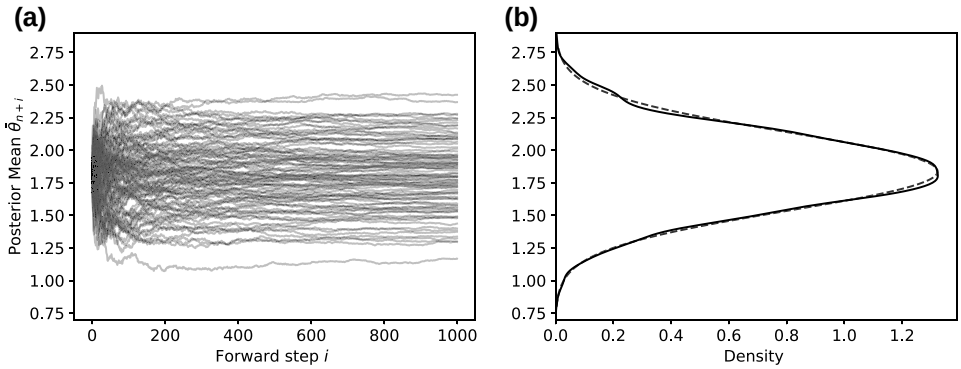

Figure 1. (a) Sample paths of $\bar{\theta}_{n+i}$ through forward sampling; (b) Kernel density estimate of $\bar{\theta}_N$ samples (——) and analytical posterior density $\pi(\theta \mid y_{1:n})$ (- - -).

Bayesian model, and as such generalizes the traditional Bayesian posterior to the martingale posterior.

2.2.1 *Sampling the missing data*

For the predictive Bayesian, the role of the posterior $\pi(\theta \mid y_{1:n})$ is to aid in the updating of the predictive density, $p(\cdot \mid y_{1:N-1}) \mapsto p(\cdot \mid y_{1:N})$ after observing Y_N, and the likelihood and prior can be viewed as merely intermediate tools to construct the sequence of predictives (Roberts, 1965). To obviate the need of a likelihood-prior specification, our proposal is to specify the sequence of 1-step ahead predictive densities $\{p(\cdot \mid y_{1:N})\}_{N \geq n}$ directly, which implies a joint density through the factorization

$$p(y_{n+1:N} \mid y_{1:n}) = \prod_{i=n+1}^{N} p(y_i \mid y_{1:i-1}). \tag{2.3}$$

However, we must take care in our elicitation of $\{p(\cdot \mid y_{1:N})\}_{N \geq n}$ to ensure the existence of the limit θ_∞. As this is technical, we defer a formal discussion of this choice and the conditions required to Section 3. For now, we point out that a sufficient condition is for the 1-step ahead predictive densities to satisfy a martingale condition similar to that of Doob, with details given in Section 3.2. It may seem that constructing this sequence will incur too much complexity, but we will show this is in fact feasible and desirable. One key idea is to utilize a general sequential updating procedure whereby given an observed $Y_N = y_N$, we have a direct and tractable iterative update $\{p(\cdot \mid y_{1:N-1}), y_N\} \mapsto p(\cdot \mid y_{1:N})$.

2.2.2 *Recovering the quantity of interest*

We now discuss the second task: given a sample $Y_{n+1:\infty}$, we require a procedure to recover the quantity of interest. In a traditional parametric Bayesian model, the quantity of interest is usually the unknown parameter θ that indexes the sampling density, and as shown by Doob, the limiting posterior mean $\bar{\theta}_\infty$ serves this purpose. A more general framework is the decision task discussed in Bissiri et al. (2016), where the aim is to minimize a functional of an unknown distribution function F_0 from which samples $Y_{1:n}$ are i.i.d.. For some loss function $\ell(\theta, y)$, the quantity of interest θ is now defined as

$$\theta_0 = \arg\min_{\theta} \int \ell(\theta, y)\, \mathrm{d}F_0(y). \tag{2.4}$$

More details can be found, for example, in Huber (2004) and Bissiri et al. (2016). Typical examples are $\ell(\theta, y) = |\theta - y|$ for the median, $\ell(\theta, y) = (\theta - y)^2$ for the mean, and $\ell(\theta, y) = -\log f_\theta(y)$ for the Kullback–Leibler minimizer between some parametric density f_θ and the sampling density f_0. The choice of the negative log-likelihood is also interesting as it allows us to target the parameters of a parametric model without the assumption that the model is well-specified (Bissiri et al., 2016; Walker, 2013). While misspecification under our framework is still an open question, the Bayesian bootstrap has particularly desirable theoretical and practical properties under misspecification (Fong et al., 2019; Lyddon et al., 2018, 2019). We will also consider more general forms of θ_0, e.g., the density of F_0.

Working now in the space of probability distributions, the traditional Bayesian approach would be to elicit a prior on F, perhaps nonparametric, and derive the posterior $\Pi(\mathrm{d}F \mid y_{1:n})$. Here, F represents the Bayesian's subjective belief in the unknown true F_0. A posterior sample of θ is then obtained as follows: draw $F \sim \Pi(\mathrm{d}F \mid y_{1:n})$ and compute the θ minimizing $\int \ell(\theta, y)\, \mathrm{d}F(y)$. For our generalization beyond the likelihood-prior construction, we do not have a posterior mean nor a posterior F and thus require an alternative to recover the quantity of interest given a sample of $Y_{n+1:\infty}$ conditioned on $y_{1:n}$. Our proposal is to construct the random limiting empirical distribution function

$$F_\infty(y) = \lim_{N \to \infty} \frac{1}{N} \left\{ \sum_{i=1}^{n} \mathbb{1}(y_i \leq y) + \sum_{i=n+1}^{N} \mathbb{1}(Y_i \leq y) \right\}$$

and take θ to minimize $\int \ell(\theta, y)\, \mathrm{d}F_\infty(y)$. Here, our F_∞ takes the place of the posterior draw of F, and its existence will rely on the martingale condition as mentioned above. We can write θ_∞, $\theta(F_\infty)$ or $\theta(Y_{1:\infty})$ interchangeably for the parameter of interest computed from the completed information. If we specify $p(\cdot \mid y_{1:n})$ through the usual likelihood-prior construction, then sampling F from the posterior, in fact, yields the same random distribution function as F_∞ almost surely; this theoretical justification for the limiting empirical distribution function F_∞ is in Online Supplementary Material, Appendix C.2.

2.3 The martingale posterior

Our framework for predictive inference is summarized as follows. Suppose we observe $Y_{1:n}$ i.i.d. from some unknown F_0 and are interested in the θ_0 defined by (2.4). We specify a sequence of predictive densities $\{p(\cdot \mid y_{1:n})\}_{n\geq 0}$ which satisfies the martingale condition to be discussed in Section 3.2 and implies a joint distribution through (2.3). We then impute an infinite future dataset through

$$Y_{n+1} \sim p(\cdot \mid y_{1:n}), \quad Y_{n+2} \sim p(\cdot \mid y_{1:n+1}), \ \ldots, Y_N \sim p(\cdot \mid y_{1:N-1})$$

for $N \to \infty$. Given the infinite random dataset $Y_{n+1:\infty}$ and the corresponding empirical distribution function F_∞, we compute $\theta_\infty = \theta(F_\infty)$. We designate the distribution of θ_∞ as the martingale posterior, where we use the notation Π_∞ for comparability to traditional Bayes.

Definition 1 (Martingale posterior). The martingale posterior distribution is defined as

$$\Pi_\infty(\theta_\infty \in A \mid y_{1:n}) = \int \mathbb{1}\{\theta(F_\infty) \in A\}\, \mathrm{d}\Pi(F_\infty \mid y_{1:n}), \tag{2.5}$$

for measurable set A, which is a subset of the parameter space.

Drawing samples of θ_∞ from the martingale posterior involves repeating the above simulation procedure given above. We refer to this Monte Carlo scheme as predictive resampling, which has strong connections with the Bayesian bootstrap of Rubin (1981), as we will see in Section 2.4. In practice, however, we may be unable to simulate $N \to \infty$, or the study population may be of finite size N. In this case, we can instead impute $Y_{n+1:N}$ for finite N, giving us the analogous empirical distribution function F_N and parameter $\theta_N = \theta(F_N)$ or $\theta(Y_{1:N})$.

Definition 2 (Finite martingale posterior). The finite martingale posterior is similarly defined as

$$\Pi_N(\theta_N \in A \mid y_{1:n}) = \int \mathbb{1}\{\theta(y_{1:N}) \in A\} p(y_{n+1:N} \mid y_{1:n})\, \mathrm{d}y_{n+1:N}.$$

In the finite form, the role of the two constituent elements, $p(y_{n+1:N} \mid y_{1:n})$ and $\theta(y_{1:N})$, is even clearer. For infinite populations, we also highlight that the value of θ_N varies around θ_∞, but this may be negligible for sufficiently large N. If the population is actually finite and of size N, then θ_N would be the actual target and thus not an approximation. Finally, we reiterate that the martingale posterior (2.5) is equivalent to the traditional Bayesian posterior when using (1.2) as the predictive. A summary of the notation and an illustration of the imputation scheme is provided, respectively, in Online Supplementary Material, Appendices A and B.

2.4 The Bayesian bootstrap

The resemblance of the martingale posterior to a bootstrap estimator should not have gone unnoticed, as both involve repeated sampling of observables followed by computing estimates from the sampled dataset. The Bayesian bootstrap of Rubin (1981) is often described as the Bayesian version of the frequentist bootstrap. After observing $y_{1:n}$, one draws a random

distribution function from the posterior through

$$w_{1:n} \sim \text{Dirichlet}(1, \ldots, 1), \quad F(y) = \sum_{i=1}^{n} w_i \mathbb{1}(y_i \leq y).$$

A posterior sample of the statistic of interest can then be computed as $\theta(F)$. One interpretation of the Dirichlet weights is to generate uncertainty through the randomization of the objective function (Jin et al., 2001; Newton et al., 2020; Newton & Raftery, 1994; Ng & Newton, 2022). Closer to our perspective are the connections to BNP inference, which have been explored in much detail within the literature as it is the non-informative limit of a posterior Dirichlet process (Ghosal & van der Vaart, 2017; Lo, 1987; Muliere & Secchi, 1996). Recent work has exploited the computational advantages of the Bayesian bootstrap for scalable nonparametric inference; see Saarela et al. (2015), Lyddon et al. (2018), Fong et al. (2019), Newton et al. (2020), Knoblauch and Vomfell (2020), and Nie and Ročková (2023).

2.4.1 *The empirical predictive*

Within the framework of martingale posteriors, the Bayesian bootstrap has a particularly elegant interpretation that follows from the equivalence to the Pólya urn scheme (Blackwell & MacQueen, 1973; Lo, 1988). The Bayesian bootstrap is equivalent to the martingale posterior if we define our sequence of predictive probability distribution functions to be the sequence of empirical distribution functions, that is

$$P(Y_{n+1} \leq y \mid y_{1:n}) = F_n(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(y_i \leq y). \tag{2.6}$$

This is easy to see as sampling $Y_{n+1} \sim F_n(y)$ amounts to drawing with replacement 1 of n colours with probability $1/n$ from the urn, and updating to $F_{n+1}(y)$ is equivalent to reinforcing the urn, that is

$$F_{n+1}(y) = \frac{n}{n+1} F_n(y) + \frac{1}{n+1} \mathbb{1}(y_{n+1} \leq y).$$

Continuing on to ∞, the proportions of colours converge in distribution to the Dirichlet distribution. Interestingly, this choice of predictive implies an exchangeable future sequence from the connection to the Dirichlet process. The atomic support of the predictive is however slightly problematic if F_0 is continuous, as any new observations from F_0 will be assigned a predictive probability of zero; we will introduce a methodology that remedies this in Section 4. Generalizations to other atomic predictives can, for example, be found in Zabell (1982) and Muliere et al. (2000).

One can consider the empirical distribution function as the simplest nonparametric predictive for i.i.d. data and can thus regard the Bayesian bootstrap as the simplest BNP model. The uncertainty from the Bayesian bootstrap arises not from the random weights, but from the sequence of empirical predictive distributions. We resample with replacement, treating each resampled point as a new observed datum; this fundamental observation is our motivation for the term predictive resampling.

2.4.2 *Comparison to the frequentist bootstrap*

Throughout this section, we have assumed the existence of an underlying F_0 from which $Y_{1:n}$ are i.i.d., which, in turn, implies the existence of an unknown true θ_0 much like the frequentist. This has some connections to frequentist consistency under our framework, which we discuss in Section 6.3. The posterior random variable θ_∞ then represents our subjective uncertainty in θ_0 after observing $Y_{1:n} = y_{1:n}$. The Bayesian bootstrap and Efron's bootstrap (Efron, 1979) are then ideal vessels for the contrasting of Bayesian and frequentist uncertainty. Both methods are nonparametric and begin by constructing the empirical predictive F_n as in (2.6) from the atoms of $y_{1:n}$ as an

estimate of F_0, and both involve resampling. The key difference lies in how the resampling is carried out.

The frequentist draws a dataset of size n i.i.d. from F_n, which we write as $Y^*_{1:n}$ with corresponding empirical distribution function F^*_n, and computes $\theta(F^*_n)$ as a random sample of the estimator. The Bayesian, on the other hand, draws an infinite future dataset $Y_{n+1:\infty}$ through predictive resampling and computes $\theta(F_\infty)$ as a random sample of the estimand, where F_∞ is the limiting empirical distribution function of $\{y_{1:n}, Y_{n+1:\infty}\}$, noting again that the Bayesian holds $y_{1:n}$ fixed. This is summarized in Algorithms 1 and 2. Notably, the specification in both bootstraps are equivalent: it is merely the elicitation of $F_n(y)$, which entirely characterizes both types of uncertainty.

Algorithm 1: Bayesian bootstrap

Set F_n from the observed data $y_{1:n}$
for $j \leftarrow 1$ **to** B **do**
  for $i \leftarrow n+1$ **to** ∞ **do**
    Sample $Y_i \sim F_{i-1}$
    Update $F_i \leftarrow \{F_{i-1}, Y_i\}$
  end
  Compute F_∞ from $\{y_{1:n}, Y_{n+1:\infty}\}$
  Evaluate $\theta^{(j)}_\infty = \theta(F_\infty)$
end
Return $\{\theta^{(1)}_\infty, \ldots, \theta^{(B)}_\infty\}$

Algorithm 2: Efrons bootstrap

Set F_n from the observed data $y_{1:n}$
for $j \leftarrow 1$ **to** B **do**
  for $i \leftarrow 1$ **to** n **do**
    Sample $Y^*_i \sim F_n$
    No update to F_n
  end
  Compute F^*_n from $\{Y^*_{1:n}\}$
  Evaluate $\theta^{(j)}_n = \theta(F^*_n)$
end
Return $\{\theta^{(1)}_n, \ldots, \theta^{(B)}_n\}$

2.5 Related work

There have been many others that shared de Finetti's view on the emphasis on observables for inference. The work of Dawid (1984, 1992a, 1992b) on prequential statistics, a portmanteau of probability/predictive and sequential, is one such example. In his work, Dawid focuses on the importance of forecasting and introduces a statistical methodology that assigns predictive probabilities and assesses these methods on their agreement with the observed data. In particular, Dawid (1984) recommends eliciting a sequence of 1-step ahead predictive distributions as we do but motivates this by arguing that forecasting is the main statistical task. As pointed out in Section 1, this is in contrast to our case where parameter inference is the main task of interest and the sequence of predictives is mainly a convenient tool to construct the joint predictive on future observations. We will see in Section 3.2 that stricter conditions are required on this sequence of predictives for inference. Another strong proponent of the predictive approach is the work of Geisser: he believed that the prediction of observables was of much greater importance than the estimation of parameters, which he described as 'artificial constructs' (Geisser, 1975). His emphasis is on the predictive motivated cross-validation (Geisser, 1974; Stone, 1974), which is now popular for Bayesian model

evaluation (Gelman et al., 2014; Vehtari & Lampinen, 2002). Works such as Dawid (1985) and Lauritzen (1988) also consider parameters as functions of the infinite sequence of observations using the notion of repetitive structures. Finally, the work of Rubin on both the potential outcomes model (Rubin, 1974) and multiple imputation (Rubin, 2004) highlights the idea of inference via imputation.

An early application of what is essentially finite predictive resampling and martingale posteriors is Bayesian inference for finite populations, first discussed in Roberts (1965) and Ericson (1969) and later by Geisser (1982, 1983). A finite population Bayesian bootstrap is described in Lo (1988), in which a finite Pólya urn is used to simulate from the posterior. The 'Pólya posterior' of Ghosh and Meeden (1997) uses the same approach following an admissibility argument. These methods have applications in survey sampling or the interim monitoring of clinical trials (Saville et al., 2014).

There have been recent exciting directions of work that investigate the predictive view of BNP. Fortini et al. (2000) investigate under what conditions parametric models arise from the sequence of predictives using the concept of predictive sufficiency and derive conditions such that the joint distribution is exchangeable. Fortini and Petrone (2012, 2014) discuss the construction of a range of popular exchangeable BNP priors through a sequence of predictive distributions, motivated through a predictive de Finetti's representation theorem (Fortini & Petrone, 2012, Theorem 2). Berti et al. (2020) then generalize the nonparametric approach to c.i.d. sequences; we will later see that c.i.d. sequences, as introduced in Berti et al. (2004), play a crucial role in our work. However, the previously described methods are mostly constrained to the discrete case. Hahn (2015) and Hahn et al. (2018) construct c.i.d. models through a predictive sequence for univariate density estimation, respectively, utilizing the kernel density estimator and the bivariate copula. Hahn (2015) also discusses the connection of Bayesian uncertainty and prediction with a weaker argument and gives a similar example to our Example 1. Predictive resampling is then used to sample nonparametric densities from a finite martingale posterior; however, Hahn (2015) instead specifies the predictive distribution P_N for large N and works backwards to find the sequence of predictives. Fortini and Petrone (2020) analyse the predictive recursion algorithm of Newton et al. (1998) and the implied underlying quasi-Bayesian model. In their work, they carry out predictive resampling to simulate from the prior law of the mixing distribution in an example and obtain its asymptotic distribution under the c.i.d. model, that is, an asymptotic approximation to the martingale posterior. An interesting aside is the recent work of Waudby-Smith and Ramdas (2023) which utilizes adaptive betting with martingale conditions for the purpose of constructing frequentist confidence intervals. We aim to unify these related strands of research under a single framework.

3 Predictive resampling for martingale posteriors

For the martingale posterior, we now embark on the task of eliciting the general 1-step ahead predictive distributions, with the traditional Bayesian posterior predictive as a special case. For notational convenience, we write the sequence of predictive probability distribution functions estimated after observing $Y_{1:i} = y_{1:i}$ as

$$P_i(y) := P(Y_{i+1} \leq y \mid y_{1:i}), \quad i \in \{1, 2, \ldots\} \tag{3.1}$$

which may have corresponding density functions $p_i(y)$. The subscript indicates the length of the conditioning sequence, and there may be a $P_0(y)$ as some initial choice. For a general sequence of predictives, where exchangeability no longer necessarily holds, we instead define our joint distribution on $y_{1:N}$ through this sequence of 1-step ahead predictives and the chain rule as in (2.3). The Ionescu-Tulcea theorem (Kallenberg, 1997, Theorem 5.17) guarantees the existence of such a joint distribution as we take $N \to \infty$, which has been pointed out by works such as Dawid (1984), Fortini and Petrone (2012), and Berti et al. (2020).

Beyond the traditional Bayesian posterior predictive, there is good justification for specifying the model with 1-step ahead predictives, instead of, say m-step ahead. It is simple to interpret and estimate a 1-step ahead predictive as the decision-maker's best estimate of the unknown sampling distribution function F_0, and methods such as maximum likelihood estimation already do

this. Finally, we will see that a 1-step update of the predictive allows for the enforcing of the c.i.d. condition for predictive coherence.

While the prescription of (3.1) remains a subjective task, we find it to be no more subjective than the selection of a sampling density. There is no longer a need to elicit subjective distributions on parameters which merely index the sampling distribution with no physical meaning, which has been described as 'intrinsic' (Dawid, 1985). In nonparametric inference, we also do not need to elicit priors directly on the space of probability distributions, which can be cumbersome. The uncertainty arises simply from the elicitation of (3.1). It is clear that we can still use external information and subjective judgement not provided by the data $y_{1:n}$ in this construction.

3.1 A practical algorithm for uncertainty

Given the model specification (3.1), suppose we wish to undertake inference on a statistic of interest $\theta(F_0)$, defined through a loss function $\ell(\theta, y)$ as in (2.4). We can obtain finite martingale posterior samples through predictive resampling given in Algorithm 3, noting the similarity to the Bayesian bootstrap algorithm.

Algorithm 3: Predictive resampling

Compute P_n from the observed data $y_{1:n}$
$N > n$ is a large integer
for $j \leftarrow 1$ **to** B **do**
  for $i \leftarrow n+1$ **to** N **do**
    Sample $Y_i \sim P_{i-1}$
    Update $P_i \leftarrow \{P_{i-1}, Y_i\}$
  end
  Compute F_N from $\{y_{1:n}, Y_{n+1:N}\}$
  Evaluate $\theta_N^{(j)} = \theta(F_N)$ or $\theta_N^{(j)} = \theta(P_N)$
end
Return $\{\theta_N^{(1)}, \ldots, \theta_N^{(B)}\} \overset{\text{iid}}{\sim} \Pi_N(\cdot \mid y_{1:n})$

In summary, we run a forward simulation starting at $P_n(y)$ by consecutively sampling from the 1-step ahead predictives and updating as we go. For large N, we now have a random dataset $\{y_{1:n}, Y_{n+1:N}\}$ from which we can compute the empirical distribution function $F_N(y)$ and statistic of interest $\theta(F_N)$. In particular, when the sequence of predictives takes on the form (1.2), combined with the self-information loss, $-\log f_\theta(y)$, is this procedure equivalent to traditional Bayesian inference.

The empirical distribution is atomic, which may be problematic if the object of interest θ_0 requires the limiting F_∞ to be continuous, for example, if θ_0 is the probability density of F_0 or a tail probability. In this case, we can instead compute $\theta(P_N)$, where P_N is the random predictive distribution function conditioned on $\{y_{1:n}, Y_{n+1:N}\}$, which would typically be continuous. We can regard P_N as the finite approximation to the limiting predictive distribution function $P_\infty := \lim_{N\to\infty} P_N$, which serves the same purpose as the limiting empirical F_∞ in Section 2.2.2. In fact, P_∞ and F_∞ coincide for traditional Bayesian models, and even for the more general c.i.d. sequence of predictives that we will consider shortly. We discuss this in Online Supplementary Material, Appendix C, borrowing results from Doob (1949), Berti et al. (2004), and Lijoi et al. (2004).

Some experimental and theoretical guidance for selecting a sufficiently large N to estimate P_∞ is given in Sections 5 and 6. However, it is also interesting to consider a finite population, where the F_0 of interest is indeed the empirical distribution function of a population of size N, as discussed in Sections 2.3 and 2.5. In this case, truncating predictive resampling at N indeed returns the correct uncertainty in any parameter of interest $\theta(Y_{1:N})$ of the finite population.

3.2 Predictive coherence and conditionally identically distributed sequences

The notion of coherence in one's belief on the parameter θ is key to the subjective Bayesian, where coherence may be defined in a decision-theoretic sense (Bernardo & Smith, 2009, Chapter 2.3) or

through Dutch book arguments (e.g., Heath & Sudderth, 1978). Extensions of coherence to forecasting are given in Lane and Sudderth (1984), Berti et al. (1998), and more examples of coherence in general can be found in Robins and Wasserman (2000), Eaton and Freedman (2004). More recently, the notion of coherence of belief updating was introduced in Bissiri et al. (2016), where a belief update on a statistic of interest θ is coherent if the update is equivalent whether computed sequentially with y_1 followed by y_2 or with $\{y_1, y_2\}$ in tandem through an additive loss condition. In bypassing the traditional likelihood-prior construction, we must forsake the usual coherence of belief updating and exchangeability. Instead, we specify conditions for a valid martingale posterior entirely in terms of the predictive distribution function, which we term *predictive coherence*.

Suppose we observe $Y_{1:n}$ i.i.d. from some F_0 and construct $P_n(y)$ as in (3.1). We can then view the predictive machine $P_n(y)$ as the best estimate of the unknown distribution function F_0 from which the data arose, incorporating all observed data and any possible subjective knowledge. The first minimal condition is that the sequence of predictive distribution functions $P_{n+1}(y)$, $P_{n+2}(y)$… converges to a random distribution function. Secondly, we would ensure that predictive resampling does not introduce any new information or bias, as P_n is already our best summary of the observed $y_{1:n}$, and the procedure should merely return uncertainty. Formally, we write these conditions, respectively, as follows:

Condition 1 (Existence). The sequence $P_{n+1}(y)$, $P_{n+2}(y)$, … converges to a random $P_\infty(y)$ almost surely for each $y \in \mathbb{R}$, where P_∞ is a random probability distribution function.

Condition 2 (Unbiasedness). The posterior expectation of the random distribution function satisfies

$$E[P_\infty(y) \mid y_{1:n}] = P_n(y)$$

almost surely for each $y \in \mathbb{R}$.

Under Condition 1, P_∞ is defined through the sequence of predictives, and we can thus treat P_∞ directly as the random distribution function without the need for an underlying Bayes' rule representation. This, in turn, gives us the posterior uncertainty in any statistic $\theta(P_\infty)$. Condition 2 is stricter and implies that P_n is our best estimate of F_0 and is equal to the posterior mean.

Fortunately, Conditions 1 and 2 are satisfied if the sequence Y_{n+1}, Y_{n+2}, … is *conditionally identically distributed* (c.i.d.), as introduced and studied in Berti et al. (2004). Many useful properties of c.i.d. sequences have been shown in their work, which we now summarize. The sequence Y_{n+1}, Y_{n+2}, … is c.i.d if we have

$$P(Y_{i+k} \leq y \mid y_{1:i}) = P_i(y), \quad \forall\, k > 0$$

almost surely for each $y \in \mathbb{R}$. This states that conditional on $y_{1:i}$, any future data points will be identically distributed according to the predictive P_i. This predictive invariance is particularly natural as a minimal predictive coherence condition and serves as an analogue to de Finetti's exchangeability assumption in the predictive framework. In fact, as shown in Kallenberg (1988), the c.i.d. condition is a weakening of exchangeability, and Berti et al. (2004) also show that c.i.d. sequences are asymptotically exchangeable, which we state formally in Theorem 3 in Section 6.1.

An equivalent formulation of c.i.d. sequences which connects closely to the predictive coherency conditions is that $P_i(y)$ is a martingale for $i \in \{n+1, n+2, \ldots\}$, that is

$$E[P_i(y) \mid y_{1:i-1}] \equiv \int P_i(y)\, \mathrm{d}P_{i-1}(y_i) = P_{i-1}(y) \tag{3.2}$$

almost surely for each $y \in \mathbb{R}$, noting that P_i depends on y_i as in (3.1). Relying again on Doob's martingale convergence theorem (Doob, 1953), the sequence $P_n(y)$, $P_{n+1}(y)$, … converges to $P_\infty(y)$ almost surely for each $y \in \mathbb{R}$, and P_∞ can be shown to be a random probability distribution function (Berti

et al., 2004); we state this formally in Theorem 4 in Section 6.1. In this case, we also designate the distribution of P_∞ as the martingale posterior when we do not specify θ_∞. Condition 2 is then satisfied as the sequence $P_{n+1}(y)$, $P_{n+2}(y)$, … is uniformly integrable. Furthermore, we are guaranteed the existence of the limiting empirical distribution function F_∞ as required in Section 2.2.2, and in fact $F_\infty(y) = P_\infty(y)$ almost surely so the interchangeability of $\theta(F_\infty)$ and $\theta(P_\infty)$ is justified. This equivalence, as well as the convergence of $\theta(Y_{1:N})$ with N for a certain class of parameters, is discussed in Online Supplementary Material, Appendix C.1. Although not explored here, connections of the c.i.d. property to other notions of coherence, such as those given at the start of this subsection, would be interesting to investigate especially given the absence of the prior distribution.

Although the above predictive coherence conditions are for a valid martingale posterior, we still need to specify a sequence of predictive distributions. Clearly, the traditional Bayesian posterior predictive satisfies the above conditions, but in the interest of computational expediency or the desire to bypass the likelihood-prior construction, we may wish to consider more general predictive machines. The remainder of this paper will consider recursive predictive densities using bivariate copulas.

4 Recursive predictives with bivariate copulas

In this section, we focus primarily on the elicitation of the sequence of predictives (3.1) in the continuous case, where $p_i(y)$ is the density of $P_i(y)$ in (3.1). Analogous predictives are derivable for the discrete case, and these are obtained in Berti et al. (2020). In particular, we investigate the prescription of this sequence of predictives through a recursive manner, that is for $i \in \{0, 1, \ldots\}$

$$p_{i+1}(y) = \psi^{\rho}_{i+1}\{p_i(y), y_{i+1}\}$$

where ψ^{ρ}_i is a sequence of update functions, possibly parameterized by a hyperparameter ρ. In this case, we require an initial guess $p_0(y)$ for our recursion, which plays the role of a prior guess on f_0. A recursive update of this form is not necessary for a martingale posterior, but it allows for simple satisfaction of conditions for predictive coherence, as discussed in Section 3.2, and computations for predictive resampling will also be significantly easier. Furthermore, when one is only interested in estimating $p_n(y)$, recursive updates may have computational advantages as one does not need to explicitly estimate the posterior.

Recursive updates have previously been motivated as a fast alternative to MCMC in Dirichlet process mixture models (DPMM). The predictive recursion algorithm was first introduced by Newton et al. (1998), which estimates the mixing distribution through a recursive update, and its properties have been studied in detail in the literature; see Martin (2021) for a thorough review. One interesting property shown in Fortini and Petrone (2020) is that the sequence of observables in Newton's algorithm is c.i.d.; however, the computation of the predictive densities is intractable and requires numerical integration, so we will not discuss this method further here. Direct recursive updates for the predictive density were then introduced in Hahn (2015), Hahn et al. (2018), Berti et al. (2020), all of which satisfy the c.i.d. condition. The bivariate copula method of Hahn et al. (2018) is particularly tractable and well motivated, and we will now build on this method in this section.

4.1 Bivariate copula update

To satisfy the c.i.d. condition required for predictive coherence, we can extend the martingale condition to hold for the sequence of densities p_n, p_{n+1}, … such that for $i \in \{n+1, n+2, \ldots\}$

$$E[p_i(y) \mid y_{1:i-1}] \equiv \int p_i(y) p_{i-1}(y_i)\, \mathrm{d}y_i = p_{i-1}(y) \tag{4.1}$$

for each $y \in \mathbb{R}$, assuming the expectations exist. We highlight again that p_i depends on y_i as it is the density of (3.1). The above is a sufficient condition for (3.2) to hold, so our sequence is c.i.d. and the existence and unbiasedness conditions are satisfied giving us a valid martingale posterior. In fact, the martingale convergence theorem shows that $p_i(y) \to p_\infty(y)$ almost surely for each $y \in \mathbb{R}$, but more assumptions are needed to show that p_∞ is the density of $P_\infty(y)$; we explore this in Theorem 5 in Section 6.1.

One particular tractable form of update rule ψ_i^ρ that satisfies (3.2) is the bivariate copula (Nelsen, 2007) update interpretation of Bayesian inference first introduced in Hahn et al. (2018) for univariate data. A bivariate copula is a bivariate cumulative distribution function $C:[0, 1]^2 \to [0, 1]$ with uniform marginal distributions, and in the cases, we consider it will have a probability density function $c: [0, 1]^2 \to \mathbb{R}$. The bivariate copula can be regarded as characterizing the dependence between two random variables independent of their marginals, which can be seen through Sklar's theorem in the bivariate case.

Theorem 2 (Sklar (1959)). For a bivariate cumulative distribution function $F(y_1, y_2)$ with continuous marginals $F_1(y_1)$, $F_2(y_2)$, there exists a unique bivariate copula C such that

$$F(y_1, y_2) = C\{F_1(y_1), F_2(y_2)\}.$$

Furthermore, if F has a density f with marginal densities f_1, f_2, we can write

$$f(y_1, y_2) = c\{F_1(y_1), F_2(y_2)\}f_1(y_1)f_2(y_2)$$

where c is the density of C.

This holds for higher dimensions, but we state it for $d = 2$ as this is what we will be working with. From this, we can see that the bivariate copula can model the dependence structure between consecutive predictive densities, and thus we have the following corollary, with the proof given in Online Supplementary Material, Appendix D.1.

Corollary 1 The sequence of conditional densities $p_0, p_1, \ldots$ satisfies the martingale condition (4.1) if and only if there exists a unique sequence of bivariate copula densities $c_1, c_2, \ldots$ such that

$$p_{i+1}(y) = c_{i+1}\{P_i(y), P_i(y_{i+1})\}p_i(y) \tag{4.2}$$

for $i \in \{0, 1, \ldots\}$ and P_i is the distribution function of p_i.

In the univariate case, we can thus elicit a c.i.d. model through a sequence of copulas, that is we have (4.2) as our update function ψ_{i+1}^ρ. We highlight that c_{i+1} is the bivariate copula density that models the dependence between $\{Y_{i+1}, Y_{i+2}\}$ conditioned on $Y_{1:i}$. Although the sequence c_{i+1} can technically depend arbitrarily on $y_{1:i}$ (and the sample size $i + 1$) without violating the martingale condition, we will later constrain this dependence. As all exchangeable Bayesian models are c.i.d., there exists a unique sequence of copulas which may or may not be tractable that characterize the model (Hahn et al., 2018). This sequence takes on exactly the form

$$p_{i+1}(y) = \underbrace{\frac{\int f_\theta(y)f_\theta(y_{i+1})\pi(\theta \mid y_{1:i})\,\mathrm{d}\theta}{p_i(y)p_i(y_{i+1})}}_{c_{i+1}\{P_i(y),P_i(y_{i+1})\}} p_i(y). \tag{4.3}$$

The copula density arises following Theorem 2 as the numerator in (4.3) is the joint density $p_i(y, y_{i+1})$ with marginal densities $p_i(y)$ and $p_i(y_{i+1})$. Instead of specifying the sampling distribution and prior, we will now consider the specification of the sequence of copulas c_i directly. The form for c_i inspired by the DPMM is particularly attractive and serves well as the canonical extension of the Bayesian bootstrap predictive to continuous random variables. In the remainder of this section, we will first review the method of Hahn et al. (2018) for univariate density estimation and extend the methodology to include predictive resampling and hyperparameter selection. We then introduce analogous copula updates for more advanced data settings, including multivariate density estimation, regression and classification.

4.2 Univariate case

Tractable forms of this sequence of copulas in Bayesian models are investigated in Hahn et al. (2018), which correspond to conjugate priors. The update of particular interest is that of the DPMM (Escobar & West, 1995) of the particular form

$$f_G(y) = \int \mathcal{N}(y \mid \theta, 1)\, \mathrm{d}G(\theta), \quad G \sim \mathrm{DP}(a, G_0), \quad G_0 = \mathcal{N}(\theta \mid 0, \tau^{-1}),$$

where $a > 0$ is the scalar precision parameter that we set to $a = 1$. The model is nonparametric, making it a strong candidate for a predictive update, but only the copula update for $i = 0$ is tractable. Inspired by this first update step, Hahn et al. (2018) suggest that the general update to compute the density $p_i(y)$ after observing $y_{1:i}$ for $i \in \{0, \ldots, n-1\}$ takes on the form

$$\begin{aligned} p_{i+1}(y) &= (1 - \alpha_{i+1}) p_i(y) + \alpha_{i+1} c_\rho\{P_i(y), P_i(y_{i+1})\} p_i(y) \\ P_{i+1}(y) &= (1 - \alpha_{i+1}) P_i(y) + \alpha_{i+1} H_\rho\{P_i(y), P_i(y_{i+1})\} \end{aligned} \tag{4.4}$$

where $P_i(y)$ is the distribution function of $p_i(y)$. Here $c_\rho(u, v)$ is the bivariate Gaussian copula density and $H_\rho(u, v)$ is the conditional Gaussian copula of the forms:

$$c_\rho(u, v) = \frac{\mathcal{N}_2\{\Phi^{-1}(u), \Phi^{-1}(v) \mid 0, 1, \rho\}}{\mathcal{N}\{\Phi^{-1}(u) \mid 0, 1\}\mathcal{N}\{\Phi^{-1}(v) \mid 0, 1\}}, \quad H_\rho(u, v) = \Phi\left\{\frac{\Phi^{-1}(u) - \rho\Phi^{-1}(v)}{\sqrt{1 - \rho^2}}\right\} \tag{4.5}$$

where Φ^{-1} is the standard inverse normal distribution function and $\mathcal{N}_2$ is the standard bivariate density with correlation $\rho \in (0, 1)$. The role of ρ as a bandwidth will be explored shortly. The update (4.4) is then a mixture of the independent copula density and the Gaussian copula density, and the sequence $\alpha_i = \mathcal{O}(i^{-1})$ ensures the update approaches the independent copula as $i \to \infty$. Although asymptotic independence is not necessary for the martingale condition, this property holds for Bayesian sequences of copulas (Hahn et al., 2018) and is indeed important for frequentist consistency when estimating p_n as we will see in Section 6.3. We will see the specific suggested form of α_i at the end of this section.

Note the similarity of the update in (4.4) to the generalized Pólya urn for the Dirichlet process, which for $c = 1$ has the update $P_{i+1}(y) = (1 - \alpha_{i+1})P_i(y) + \alpha_{i+1}\mathbb{1}(y_{i+1} \le y)$. We can thus interpret (4.4) as a smooth generalization of the Bayesian bootstrap update for continuous distributions. One can also interpret (4.4) as a Bayesian kernel density estimate (KDE) that satisfies the c.i.d. condition, as the regular KDE cannot satisfy this condition (West, 1991). The update can be visualized in Figure 2, where for convenience we write $u_i = P_i(y)$, $v_i = P_i(y_{i+1})$. The Gaussian copula kernel $c_\rho(u_i, v_i)p_i(y)$ is a data-dependent kernel roughly centred at y_{i+1}, as shown in the left. The kernel becomes sharper as ρ increases, and we recover the Bayesian bootstrap in the limit of $\rho \to 1$ (with $\alpha_i = 1/i$). The update is then a mixture of $p_i(y)$ and the copula kernel, which gives us $p_{i+1}(y)$ in the right panel.

The recursive update was first introduced to compute $p_n(y)$, but properties of the update make it a highly suitable candidate for predictive resampling. Firstly, by Corollary 1, this update is guaranteed to provide a c.i.d. sequence and hence satisfy the existence and unbiasedness conditions. Secondly, the update of the predictive distribution is online and does not require an expensive recomputation of the predictive distribution at each step. Finally, the predictive resampling update is particularly computationally elegant as $y_{i+1} \sim P_i(y)$ implies that $P_i(y_{i+1}) \sim \mathcal{U}[0, 1]$, so all that is required is the simulation of uniform random variables. The forward sampling step then involves simulating $V_i \sim \mathcal{U}[0, 1]$ and computing

$$\begin{aligned} p_{i+1}(y) &= [1 - \alpha_{i+1} + \alpha_{i+1} c_\rho\{P_i(y), V_i\}] p_i(y) \\ P_{i+1}(y) &= (1 - \alpha_{i+1}) P_i(y) + \alpha_{i+1} H_\rho\{P_i(y), V_i\} \end{aligned}$$

iterated over $i \in \{n, \ldots, N\}$, which gives us a random $p_N(y)$ at the end. There is no need to actually sample $Y_{i+1} \sim P_i(y)$, which is possible but is more computationally expensive. In Section 6, we will see that this update form allows easy analysis of the theoretical properties of predictive resampling.

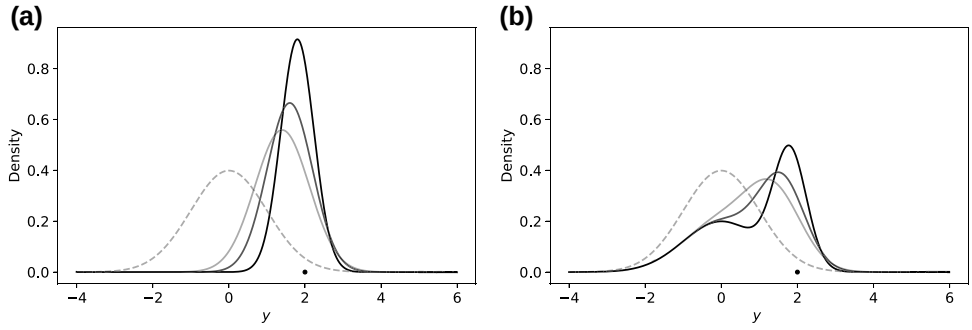

Figure 2. Current predictive density $p_i(y)$ (- - -) and new datum y_{i+1} (•); (a) Copula kernel $c_\rho(u_i, v_i)p_i(y)$ for correlation $\rho = 0.7, 0.8, 0.9$ (——,——,——); (b) Corresponding updated predictive density $p_{i+1}(y)$ (——,——,——) for $\alpha_{i+1} = 0.5$; note that we write $u_i = P_i(y)$, $v_i = P_i(y_{i+1})$.

The bandwidth ρ controls the smoothness of the density estimate, which we can set in a data-dependent manner, as we show in Section 4.5.2. On the other hand, the sequence α_i is responsible for the uncertainty as we will see in Section 6, so extra care must be taken when eliciting this. Hahn et al. (2018) suggest the form $\alpha_i = (i+1)^{-1}$ inspired from the stick-breaking process of the posterior DP as in the Bayesian bootstrap, which works well for estimating $p_n(y)$ but we find this performs poorly when predictive resampling, giving too little uncertainty. This was also observed in Fortini and Petrone (2020) in the case of Newton's recursive method. However, it should be observed that the posterior over the mixing distribution G is actually a mixture of DPs, that is

$$[G \mid \theta_{1:n}, y_{1:n}] \sim \mathrm{DP}\left(a+n, \frac{aG_0 + \sum_{i=1}^n \delta_{\theta_i}}{a+n}\right), \quad [\theta_{1:n} \mid y_{1:n}] \sim \pi(\theta_{1:n} \mid y_{1:n})$$

where $\pi(\theta_{1:n} \mid y_{1:n})$ is intractable. As shown in Online Supplementary Material, Appendix E.1.1, we only require the simplifying assumption of $\pi(\theta_{1:n} \mid y_{1:n}) = \prod_{i=1}^n G_0(\theta_i)$, which corresponds to each datum belonging to its own cluster in a similar spirit to the KDE. This then returns us the same copula update as (4.4) with

$$\alpha_i = \left(2 - \frac{1}{i}\right)\frac{1}{i+1}. \tag{4.6}$$

Intuitively, the additional mixing over $\theta_{1:n}$ results in the inflated value compared to $\alpha_i = (i+1)^{-1}$. Note this is still $\mathcal{O}(i^{-1})$, matches with initial update step for $i = 1$, and works much better in practice as it approaches 0 more slowly. We use this sequence for the remainder of the copula methods.

4.3 Multivariate case

In this section, we extend the univariate method to multivariate data $\mathbf{y} \in \mathbb{R}^d$, allowing us to both learn $p_n(\mathbf{y})$ recursively and retain the c.i.d. sequence so we can predictively resample to obtain uncertainty. Even without predictive resampling, a general multivariate density estimator $p_n(\mathbf{y})$ is of interest, as the KDE is known to perform poorly in high dimensions; see Wang and Scott (2019) for a review. Computation for the multivariate DPMM (Escobar & West, 1995; MacEachern, 1994; Neal, 2000) may scale poorly as the number of dimensions grows. Variational inference (VI) is a quicker approximation, as demonstrated in Blei and Jordan (2006), but there is a strong dependence on the optimization procedure, which may impair performance in high dimensions. A copula method for bivariate data is suggested in the appendix of Hahn et al. (2018), but it does not scale well with dimensionality and is not c.i.d.. A recursive method for multivariate density estimation is introduced in Cappello and Walker (2018), but numerical integration on a grid is still required, which scales exponentially with d, or a Monte Carlo scheme is required. Fortini and Petrone

(2020) propose a multivariate extension of Newton's recursive method, but it also requires an approximate Monte Carlo scheme to evaluate the predictive density.

Extending the above argument in Corollary 1 to multivariate data is not as straightforward, as we would like to factorize the joint density into $p_i(\mathbf{y}, \mathbf{y}_{i+1}) = k(\mathbf{y}, \mathbf{y}_{i+1})p_i(\mathbf{y})p_i(\mathbf{y}_{i+1})$, which does not have the copula interpretation like in the 2-dimensional case. Furthermore, building high-dimensional copulas are a difficult task, and bivariate copulas are good building blocks for higher dimensional dependency (Aas et al., 2009; Bedford & Cooke, 2001; Joe & Xu, 1996).

4.3.1 Factorized kernel

With the above in mind, we now consider the first step update of a multivariate DPMM below

$$f_G(\mathbf{y}) = \int \prod_{j=1}^{d} \mathcal{N}(y^j \mid \theta^j, 1)\, \mathrm{d}G(\boldsymbol{\theta}), \quad G \sim \mathrm{DP}(a, G_0), \quad G_0(\boldsymbol{\theta}) = \prod_{j=1}^{d} \mathcal{N}(\theta^j \mid 0, \tau^{-1})$$

where y^j is the jth dimension of $\mathbf{y}$, and likewise for θ^j. Note the factorized normal kernel and independent priors for each θ^j. From this, we see that we can factorize $p_0(\mathbf{y}) = \prod_{j=1}^{d} p_0(y^j)$. It is shown in Online Supplementary Material, Appendix E.1.2 that the first update step takes on the form

$$p_1(\mathbf{y}) = \left[1 - \alpha_1 + \alpha_1 \prod_{j=1}^{d} c_\rho\{P_0(y^j), P_0(y_1^j)\}\right] p_0(\mathbf{y})$$

where y_i^j is the jth dimension of the ith data point. However, naively using this update for $i > 1$ will result in the sequence $p_i(\mathbf{y})$ no longer satisfying the martingale condition in (4.1), and we also find that it performs poorly empirically. A simple but key extension allows us to retain the c.i.d. sequence:

$$p_{i+1}(\mathbf{y}) = \left\{1 - \alpha_{i+1} + \alpha_{i+1} \prod_{j=1}^{d} c_\rho(u_i^j, v_i^j)\right\} p_i(\mathbf{y}) \tag{4.7}$$

where

$$u_i^j = P_i(y^j \mid y^{1:j-1}), \quad v_i^j = P_i(y_{i+1}^j \mid y_{i+1}^{1:j-1}).$$

The input to the bivariate normal copula is now the *conditional* cumulative distribution function at $\mathbf{y}$ and $\mathbf{y}_{i+1}$ for a particular dimension ordering, and this change ensures many desirable properties. Firstly, we can verify that the martingale condition (4.1) now holds through a multivariate change of variables from $\mathbf{y}_{i+1}$ to $v_i^{1:d}$, so the c.i.d. condition is satisfied. By marginalizing y^d, y^{d-1}, ..., y^{k+1} in descending order, we also have that the marginals for a single ordering of dimensions has the same update

$$p_{i+1}(y^{1:k}) = \left\{1 - \alpha_{i+1} + \alpha_{i+1} \prod_{j=1}^{k} c_\rho(u_i^j, v_i^j)\right\} p_i(y^{1:k}). \tag{4.8}$$

From this, we can update the conditional distribution functions via

$$u_{i+1}^k = \left\{(1 - \alpha_{i+1})u_i^k + \alpha_{i+1} H_\rho(u_i^k, v_i^k) \prod_{j=1}^{k-1} c_\rho(u_i^j, v_i^j)\right\} \frac{p_i(y^{1:k-1})}{p_{i+1}(y^{1:k-1})} \tag{4.9}$$

and likewise for v_{i+1}^k. As a result, all terms in the update (4.7) can be computed tractably, with no need for numerical integration or approximations, allowing us to extend this method to any number of dimensions as computation complexity is linear in d. Notably, we must specify an ordering of the dimensions of $\mathbf{y}$, which at first may seem undesirable. However, it is not an assumption on dependence, and the only implication is that the subset of ordered marginal distributions continue to satisfy (4.8), which is a sort of marginal coherence. Interestingly, the form of (4.8) suggests that $p_i(y^{1:k})$ depends only on the first k dimensions of $\mathbf{y}_{1:i}$. Practically, we find the dimension ordering makes little difference, and we recommend selecting the ordering such that any conditional or marginal distributions of interest remain tractable. In Online Supplementary Material, Appendix E.1.3, we provide an extension to the above for mixed-type data.

Predictive resampling again takes on a simple form due to the nature of the update (4.7). We can imagine drawing each dimension of $\mathbf{Y} \sim P_i(\cdot)$ in a sequential nature, that is

$$[Y^1] \sim P_i(y^1), \quad [Y^2 \mid y^1] \sim P_i(y^2 \mid y^1), \ \ldots, [Y^d \mid y^{1:d-1}] \sim P_i(y^d \mid y^{1:d-1}). \tag{4.10}$$

Letting V_i^j denote $P_i(Y^j \mid Y^{1:j-1})$, we then have that $V_i^j \overset{\text{iid}}{\sim} \mathcal{U}[0, 1]$ for $j = \{1, \ldots, d\}$, which we can substitute into (4.7) and (4.9), similar to the univariate case. Predictive resampling again only requires sampling d independent uniform random variables for each forward step and computing the update.

4.4 Regression

We now consider extending the copula method and predictive resampling to the regression setting, where we have univariate $y_i \in \mathbb{R}$ (which can be easily extended to multivariate) with corresponding covariates $\mathbf{x}_i \in \mathcal{X}$, where, for example, $\mathcal{X} = \mathbb{R}^d$. We will later also consider binary regression, where $y_i \in \{0, 1\}$. One assumption is that the covariates are random, where we write $\{y_i, \mathbf{x}_i\} \overset{\text{iid}}{\sim} f_0(y, \mathbf{x})$, and we are interested in $f_0(y_i \mid \mathbf{x}_i)$. We term this the 'joint method', as we infer the full joint $f_0(y_i, \mathbf{x}_i)$ from which the conditional then follows. Examples of this are Müller et al. (1996), Shahbaba and Neal (2009), and Hannah et al. (2011), where the prior on $f_0(y_i, \mathbf{x}_i)$ is a DPMM. The second type of assumption, which we call the 'conditional method', is the more common framework. Here we assume that $\mathbf{x}_{1:n}$ are fixed design points and the randomness arises from the response $y_{1:n}$, so we infer a family of conditional densities $\{f_{\mathbf{x}}(y) : \mathbf{x} \in \mathcal{X}\}$. The most common framework is the additional assumption of $y_i = g(\mathbf{x}_i) + \epsilon_i$, where ϵ_i are independent zero-mean noise, and a prior on the mean function g is assumed, e.g., a Gaussian process (Rasmussen, 2003). Alternatively, one can elicit a prior on $\{f_{\mathbf{x}}(y) : \mathbf{x} \in \mathcal{X}\}$ directly, for example, with mixture models based on the dependent Dirichlet process (MacEachern, 1999). We recommend Wade (2013), Wade et al. (2014), and Quintana et al. (2022) for thorough reviews.

4.4.1 Joint method

The joint method follows easily from the multivariate: we first estimate the joint predictive density $p_{i+1}(y, \mathbf{x})$, then compute the conditional $p_{i+1}(y \mid \mathbf{x}) = p_{i+1}(y, \mathbf{x})/p_{i+1}(\mathbf{x})$. Utilizing (4.8), we have the tractable update for the conditional density

$$p_{i+1}(y \mid \mathbf{x}) = p_i(y \mid \mathbf{x}) \frac{\{1 - \alpha_{i+1} + \alpha_{i+1} c_{\rho_y}(q_i, r_i) \prod_{j=1}^d c_{\rho_x}(u_i^j, v_i^j)\}}{\{1 - \alpha_{i+1} + \alpha_{i+1} \prod_{j=1}^d c_{\rho}(u_i^j, v_i^j)\}} \tag{4.11}$$

where

$$\begin{aligned} q_i &= P_i(y \mid \mathbf{x}), \qquad r_i = P_i(y_{i+1} \mid \mathbf{x}_{i+1}) \\ u_i^j &= P_i(x^j \mid x^{1:j-1}), \; v_i^j = P_i(x_{i+1}^j \mid x_{i+1}^{1:j-1}). \end{aligned} \tag{4.12}$$

Here, we can have separate bandwidths for y and $\mathbf{x}$, and even one for each dimension of $\mathbf{x}$. The updates for $q_{i+1}, r_{i+1}, u_{i+1}^j, v_{i+1}^j$ are the same as in (4.9), and again all terms are tractable. Predictive resampling, in this case, requires simulating both $\{Y, \mathbf{X}\} \sim P_i(y, \mathbf{x})$ just like in (4.10).

4.4.2 *Conditional method*

When $\mathbf{x}$ is high-dimensional, it may be cumbersome to model $p_n(\mathbf{x})$ when we are only interested in the conditional density. The conditional method models $p(y \mid \mathbf{x})$ directly, and we turn to the dependent Dirichlet process (DDP) and its extensions for inspiration. In particular, consider the general covariate-dependent stick-breaking mixture model

$$f_{G_{\mathbf{x}}}(\mathbf{y}) = \int \mathcal{N}(y \mid \theta, 1)\, \mathrm{d}G_{\mathbf{x}}(\theta), \quad G_{\mathbf{x}} = \sum_{k=1}^{\infty} w_k(\mathbf{x})\delta_{\theta_k^*}$$

where $w_k(\mathbf{x})$ follows an $\mathbf{x}$-dependent stick-breaking process, and $\theta_k^* \overset{\text{iid}}{\sim} \mathcal{N}(\theta \mid 0, \tau^{-1})$. A full derivation is provided in Online Supplementary Material, Appendix E.2.2. We can show that the update step of the predictive takes the form

$$p_{i+1}(y \mid \mathbf{x}) = \{1 - \alpha_{i+1}(\mathbf{x}, \mathbf{x}_{i+1}) + \alpha_{i+1}(\mathbf{x}, \mathbf{x}_{i+1})c_{\rho_y}(q_i, r_i)\}p_i(y \mid \mathbf{x}) \tag{4.13}$$

where $\alpha_1(\mathbf{x}, \mathbf{x}') = \sum_{k=1}^{\infty} E[w_k(\mathbf{x})w_k(\mathbf{x}')]$, $\rho_y = 1/(1+\tau)$ and q_i, r_i are as in (4.12). The term $\alpha_1(\mathbf{x}, \mathbf{x}')$ is tractable for some choices of the construction of $w_k(\mathbf{x})$, e.g., the kernel stick-breaking process (Dunson & Park, 2008). Unfortunately, this does not provide guidance on how to generalize to $\alpha_i(\mathbf{x}, \mathbf{x}')$. Instead, we turn to the joint copula method in the previous section for inspiration, which can be written as (4.13) with

$$\alpha_i(\mathbf{x}, \mathbf{x}') = \frac{\alpha_i \prod_{j=1}^{d} c_{\rho_x}(u_{i-1}^j, v_{i-1}^j)}{1 - \alpha_i + \alpha_i \prod_{j=1}^{d} c_{\rho_x}(u_{i-1}^j, v_{i-1}^j)}.$$

This form of $\alpha_i(\mathbf{x}, \mathbf{x}')$ can be viewed as a distance measure between $\mathbf{x}$ and $\mathbf{x}'$ that is dependent on $P_n(\mathbf{x})$ which is updated in parallel. To avoid modelling $P_n(\mathbf{x})$, we can simplify the above and consider the following as a distance function directly:

$$\alpha_i(\mathbf{x}, \mathbf{x}') = \frac{\alpha_i \prod_{j=1}^{d} c_{\rho_{x,j}}\{\Phi(x^j), \Phi(x'^j)\}}{1 - \alpha_i + \alpha_i \prod_{j=1}^{d} c_{\rho_{x,j}}\{\Phi(x^j), \Phi(x'^j)\}} \tag{4.14}$$

which is equivalent to the joint method but leaving $P_n(\mathbf{x}) = P_0(\mathbf{x})$ without updating, providing us an increase in computational speed. This form requires $\mathbf{x}_{1:n}$ to be standardized for good performance, and we find that specifying independent bandwidths for each dimension in $\mathbf{x}$ works well. This method is similar to the normalized covariate-dependent weights of Antoniano-Villalobos et al. (2014).

If $\mathbf{x}_{1:n}$ is indeed a subsequence of a deterministic sequence of design points $\mathbf{x}_1, \mathbf{x}_2, \ldots$, then predictive resampling simply involves selecting $\mathbf{x}_i$ for $i > n$ from this sequence and drawing $[Y_{i+1} \mid \mathbf{x}_{i+1}] \sim P_i(y \mid \mathbf{x}_{i+1})$. If $\mathbf{X}_{1:n}$ is actually random and we have chosen the conditional approach simply for convenience, then we can draw the future $\mathbf{X}_{n+1:N}$ from the sequence of empirical predictives as in the Bayesian bootstrap. We have, however, noticed some numerical sensitivity to this choice of $P_n(\mathbf{x})$ in the uncertainty in $p_n(y \mid \mathbf{x})$ for $\mathbf{x}$ far from the observed dataset; this is illustrated in Online Supplementary Material, Appendices G.5 and G.6. Once again, conditional on $\mathbf{X}_{i+1} = \mathbf{x}_{i+1}$, we have that $P_i(Y_{i+1} \mid \mathbf{x}_{i+1}) \sim \mathcal{U}[0, 1]$, so predictive resampling only consists of simulating independent uniform random variables and updating. An example of using the Bayesian bootstrap for the covariates is provided in Online Supplementary Material, Appendix G.6.

4.4.3 *Classification*

For classification, both the joint and conditional approach generalize easily to when $y_i \in \{0, 1\}$. To this end, we can derive the copula update for a beta-Bernoulli mixture. As shown in Online

Supplementary Material, Appendix E.3, this gives

$$d_{\rho_y}\{q_i, r_i\} = \begin{cases} 1 - \rho_y + \rho_y \frac{q_i \wedge r_i}{q_i r_i} & \text{if } y = y_{i+1} \\ 1 - \rho_y + \rho_y \frac{q_i - \{q_i \wedge (1 - r_i)\}}{q_i r_i} & \text{if } y \neq y_{i+1} \end{cases}$$

where $q_i = p_i(y \mid \mathbf{x})$, $r_i = p_i(y_{i+1} \mid \mathbf{x}_{i+1})$ and $\rho_y \in (0, 1)$. We can simply replace the bivariate Gaussian copula density $c_{\rho_y}(q_i, r_i)$ in (4.11) and (4.13) with $d_{\rho_y}(u_i, v_i)$. One can check that q_i is indeed a martingale when predictive resampling, and forward sampling can be done directly as drawing binary Y_{n+1} from the Bernoulli predictive is straightforward. Unfortunately, we do not have the useful property of $P_i(y_{i+1}) \sim \mathcal{U}[0, 1]$ in the discrete case, so predictive resampling beyond the Bayesian bootstrap for $\mathbf{X}_{n+1:N}$ is computationally expensive at $\mathcal{O}(N^2)$, or approximation via a grid is required. The Bayesian bootstrap for $\mathbf{X}_{n+1:N}$ is still feasible as we only need to compute $p_N(y \mid \mathbf{x})$ at the observed $\mathbf{x}_{1:n}$. An example of this method is provided in Online Supplementary Material, Appendix G.5.

4.5 Practical considerations

In this section, we discuss some practical considerations. Further details, such as those regarding sampling and optimization, are given in Online Supplementary Material, Appendix F.

4.5.1 Initial density

For the copula methods, we require an initial guess $p_0(\mathbf{y})$ to begin our recursive updates, which can contain prior information. As it is a statement on observables, it is easier to elicit than a traditional Bayesian prior. In practice, we recommend standardizing each variable in the data $y^j_{1:n}$ to have mean 0 and variance 1 and using the default initialization $\mathcal{N}(y^j \mid 0, 1)$ for each dimension in an empirical Bayes fashion. For discrete variables, a suitable default choice is the uniform distribution over the classes. Finally, in the regression case, we can include prior information on the regression function, e.g., $p_0(y \mid \mathbf{x}) = \mathcal{N}(y \mid \beta^{\mathrm{T}}\mathbf{x}, 1)$. However, $p_0(y \mid \mathbf{x}) = \mathcal{N}(y \mid 0, 1)$ tends to work well as a default choice.

4.5.2 Hyperparameters

As we recommend the fixed form of α_i in (4.6), the only hyperparameter in the copula update is the constant ρ which parameterizes the bivariate normal copula in (4.5). While Hahn et al. (2018) suggest a default choice for ρ, we prefer a data-driven approach. Fortunately, there is an obvious method to select ρ using the prequential log score of Dawid (1984), that is to maximize $\sum_{i=1}^n \log p_{i-1}(\mathbf{y}_i)$ for density estimation or $\sum_{i=1}^n \log p_{i-1}(y_i \mid \mathbf{x}_i)$ for regression, which is related to a cross-validation metric (Fong & Holmes, 2020; Gneiting & Raftery, 2007). This fits nicely into our simulative framework, as ρ is selected on how well the sequence of predictives forecasts consecutive data points, which then informs us on the future predictives for predictive resampling. We can also specify a separate ρ_j for each dimension, which corresponds to differing length scales for the update from each conditional distribution. For optimization, gradients with respect to ρ can be computed quickly using automatic differentiation.

4.5.3 Permutations

Due to our relaxation of exchangeability in Section 3.2, one downside to the copula update and c.i.d. sequences, in general, is the dependence of p_n on the permutation of $y_{1:n}$ when there is no natural ordering of the data. For permutation invariance, we can average p_n and the corresponding prequential log-likelihood over M random permutations of $y_{1:n}$. We find in practice that $M = 10$ is sufficient, which is computationally feasible for moderate n due to the speed of the copula update, and the method is also parallelizable over permutations. For predictive resampling, we then begin with the permutation averaged p_n and forward sample with the copula update. From asymptotic exchangeability in Theorem 3 in Section 6.1, averaging over permutations is not required for forward sampling provided N is chosen to be sufficiently large. Theoretical properties of permutation averaging are explored in Tokdar et al. (2009), Dixit and Martin (2019), which we do not consider here.

4.5.4 Computational complexity

For computing $p_n(\mathbf{y})$ in the multivariate copula method, there is an overhead of first computing v_i^j for $j \in \{1, \ldots, d\}$, $i \in \{0, \ldots, n-1\}$ using (4.9), which requires $\mathcal{O}(n^2 d)$ operations, followed by $\mathcal{O}(nd)$ operations to compute $p_n(\mathbf{y})$ at a single $\mathbf{y}$ (which is then parallelizable). After computing $p_n(\mathbf{y})$, predictive resampling N future observables requires $\mathcal{O}(Nd)$ for each sample of $p_N(\mathbf{y})$; this is fully parallelizable across test points and posterior samples. Interestingly, we first compute $p_n(\mathbf{y})$ and only predictively resample after if uncertainty is desired, allowing for large computational savings if we are only interested in prediction. The regression methods have a similar computational cost.

5 Illustrations

In this section, we demonstrate the martingale posteriors induced by the copula methods of the previous section. Code for all experiments is available online at https://github.com/edfong/MP. We will demonstrate the copula method on examples where θ_0 is the density itself or the loss function induces a simple parameter, e.g., quantiles. However, any θ_0 of interest (as in Section 2.2.2) can technically be computed directly from the density or from $y_{1:n}$ and samples of $Y_{n+1:\infty}$, although this may require a high-dimensional grid or relatively expensive sampling. As a result, for cases with complex loss functions that do not rely on the smoothness of F_∞ (e.g., a parametric log-likelihood), we recommend the Bayesian bootstrap instead as a computationally efficient predictive resampling approach. For examples regarding the Bayesian bootstrap, we refer the reader to the references in Section 2.4, and we qualitatively compare the Bayesian bootstrap and the copula methods in Section 7.

For all examples, we follow the recommendations of Section 4.5 for P_0 and averaging over permutations. We will demonstrate the monitoring of convergence to P_∞, but we set $N = n + 5000$ as a standard default for the number of forward samples, where n is the size of the dataset. All copula examples are implemented in JAX (Frostig et al., 2018), which is a Python package popular in the machine learning community. JAX is ideal for our copula updates: its just-in-time compilation facilitates a dramatic speed-up for our iterative updates especially on a GPU, and its efficient automatic differentiation allows for quick hyperparameter selection. Note that the first execution of code induces an overhead compilation time of between 10–20 s for all examples. We carry out all copula experiments on an Azure NC6 Virtual Machine, which has a one-half Tesla K80 GPU card. The copula methods consist of many parallel simple computations on a matrix of density values, which is very suitable for a GPU, unlike traditional MCMC. The DPMM with MCMC examples are implemented in the `dirichletprocess` package (Ross & Markwick, 2018), which utilizes Gibbs sampling. Other benchmarks are implemented in `sklearn` (Pedregosa et al., 2011). Unless otherwise stated, default hyperparameter values are set for baselines. As the baseline packages are designed for CPU usage, we run them on a 2.6 GHz 6-Core Intel Core i7-8850H CPU. Further details can be found in Online Supplementary Material, Appendix G.2.

5.1 Density estimation

5.1.1 Univariate Gaussian mixture model

We begin by demonstrating the validity of the martingale posterior uncertainty returned from predictive resampling by comparing to a traditional DPMM in a simulated example, where the true density is known. We also discuss the monitoring of convergence of predictive resampling. For the data, we simulate $n = 50$ and $n = 200$ samples from a Gaussian mixture model:

$$f_0(y) = 0.8\mathcal{N}(y \mid -2, 1) + 0.2\mathcal{N}(y \mid 2, 1).$$

For all plots, we compute the copula predictive $p_n(y)$ on an even grid of size 160. Figures 3 and 4 show the martingale posterior density using the copula method for $n = 50$ and $n = 200$, respectively, compared to the traditional DPMM of Escobar and West (1995) with MCMC. We draw $B = 1000$ samples for both methods. We see that the resulting uncertainty and posterior means are comparable between the copula and DPMM, and the uncertainty decreases as n increases. The true density is largely contained within the 95% credible intervals.

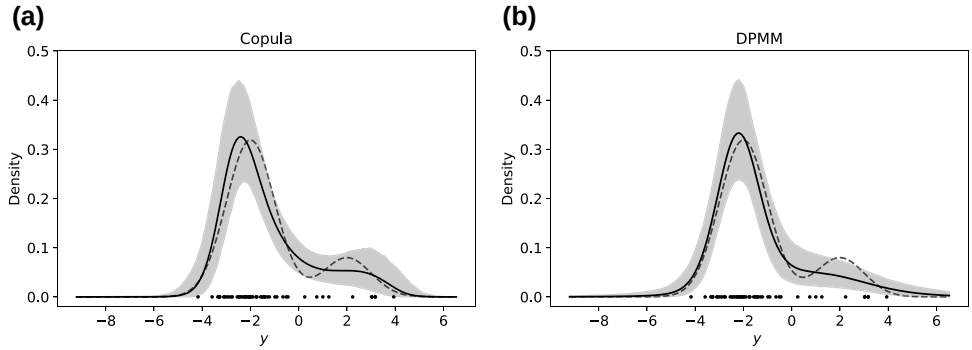

Figure 3. Posterior mean (—) and 95% credible interval ( ) of (a) $p_N(y)$ for the copula method and (b) $p_\infty(y)$ for the DPMM, for $n = 50$ with true density (- - -) and data (•).

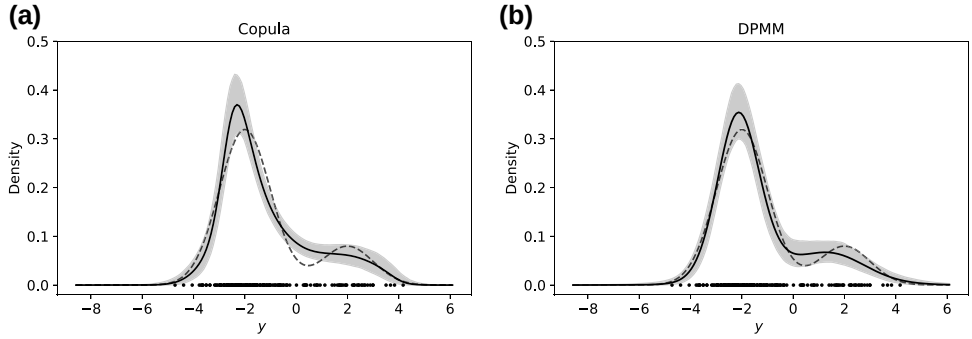

Figure 4. Posterior mean (—) and 95% credible interval ( ) of (a) $p_N(y)$ for the copula method and (b) $p_\infty(y)$ for the DPMM, for $n = 200$ with true density (- - -) and data (•).

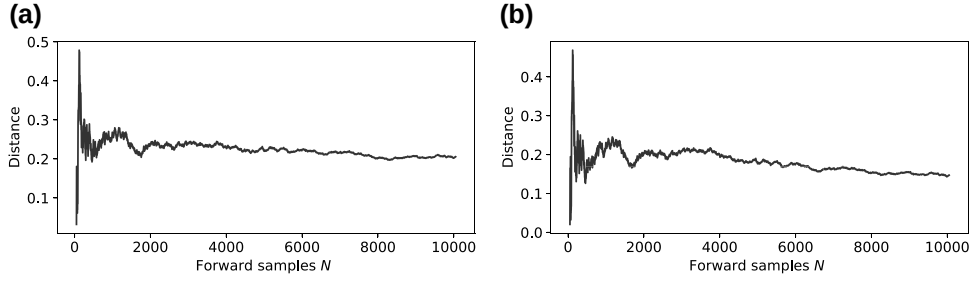

Figure 5. Estimated L_1 distance (a) $\|p_N - p_n\|_1$ and (b) $\|P_N - P_n\|_1$ for a single forward sample for $n = 50$.

For predictive resampling with the copula method, we judge convergence by considering the L_1 distance between the forward sampled p_N and initial p_n. This is demonstrated in Figure 5 for a single forward sample for $n = 50$. On the left, we have a numerical estimate of $\|p_N - p_n\|_1$ which converges to a constant, and likewise for $\|P_N - P_n\|_1$ on the right, where $\|\cdot\|_1$ is the L_1 norm and is computed on the grid. We see in this example that $N = n + 5000$ is sufficiently large for p_N to approximate p_∞. When we are not plotting on a grid and instead predicting over some test set, we may instead monitor

$$\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} |p_N(y_i) - p_n(y_i)|.$$

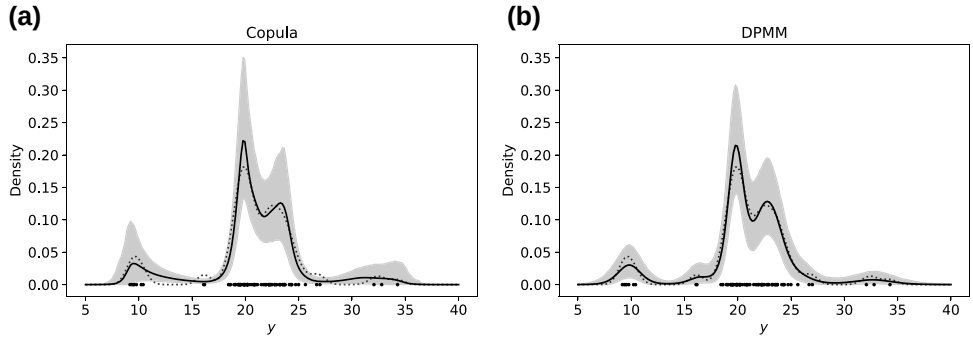

Figure 6. Posterior mean (—) and 95% credible interval (▒) of (a) $p_N(y)$ for the copula method and (b) $p_\infty(y)$ for the DPMM, with KDE (·····) and data (•).

Optimization of the prequential log-likelihood gives us the optimal hyperparameter $\rho = 0.77$ and 0.78 for $n = 50$ and 200, respectively. The prequential log-likelihood is returned easily from the copula method, allowing for easy hyperparameter selection. However, computing the marginal likelihood for the DPMM is non-trivial, and thus setting the hyperparameters of the priors in a data-driven way, that is empirical Bayes, remains a difficult task. Here, we select the DPMM hyperparameters to match the smoothness of the posterior mean of the copula method for comparability of the uncertainty.

5.1.2 Univariate galaxy dataset

We now demonstrate the martingale posterior sampling of a parameter of interest that requires a smooth density, through predictive resampling and the computation of $\theta(P_N)$. We analyse the classic 'galaxy' dataset (Roeder, 1990), thereby extending the example of Hahn et al. (2018) to the predictive resampling framework. The dataset consists of $n = 82$ velocity measurements of galaxies in the Corona Borealis region. For all plots, we compute $p(y)$ on an even grid of size 200, and unnormalize after the copula method so that the scale of y is in km/s.

Figure 6 compares predictive resampling with the copula method for $B = 1000$ posterior samples of p_N, where the selected bandwidth is $\rho = 0.93$. The bandwidth for KDE was computed through 10-fold cross-validation, and DPMM hyperparameters are set to the suggested values in West (1991). The 95% credible intervals and posterior mean of the copula approach are comparable with that of the DPMM. Excluding compilation times, the optimization for ρ and computation of $p_n(y)$ on the grid of size 200 took 0.5 s, and predictive resampling took 2 s. In comparison, DPMM with MCMC took 25 s for the same number of samples ($B = 1000$), where the samples are not independent; the plots for MCMC are thus produced with $B = 2000$. Given this random density, we can also compute the statistics of interest θ directly from the grid of density values. Martingale posterior samples of the number of modes and 10% quantiles of the random density are shown in Figure 7, with comparison to the DPMM. Here the copula method tends to prefer 4 modes, whereas the DPMM prefers 5.

5.1.3 Bivariate air quality dataset

We demonstrate the martingale posterior for bivariate data using the method of Section 4.3.1, which has large computational gains over posterior sampling with DPMM when the density is of interest, where the latter is expensive due to dimensionality. For this, we look at the 'airquality' dataset (Chambers, 2018) from `DPpackage`. The dataset consists of daily ozone and solar radiation measurements in New York, with $n = 111$ completed data points. For all plots, we compute $p_n(\mathbf{y})$ on a grid of size 25×25.

We fit the multivariate copula method of Section 4.3.1 with one bandwidth per dimension, and optimizing the prequential log-likelihood returns $\rho = [0.47, 0.82]$. Predictive resampling $B = 1000$ martingale posterior samples returns us the martingale posterior mean and standard deviation of

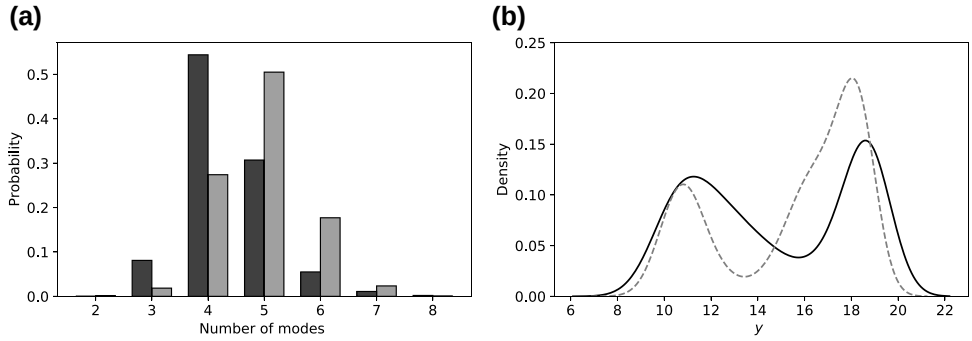

Figure 7. (a) Posterior samples of number of modes for the copula method (■) and DPMM (■); (b) Posterior density of 10% quantiles for the copula method (—) and the DPMM (- - -).

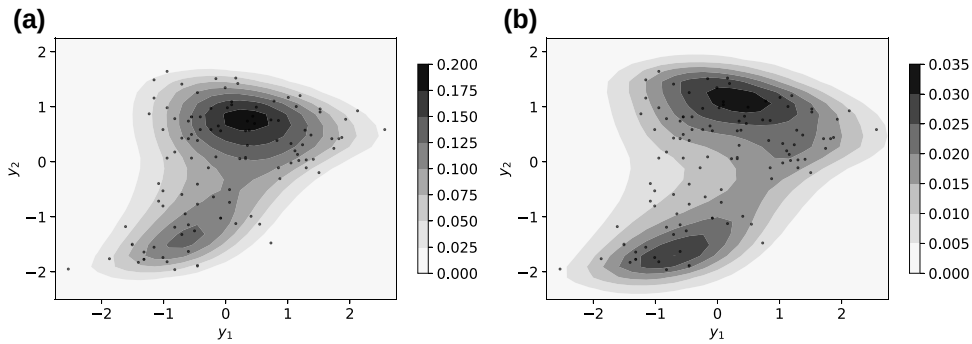

Figure 8. Posterior (a) mean and (b) standard deviation of $p_N(\mathbf{y})$ for the copula method with scatter plot of data (•).

the bivariate density as shown in Figure 8. Again excluding compilation times, the optimization for ρ and computation of $p_n(y)$ on the grid of size 625 took 1 s, and predictive resampling took 10 s in total. For comparison, the DPMM with MCMC required 4 min for the same number of samples. Further details and comparisons to the DPMM are given in Online Supplementary Material, Appendix G.4.

Figure 9 plots a martingale posterior sample of the density, with the corresponding L_1 distance convergence plot. We see that $N = 5000$ is again sufficient, which suggests a dimension independent convergence rate of $P_N \to P_\infty$. This is justified in the theory in Section 6.

5.1.4 Multivariate UCI datasets

In this section, we demonstrate the multivariate copula method of Section 4.3.1 as a highly effective density estimator compared to the usual DPMM, as we do not need to deal with the posterior sampling or integration over high-dimensional parameters. We demonstrate on multivariate datasets from the UCI Machine Learning Repository (Asuncion & Newman, 2007). To prevent misleadingly high-density values, we remove non-numerical variables and one variable from any pairs with Pearson correlation coefficient greater than 0.98 (e.g., see Tang et al., 2012). We compare to the KDE, DPMM and multivariate Gaussian and evaluate the methods with a 50-50 test-train split and average the test log-likelihoods over 10 random splits.

For the copula method, we use a single value of ρ for all dimensions for a fair comparison to the KDE. We find that having distinct $\rho_{1:d}$ slightly improves predictive performance at the cost of higher optimization times. For the KDE, we use a single scalar bandwidth set through 10-fold cross-validation. For the DPMM, we set the Gaussian kernel to have diagonal covariance matrices and use VI (Blei & Jordan, 2006). Using a full covariance matrix kernel is unreliable likely due

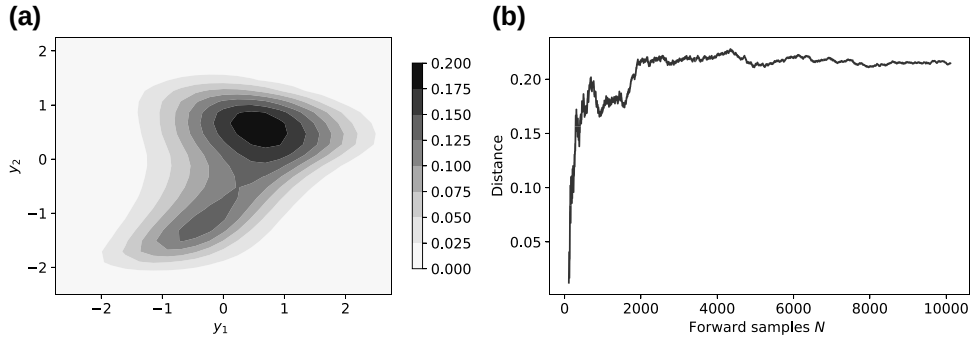

Figure 9. (a) Random sample of $p_N(\mathbf{y})$; (b) Corresponding estimated $\|p_N - p_n\|_1$.

Table 1. Average test log-likelihood, standard errors (in brackets) and best performance in bold

Dataset	n	d	Gaussian	KDE	DPMM (VI)	Copula
Breast cancer	569	26	−17.8 (0.61)	−25.6 (0.29)	−33.4 (0.80)	**−13.0** (0.26)
Ionosphere	351	32	−49.4 (1.97)	−32.3 (0.79)	−36.5 (0.59)	**−21.5** (1.63)
Parkinsons	195	16	−14.3 (0.54)	−15.6 (0.41)	−25.7 (0.92)	**−9.9** (0.28)
Wine	178	13	−16.1 (0.26)	−15.7 (0.20)	−22.8 (0.61)	**−14.6** (0.17)

to local optima for VI, and MCMC is too computationally expensive for large d. For the multivariate Gaussian, we use the empirical mean and covariance.

As shown in Table 1, the performance is significantly better on test data for these datasets. The better performance than the KDE is likely due to the regularizing effect of $p_0(\mathbf{y})$, which is important here as n is only of moderate size. The DPMM (VI) likely performs poorly as the diagonal covariance cannot capture dependent structure, and the number of variational parameters is still high so optimization is difficult. We provide a more detailed analysis of the degradation in performance with the dimensionality of the DPMM with VI in Online Supplementary Material, Appendix G.7, where the copula method remains robust to dimensionality.

Overall, the run-times for the copula method, KDE and DPMM (VI) are similar, all of which are orders of magnitude faster than the DPMM with MCMC. For a single train-test split, the slowest example of the above (Breast cancer) for the copula method required less than 4 s in total to optimize ρ, while computing the overhead ν_i^j and predicting on the test data required less than 100ms. For the same example, the KDE and DPMM (VI) required around 1.5 and 6 s respectively.

5.2 Regression and classification

5.2.1 Regression in LIDAR dataset

We now demonstrate the joint copula regression method of Section 4.4.1 on a non-linear heteroscedastic regression example, where the copula method performs well off-the-shelf. We use the LIDAR dataset from Wasserman (2006), which consists of $n = 221$ observations of the distance travelled by the light and the log ratio of intensity of the measured light from the two lasers; the latter is the dependent variable. For the plots below, we evaluate the conditional density on a y, x grid of 200×40 points.

For the copula method, we optimize the prequential conditional log-likelihood over the $M = 10$ permutations and get $\rho_y = 0.90$, $\rho_x = 0.83$. The *predictive* mean and 95% central interval of $p_n(y \mid x)$ are shown in Figure 10, compared to the DPMM, and we observe that the copula methods handle the nonlinearity better. The optimization, fitting and prediction on the grid took under 4 s for the copula method, compared to 5 min for the DPMM with MCMC for the same number of samples.

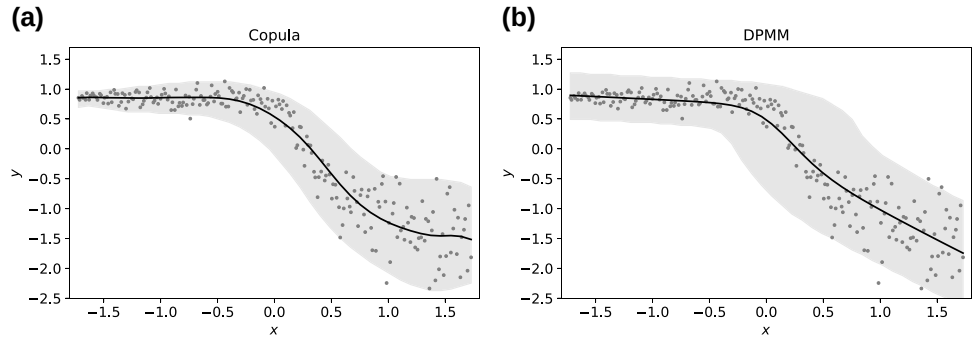

Figure 10. $p_n(y \mid x)$ (—) with 95% predictive interval ( ) for the (a) joint copula method and (b) joint DPMM, with data ( ).

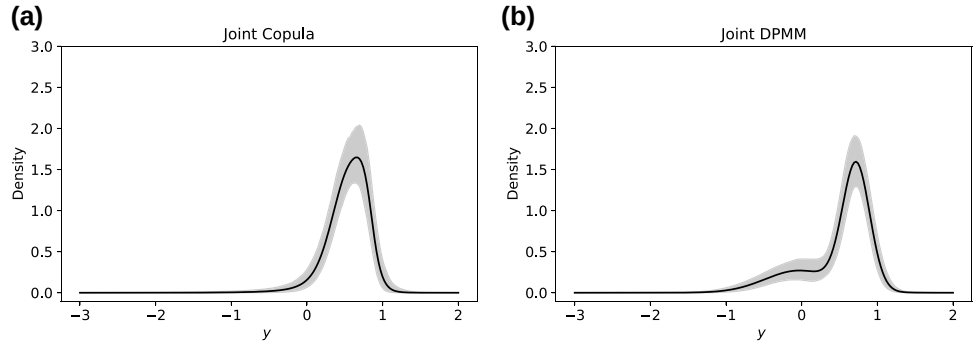

Figure 11. Posterior mean (—) and 95% credible interval ( ) of (a) $p_N(y \mid x = 0)$ for the joint copula method and (b) $p_\infty(y \mid x = 0)$ for the joint DPMM.

In Figure 11, we see martingale posterior samples of $p_N(y \mid x = 0)$ for the copula method compared to the DPMM. For reference, predictive resampling the $B = 1000$ martingale posterior samples on the y grid for a single x took under 3 s. One can see in Figure 11 that there is more posterior uncertainty in the density $p_N(y \mid x = 0)$ for the copula methods, as the DPMM has a simpler mean function (weighted sum of linear). Convergence of the conditional density under predictive resampling is now dependent on the value of x. Figure 13b shows the L_1 distances as before for $x = 0$; however, we find that more forward samples are needed for x far from the data. Figure 12 then shows martingale posterior samples of $p_N(y \mid x = -3)$ where x is far from the data, and we see that both the copula and DPMM methods have larger uncertainty as expected. However, predictive resampling for the conditional copula method of Section 4.4.2 does not always demonstrate this desirable behaviour for outlying x; the joint and conditional methods are compared in Online Supplementary Material, Appendix G.6 and this undesirable behaviour is also noted in Online Supplementary Material, Appendix G.5.

One may also be interested in the uncertainty in a point estimate for the function which we write as θ_x, in this case, the conditional median. In Figure 13a, we plot the martingale posterior mean and 95% credible interval of the conditional median of $P_N(y \mid x)$, where we see the uncertainty increasing with x . Here we predictively resample on a y, x grid of size 40×40 and compute the median numerically; this took 12 s for $B = 1000$ samples.

5.2.2 Multivariate covariates in UCI datasets

We now demonstrate the conditional copula method for prediction in the regression and classification setting with multivariate covariates, which is of particular interest to the machine learning community. For high-dimensional covariates, the conditional copula method performs better than

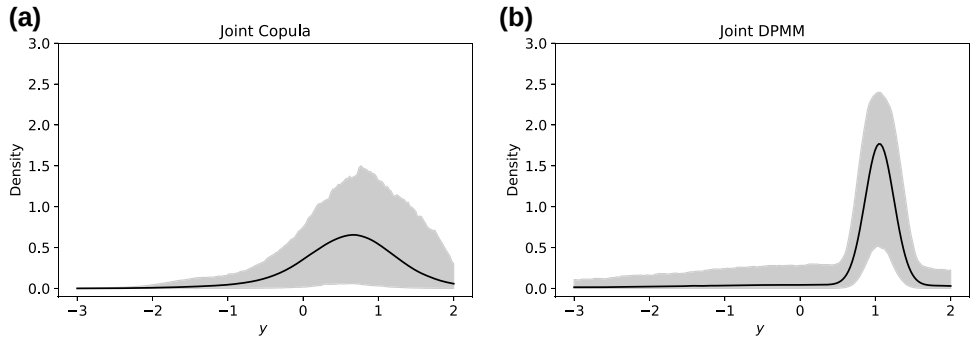

Figure 12. Posterior mean (—) and 95% credible interval (▒) of (a) $p_N(y \mid x = -3)$ for the joint copula method and (b) $p_\infty(y \mid x = -3)$ for the joint DPMM.

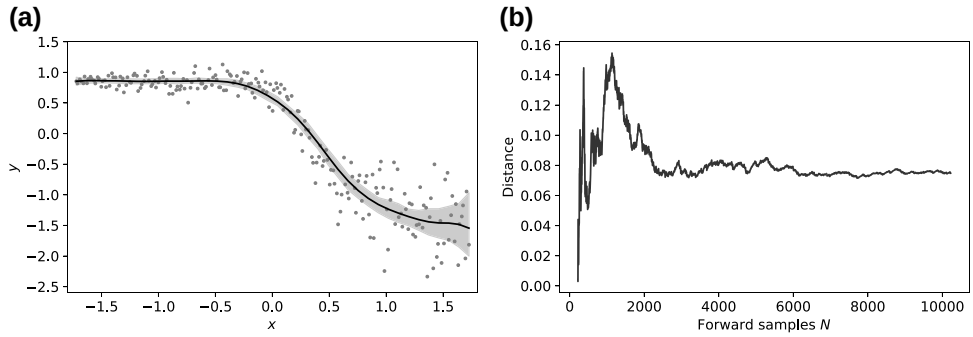

Figure 13. (a) Posterior mean (—) and 95% credible interval (▒) of the conditional median of $P_N(y \mid x)$, with data (●). (b) Estimated L_1 distance $\|p_N(\cdot \mid x) - p_n(\cdot \mid x)\|_1$ for a single forward sample with $x = 0$.

the joint method, both in terms of computational speed and test log-likelihood. This is likely due to the dominance of estimating $P_n(\mathbf{x})$ in high dimensions, which disrupts the estimate of $P_n(y \mid \mathbf{x})$.

Similar to the multivariate density estimation, we demonstrate the regression and classification conditional copula methods on UCI datasets with scalar y and multivariate $\mathbf{x}$. Again, we evaluate the methods with 10 random 50-50 test-train splits and evaluate the average test conditional log-likelihoods. We convert categorical variables into dummy variables and report the preprocessed covariate dimensionality in Table 2. We compare to Bayesian linear regression and Gaussian processes (GP) with a single length scale RBF kernel as baselines for regression, and similarly to logistic regression and GPs with the logistic link and Laplace approximation for classification. We use the Laplace approximation as it is available off-the-shelf in `sklearn`, and we found that independent kernel length scales (ARD) performed worse due to overfitting given n is moderate. For the conditional copula method, we have distinct bandwidths $\rho_{1:d}$ for each covariate, which we optimize through the prequential log-likelihood over $M = 10$ permutations.

In Table 2, we see the test log-likelihoods, where the copula method is competitive with the GP, though, in general, we find that the GP provides a better estimate for the mean function for regression. Again, optimization took the most time due to the d bandwidths, taking on average 30 s per fold for the slowest example ('Statlog'). The time for actual fitting and prediction on the test set was under 120 ms per fold for all examples. The GP on the slowest examples required around 20 s per fold for the marginal likelihood optimizations, but computation time scales as $\mathcal{O}(n^3)$.

6 Theory

In this section, we provide a theoretical analysis of the martingale posteriors and predictive resampling using the copula update introduced in Section 4. We utilize the theory of c.i.d. sequences

Table 2. Average test log-likelihood, standard errors (in brackets) and best performance in bold

	Dataset	n	d	Linear	GP	Copula
Regression	Boston	506	13	−0.842 (0.043)	−0.404 (0.040)	**−0.351** (0.025)
	Concrete	1030	8	−0.965 (0.008)	**−0.364** (0.014)	−0.445 (0.013)
	Diabetes	442	10	−1.096 (0.017)	−1.089 (0.015)	**−1.003** (0.018)
	Wine Quality	1599	11	−1.196 (0.017)	**−0.497** (0.034)	−1.143 (0.020)
Classification	Breast cancer	569	30	−0.107 (0.005)	−0.105 (0.005)	**−0.096** (0.008)
	Ionosphere	351	33	−0.348 (0.005)	**−0.304** (0.006)	−0.388 (0.016)
	Parkinsons	195	22	−0.352 (0.007)	−0.364 (0.013)	**−0.257** (0.010)
	Statlog	1000	20	**−0.530** (0.009)	−0.542 (0.011)	−0.541 (0.006)

from the works of Berti et al. (2004, 2013). We then show frequentist consistency (with little n) under relatively weak conditions for the multivariate copula update by extending the proof of Hahn et al. (2018), and we discuss its implications. All proofs are deferred to Online Supplementary Material, Appendix D.

6.1 Martingale posteriors for copula density estimation

We first analyse the properties under predictive resampling of the multivariate copula recursive update for the martingale posterior. We write $P_i(\mathbf{y})$ as the joint cumulative distribution function of the density $p_i(\mathbf{y})$ with update (4.7), and consider predictive resampling starting at $p_n(\mathbf{y})$ such that $\mathbf{Y}_{i+1} \sim P_i(\mathbf{y})$ for $i = n, n+1 \ldots, N$. As before, n corresponds to the number of observed data points, whereas $N - n$ corresponds to the number of forward samples drawn from predictive resampling. The first two results follow directly from the c.i.d. property of the sequence.

Theorem 3 (Berti et al. (2004, Theorem 2.5)). The sequence $\mathbf{Y}_{N+1}, \mathbf{Y}_{N+2}, \ldots$ is asymptotically exchangeable, that is

$$(\mathbf{Y}_{N+1}, \mathbf{Y}_{N+2}, \ldots) \xrightarrow{d} (\mathbf{Z}_1, \mathbf{Z}_2, \ldots)$$

for $N \to \infty$, where $(\mathbf{Z}_1, \mathbf{Z}_2, \ldots)$ is exchangeable.

The above justifies that we may not need to average over permutations for sufficiently large N when predictive resampling.

As mentioned in Section 3.2, we would like $P_N(\mathbf{y}) \to P_\infty(\mathbf{y})$ at each $\mathbf{y} \in \mathbb{R}^d$, which indeed holds for predictive resampling here from the c.i.d. sequence:

Theorem 4 (Berti et al., 2004, Lemmas 2.1, 2.4). There exists a random probability measure P_∞ such that P_N converges weakly to P_∞ almost surely.

Specifically for the univariate case of the copula update above, we can strengthen this to convergence in total variation, which also implies that the limiting predictive P_∞ is continuous, following from an interesting result in Berti et al. (2013).

Theorem 5 For $y \in \mathbb{R}$, suppose the sequence of probability measures P_N has density function $p_N(y)$ and cumulative distribution function $P_N(y)$ satisfying the updates (4.4). Let us assume that the initial $P_n(y)$ is continuous and its density satisfies

$$\int_K p_n^2(y)\, \mathrm{d}y < \infty$$

for all K, where K is a compact subset of $\mathbb{R}$ with finite Lebesgue measure. For the sequence

$$\alpha_i = \left(2 - \frac{1}{i}\right)\frac{1}{i+1},$$

let us assume further that $\rho < 1/\sqrt{3}$. We then have

(a) P_∞ is absolutely continuous with respect to the Lebesgue measure almost surely, with density p_∞.
(b) P_N converges in total variation to P_∞ almost surely, that is

$$\lim_{N\to\infty} \int |p_N(y) - p_\infty(y)|\, \mathrm{d}y = 0 \quad \text{a.s.}$$

The assumptions hold if $p_n(y)$ is continuous. From this, we are justified in using $p_N(y)$ as an approximate sample of the martingale posterior $p_\infty(y)$. We conjecture that the choice of $\rho < 1/\sqrt{3}$ can be relaxed, and empirically it seems the case. Furthermore, this restriction on ρ is not needed if $\alpha_i = (i+1)^{-1}$. Unfortunately, we have been unable to extend Theorem 5 to the multivariate copula update, as the update for $P(y^j \mid y^{1:j-1})$ is not as easy to bound. We also conjecture that the L_1 convergence holds true in the multivariate case, and again the empirical results suggest so.

We can also quantify to some degree the convergence rate to P_∞ as we predictively resample. We have the following result from a variant of the Azuma-Hoeffding inequality from McDiarmid (1998).

Proposition 1 For $M > N$ and any $\epsilon \geq 0$, the cumulative distribution function $P_N(\mathbf{y})$ of the density in (4.7) satisfies

$$\sup_{\mathbf{y}} \mathbb{P}(|P_M(\mathbf{y}) - P_N(\mathbf{y})| \geq \epsilon) \leq 2\exp\left(\frac{-\epsilon^2}{\frac{2\epsilon\alpha_{N+1}}{3} + \frac{1}{2}\sum_{i=N+1}^{M}\alpha_i^2}\right).$$

Taking the limit (superior) as $M \to \infty$ of the above gives insight into the quality of the approximation of P_∞ when we truncate the predictive resampling at P_N. For our choice of α_i from (4.6), we have $\sum_{i=N+1}^{\infty}\alpha_i^2 = \mathcal{O}(N^{-1})$, so the limiting probability of a difference greater than ϵ decreases roughly at rate $\exp(-\epsilon^2 cN)$ for some constant c. Notably, this rate is independent from the dimensionality d and instead depends only on the sequence α_i. Furthermore, we have some notion of posterior contraction in Proposition 1 if we instead consider N as the number of observed data points and M as the number of forward samples.

6.2 Martingale posteriors for conditional copula regression

For the regression case, where $y \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^d$, we analyse the update given in (4.13) and (4.14). Assuming we have observed $y_{1:n}$, $\mathbf{x}_{1:n}$, we draw the sequence $\mathbf{X}_{n+1:\infty}$ from the Bayesian bootstrap with $\mathbf{x}_{1:n}$. While this is no longer the traditional c.i.d. setup, we still have that $P_N(y \mid \mathbf{x})$ is a martingale under predictive resampling, so we have that $P_N(y \mid \mathbf{x})$ converges pointwise for each $\mathbf{x}$ almost surely. Fortunately, Berti et al. (2006, Theorem 2.2) assure that the martingale posterior $P_\infty(y \mid \mathbf{x})$ exists.

Theorem 6 For each $\mathbf{x} \in \mathbb{R}^d$, there exists a random probability measure $P_\infty(\cdot \mid \mathbf{x})$ such that $P_N(\cdot \mid \mathbf{x})$ converges weakly to $P_\infty(\cdot \mid \mathbf{x})$ almost surely.

We also have the appropriate extension to Proposition 1 below.

Proposition 2 For $M > N$ and any $\epsilon \geq 0$, the cumulative distribution function $P_N(y \mid \mathbf{x})$ of the density in (4.13) satisfies

$$\sup_y \mathbb{P}(|P_M(y \mid \mathbf{x}) - P_N(y \mid \mathbf{x})| \geq \epsilon) \leq 2 \exp\left(\frac{-\epsilon^2}{\frac{4\epsilon C \alpha_{N+1}}{3} + 2C^2 \sum_{i=N+1}^{M} \alpha_i^2} \right)$$

for each $\mathbf{x} \in \mathbb{R}^d$, where C depends only on ρ and $\mathbf{x}$.

It can be shown that C increases as $\mathbf{x}$ moves from the origin. Assuming $x_{1:n}$ is standardized, this implies that the number of forward samples needed for convergence may increase as $\mathbf{x}$ shifts away from the data. The above results can also be easily extended to the classification scenario.

6.3 Frequentist consistency of copula density estimation

To simulate from the martingale posterior given $\mathbf{Y}_{1:n}$, we start with the density p_n computed from (4.7), so we would like to verify that it is indeed an appropriate predictive density. In this section, we thus concern ourselves with the frequentist notion of consistency, that is we look at the properties of the density estimate p_n assuming $\mathbf{Y}_{1:n}$ is i.i.d. from some probability distribution with density function f_0 as we take $n \to \infty$. It should be noted that this is distinct from the Doob-type asymptotics of predictive resampling in the previous sections where we take $N \to \infty$.

The frequentist consistency of the univariate copula method was first discussed in Hahn et al. (2018) based on the 'almost supermartingale' of Robbins and Siegmund (1971). We will now extend the result to the multivariate copula method, of which the univariate method is a special case. The full proof can be found in Online Supplementary Material, Appendix D.6. Instead of the Kullback–Leibler divergence, we work with the squared Hellinger distance between probability density functions p_1 and p_2 on $\mathbf{y} \in \mathbb{R}^d$, defined as $H^2(p_1, p_2) := 1 - \int \sqrt{p_1(\mathbf{y}) p_2(\mathbf{y})}\, d\mathbf{y}$. We then have the main result.

Theorem 7 For $\mathbf{Y}_{1:n} \overset{\text{iid}}{\sim} f_0$, suppose the sequence of densities $p_n(\mathbf{y})$ satisfies the updates in (4.7). Assume that $\rho \in (0, 1)$, $\alpha_i = a(i+1)^{-1}$ where $a < 2/5$, and there exists $B < \infty$ such that $f_0(\mathbf{y})/p_0(\mathbf{y}) \leq B$ for all $\mathbf{y} \in \mathbb{R}^d$. We then have that p_n is Hellinger consistent at f_0, that is

$$\lim_{n \to \infty} H^2(p_n, f_0) = 0 \quad \text{a.s.}$$

Intuitively, the update (4.7) can be regarded as a stochastic gradient descent in the space of probability density functions, where α_{i+1} is the step-size. As is standard in stochastic optimization (Kushner & Yin, 2003), consistency of the copula method relies delicately on the decay of the sequence α_i, which ensures we approach the independent copula at the correct rate. A similar condition is, for example, discussed in Tokdar et al. (2009) for Newton's algorithm. On the one hand, we require $\sum_{i=1}^{\infty} \alpha_i = \infty$ to ensure that the initialization p_0 is forgotten. On the other hand, we require the sequence α_i to decay sufficiently quickly to 0, that is $\sum_{i=1}^{\infty} \alpha_i^2 < \infty$, for information to accumulate correctly. The requirement on a also ensures the information in later terms decay properly. Notably, the condition on $a < 2/5$ is different to the suggestion for predictive resampling, so a different choice of α_n may be more suitable when consistency is of primary interest. The second assumption is a regularity condition on the tails of the initial p_0 being heavier than f_0, which motivates a heavy-tailed initial density as also suggested by Hahn et al. (2018). Interestingly, the bounded condition on f_0/p_0 is the only requirement on f_0 for consistency, which follows from the nonparametric update. However, unlike the KDE there are no conditions on the bandwidth ρ, which likely follows from the data dependence of the copula kernel.

There are a number of unanswered questions when compared to the consistency of traditional Bayes. The first is whether the martingale posterior converges weakly to the Dirac measure at F_0, as we have only shown Hellinger consistency of the posterior mean measure of P_∞. We believe this is likely to be positive, as there is a notion of posterior contraction as in Proposition 1. A related

inquiry is the rate of convergence of p_n, or the martingale posterior on p_∞, to the true f_0. The second and more ambitious question is whether the above approach provides a general method to prove consistency for other copula models. For the multivariate copula method, we only require the weak tail condition on f_0, but the proof relies heavily on the nonparametric nature of the update. It is still unclear what the conditions would be if the copula sequence corresponded to a parametric Bayesian model, such as the examples given in Hahn et al. (2018). In the absence of the prior under the predictive view, a question of interest is whether an analogue to the Kullback–Leibler property of the traditional Bayesian prior (e.g., Ghosal & van der Vaart, 2017, Definition 6.15) exists, which would highlight a predictive notion of model misspecification.

7 Discussion

We see that Bayesian uncertainty, at its core, is concerned with the missing observations required to know any statistic of interest precisely. In the i.i.d. case, this is $Y_{n+1:\infty}$, and our task is to obtain the joint distribution $p(y_{n+1:\infty} \mid y_{1:n})$, which is simplified through the factorization into a sequence of 1-step ahead predictive densities. One open question is whether there are more general methods to elicit this joint beyond the likelihood-prior construction and the prequential factorization. For the more general data setting, the Bayesian would be tasked with eliciting $p(y_{\text{mis}} \mid y_{\text{obs}})$, where the missing observations y_{mis} would be specific to the setting and statistic of interest. We highlight that y_{mis} must be sufficiently large to compute the statistic precisely, unlike in multiple imputation (Rubin, 2004) where the imputed data is often finite and for computational convenience. For future work, identifying y_{mis} and extending the methodology in more complex data settings such as time series or hierarchical data is of primary interest.

In terms of practical methodology, it is worth comparing when one would prefer to use the Bayesian bootstrap versus the copula methods. When the data is high-dimensional but a low-dimensional statistic is of interest, the copula methods may not be suitable, as computing the density on a grid or sampling the data directly is required. Fortunately, the Bayesian bootstrap shines in this setting. On the other hand, the discreteness of the Bayesian bootstrap makes it unsuitable for when smoothness is required, for example, when the density is directly of interest, or in regression where we rely on smoothness with x. In these settings, the copula methods are highly suitable. Together, the predictive framework allows us to cover a wide variety of settings with practical advantages over the traditional Bayesian approach.

We believe our framework offers an interesting insight into the interplay between Bayesian and frequentist approaches. As we have seen through the lens of the Bayesian bootstrap, Bayesians and frequentists are concerned with $Y_{n+1:\infty}$ and $Y_{1:n}$, respectively. Analysis of the frequentist asymptotic properties of martingale posteriors also offers new challenges, as we must work with the predictive distribution directly, and it is unclear if the methods used in our paper generalize to other copula models. For generalizations of our martingale posterior framework, imputing aspects of the population instead of the entire population directly may also help bridge the gap between Bayesian and frequentist methods. In the hierarchical example in Section 1, we can in fact treat θ_i as the mean of population i from which we observe a single sample y_i. We would thus be imputing the means of observation populations (i.e., the random effects) instead of the entire population of observables directly. This interpretation would align well with our philosophy of only imputing what one would need to carry out the statistical task.

Acknowledgments

The authors are grateful for the detailed comments of three referees and the Associate Editor on the previous version of the paper. The authors also thank Sahra Ghalebikesabi, Brieuc Lehmann, Geoff Nicholls, George Nicholson and Judith Rousseau for their helpful comments.

Supplementary material

Supplementary material is available at *Journal of the Royal Statistical Society: Series B* online.

Conflict of interest: None declared.

Funding

E.F. was funded by The Alan Turing Institute Doctoral Studentship under the EPSRC grant EP/N510129/1. C.H. is supported by The Alan Turing Institute, the Health Data Research, UK, the Li Ka Shing Foundation, the Medical Research Council, and the U.K. Engineering and Physical Sciences Research Council.

Data availability

The data underlying this article are available in the UCI Machine Learning Repository (Asuncion & Newman, 2007), at [https://archive.ics.uci.edu/ml]. The other datasets were derived from sources in the public domain:

- 'galaxy' from the `MASS` package (Roeder, 1990) [https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/galaxies.html]
- 'airquality' from the `DPpackage` (Chambers, 2018) [https://rdrr.io/cran/DPpackage/]
- 'LIDAR' from the book (Wasserman, 2006) [https://www.stat.cmu.edu/~larry/all-of-nonpar/=data/lidar.dat]

References

Aas K., Czado C., Frigessi A., & Bakken H. (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, *44*(2), 182–198. https://doi.org/10.1016/j.insmatheco.2007.02.001

Antoniano-Villalobos I., Wade S., & Walker S. G. (2014). A Bayesian nonparametric regression model with normalized weights: A study of hippocampal atrophy in Alzheimer's disease. *Journal of the American Statistical Association*, *109*(506), 477–490. https://doi.org/10.1080/01621459.2013.879061

Asuncion A., & Newman D. (2007). UCI machine learning repository. http://www.ics.uci.edu/~mlearn/MLRepository.html

Bedford T., & Cooke R. (2001). *Mathematical tools for probabilistic risk analysis*. Cambridge University Press.

Bernardo J., & Smith A. (2009). *Bayesian theory*. Wiley Series in Probability and Statistics. Wiley.

Berti P., Dreassi E., Pratelli L., & Rigo P. (2020). A class of models for Bayesian predictive inference. *Bernoulli*, *27*(1), 702–726. https://doi.org/10.3150/20-BEJ1255

Berti P., Pratelli L., & Rigo P. (2004). Limit theorems for a class of identically distributed random variables. *The Annals of Probability*, *32*(3), 2029–2052. https://doi.org/10.1214/009117904000000676

Berti P., Pratelli L., & Rigo P. (2006). Almost sure weak convergence of random probability measures. *Stochastics and Stochastics Reports*, *78*(2), 91–97. https://doi.org/10.1080/17442500600745359

Berti P., Pratelli L., & Rigo P. (2013). Exchangeable sequences driven by an absolutely continuous random measure. *The Annals of Probability*, *41*(3B), 2090–2102.https://doi.org/10.1214/12-AOP786

Berti P., Regazzini E., & Rigo P. (1998). Well calibrated, coherent forecasting systems. *Theory of Probability & its Applications*, *42*(1), 82–102. https://doi.org/10.1137/S0040585X97975988

Bissiri P. G., Holmes C. C., & Walker S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *78*(5), 1103–1130. https://doi.org/10.1111/rssb.12158

Blackwell D., & MacQueen J. B. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, *1*(2), 353–355. https://doi.org/10.1214/aos/1176342372

Blei D. M., & Jordan M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, *1*(1), 121–143. https://doi.org/10.1214/06-BA104

Breiman L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Cappello L., & Walker S. G. (2018). A Bayesian motivated Laplace inversion for multivariate probability distributions. *Methodology and Computing in Applied Probability*, *20*(2), 777–797. https://doi.org/10.1007/s11009-017-9587-y

Chambers J. M. (2018). *Graphical methods for data analysis*. CRC Press.

Dawid A. P. (1984). Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, *147*(2), 278–290. https://doi.org/10.2307/2981683

Dawid A. P. (1985). Probability, symmetry and frequency. *The British Journal for the Philosophy of Science*, *36*(2), 107–128. https://doi.org/10.1093/bjps/36.2.107

Dawid A. P. (1992a). Prequential analysis, stochastic complexity and Bayesian inference. *Bayesian Statistics*, *4*, 109–125. https://global.oup.com/academic/product/bayesian-statistics-4-9780198522669?lang=en&cc=gb

Dawid A. P. (1992b). Prequential data analysis. In *Current issues in statistical inference: Essays in honor of D. Basu* (pp. 113–126). Hayward: Institute of Mathematical Statistics.
de Finetti B. (1937). La prévision: ses lois logiques, ses sources subjectives. In *Annales de l'institut Henri Poincaré* (Vol. 7, pp. 1–68). [English translation in Studies in Subjective Probability (1980) (H. E. Kyburg & H. E. Smokler, eds.) 53–118. Krieger, Malabar, FL.].
Dixit V., & Martin R. (2019). Permutation-based uncertainty quantification about a mixing distribution. arXiv preprint arXiv:1906.05349.
Doob J. L. (1949). Application of the theory of martingales. In *Actes du Colloque International Le Calcul des Probabilités et ses applications, Lyon, 28 Juin–3 Juillet 1948* (pp. 23–27). CNRS.
Doob J. L. (1953). *Stochastic processes*. (Vol. *101*). Wiley.
Dunson D. B., & Park J. H. (2008). Kernel stick-breaking processes. *Biometrika*, *95*(2), 307–323. https://doi.org/10.1093/biomet/asn012
Eaton M. L., & Freedman D. A. (2004). Dutch book against some objective priors. *Bernoulli*, *10*(5), 861–872. https://doi.org/10.3150/bj/1099579159
Efron B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552
Ericson W. A. (1969). Subjective Bayesian models in sampling finite populations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *31*(2), 195–224. https://doi.org/10.1111/j.2517-6161.1969.tb00782.x
Escobar M. D., & West M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*(430), 577–588. https://doi.org/10.1080/01621459.1995.10476550
Fong E., & Holmes C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, *107*(2), 489–496. https://doi.org/10.1093/biomet/asz077
Fong E., Lyddon S., & Holmes C. (2019). Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. In *Proceedings of the 36th International Conference on Machine Learning* (pp. 1952–1962). PMLR.
Fortini S., Ladelli L., & Regazzini E. (2000). Exchangeability, predictive distributions and parametric models. *Sankhyā: The Indian Journal of Statistics, Series A*, *62*(Pt. 1), 86–109.
Fortini S., & Petrone S. (2012). Predictive construction of priors in Bayesian nonparametrics. *Brazilian Journal of Probability and Statistics*, *26*(4), 423–449. https://doi.org/10.1214/11-BJPS176
Fortini S., & Petrone S. (2014). Predictive distribution (de Finetti's view). In *Wiley StatsRef: Statistics reference online* (pp. 1–9). Wiley.
Fortini S., & Petrone S. (2020). Quasi-Bayes properties of a procedure for sequential learning in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(4), 1087–1114. https://doi.org/10.1111/rssb.12385
Frostig R., Johnson M. J., & Leary C. (2018). Compiling machine learning programs via high-level tracing. *Systems for Machine Learning*, *4*(9).
Geisser S. (1974). A predictive approach to the random effect model. *Biometrika*, *61*(1), 101–107. https://doi.org/10.1093/biomet/61.1.101
Geisser S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, *70*(350), 320–328. https://doi.org/10.1080/01621459.1975.10479865
Geisser S. (1982). Aspects of the predictive and estimative approaches in the determination of probabilities. *Biometrics*, *38*, 75–85. https://doi.org/10.2307/2529856
Geisser S. (1983). *On the prediction of observables: A selective update* (Technical Report). University of Minnesota.
Gelman A., Hwang J., & Vehtari A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2
Ghosal S., & van der Vaart A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.
Ghosh M., & Meeden G. (1997). *Bayesian methods for finite population sampling*. (Vol. *79*). CRC Press.
Gneiting T., & Raftery A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437
Hahn P. R. (2015). *Predictivist Bayes density estimation* (Technical Report). Arizona State University.
Hahn P. R., Martin R., & Walker S. G. (2018). On recursive Bayesian predictive distributions. *Journal of the American Statistical Association*, *113*(523), 1085–1093. https://doi.org/10.1080/01621459.2017.1304219
Hannah L. A., Blei D. M., & Powell W. B. (2011). Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, *12*, 1923–1953.
Heath D., & Sudderth W. (1978). On finitely additive priors, coherence, and extended admissibility. *The Annals of Statistics*, *6*(2), 333–345. https://doi.org/10.1214/aos/1176344128
Hewitt E., & Savage L. J. (1955). Symmetric measures on cartesian products. *Transactions of the American Mathematical Society*, *80*(2), 470–501. https://doi.org/10.1090/S0002-9947-1955-0076206-8

Huber P. J. (2004). *Robust statistics*. (Vol. *523*). John Wiley & Sons.
Jin Z., Ying Z., & Wei L. (2001). A simple resampling method by perturbing the minimand. *Biometrika*, *88*(2), 381–390. https://doi.org/10.1093/biomet/88.2.381
Joe H., & Xu J. J. (1996). *The estimation method of inference functions for margins for multivariate models* (Technical Report). University of British Columbia.
Kallenberg O. (1988). Spreading and predictable sampling in exchangeable sequences and processes. *The Annals of Probability*, *16*(2), 508–534. https://doi.org/10.1214/aop/1176991771
Kallenberg O. (1997). *Foundations of modern probability*. (Vol. *2*). Springer.
Knoblauch J., & Vomfell L. (2020). Robust Bayesian inference for discrete outcomes with the total variation distance. arXiv preprint arXiv:2010.13456.
Kushner H., & Yin G. G. (2003). *Stochastic approximation and recursive algorithms and applications*. (Vol. *35*). Springer Science & Business Media.
Lane D. A., & Sudderth W. D. (1984). Coherent predictive inference. *Sankhyā: The Indian Journal of Statistics, Series A*, *46*(2), 166–185.
Lauritzen S. L. (1988). *Extremal families and systems of sufficient statistics*. (Vol. *49*). Springer Science & Business Media.
Lijoi A., Prünster I., & Walker S. G. (2004). Extending Doob's consistency theorem to nonparametric densities. *Bernoulli*, *10*(4), 651–663. https://doi.org/10.3150/bj/1093265634
Lo A. Y. (1987). A large sample study of the Bayesian bootstrap. *The Annals of Statistics*, *15*(1), 360–375. https://doi.org/10.1214/aos/1176350271
Lo A. Y. (1988). A Bayesian bootstrap for a finite population. *The Annals of Statistics*, *16*(4), 1684–1695. https://doi.org/10.1214/aos/1176351061
Lyddon S., Walker S., & Holmes C. C. (2018). Nonparametric learning from Bayesian models with randomized objective functions. In *Advances in neural information processing systems 31* (pp. 2075–2085). Curran Associates, Inc.
Lyddon S. P., Holmes C. C., & Walker S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, *106*(2), 465–478. https://doi.org/10.1093/biomet/asz006
MacEachern S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation*, *23*(3), 727–741. https://doi.org/10.1080/03610919408813196
MacEachern S. N. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science* (Vol. 1, pp. 50–55). American Statistical Association.
Martin R. (2021). A survey of nonparametric mixing density estimation via the predictive recursion algorithm. *Sankhya B*, *83*, 97–121. https://doi.org/10.1007/s13571-019-00206-w
McDiarmid C. (1998). Concentration. In *Probabilistic methods for algorithmic discrete mathematics* (pp. 195–248). Springer.
Muliere P., & Secchi P. (1996). Bayesian nonparametric predictive inference and bootstrap techniques. *Annals of the Institute of Statistical Mathematics*, *48*(4), 663–673. https://doi.org/10.1007/BF00052326
Muliere P., & Walker S. (2000). Neutral to the right processes from a predictive perspective: A review and new developments. *Metron*, *58*(3–4), 13–30.
Müller P., Erkanli A., & West M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, *83*(1), 67–79. https://doi.org/10.1093/biomet/83.1.67
Neal R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265.https://doi.org/10.2307/1390653
Nelsen R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.
Newton M. A., Polson N. G., & Xu J. (2020). Weighted Bayesian bootstrap for scalable posterior distributions. *Canadian Journal of Statistics*, *49*(2), 421–437. https://doi.org/10.1002/cjs.11570
Newton M. A., Quintana F. A., & Zhang Y. (1998). Nonparametric Bayes methods using predictive updating. In *Practical nonparametric and semiparametric Bayesian statistics* (pp. 45–61). Springer.
Newton M. A., & Raftery A. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *56*(1), 3–48. https://doi.org/10.1111/j.2517-6161.1994.tb01956.x
Ng T. L., & Newton M. A. (2022). Random weighting in LASSO regression. *Electronic Journal of Statistics*, *16*(1), 3430–3481. https://doi.org/10.1214/22-EJS2020
Nie L., & Ročková V. (2023). Bayesian bootstrap spike-and-slab LASSO. *Journal of the American Statistical Association*, *118*(543), 2013–2028. https://doi.org/10.1080/01621459.2022.2025815
Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., & Duchesnay E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.
Quintana F. A., Mueller P., Jara A., & MacEachern S. N. (2022). The dependent Dirichlet process and related models. *Statistical Science*, *37*(1), 24–41. https://doi.org/10.1214/20-STS819

Rasmussen C. E. (2003). Gaussian processes in machine learning. In *Summer School on Machine Learning*, (pp. 63–71). Springer.
Robbins H., & Siegmund D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics* (pp. 233–257). Elsevier.
Roberts H. V. (1965). Probabilistic prediction. *Journal of the American Statistical Association*, *60*(309), 50–62. https://doi.org/10.1080/01621459.1965.10480774
Robins J., & Wasserman L. (2000). Conditioning, likelihood, and coherence: A review of some foundational concepts. *Journal of the American Statistical Association*, *95*(452), 1340–1346. https://doi.org/10.1080/01621459.2000.10474344
Roeder K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, *85*(411), 617–624. https://doi.org/10.1080/01621459.1990.10474918
Ross G. J., & Markwick D. (2018). dirichletprocess: An R package for fitting complex Bayesian nonparametric models. https://cran.r-project.org/web/packages/dirichletprocess/index.html
Rubin D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350
Rubin D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics*, *9*(1), 130–134. https://doi.org/10.1214/aos/1176345338
Rubin D. B. (2004). *Multiple imputation for nonresponse in surveys*. (Vol. *81*). John Wiley & Sons.
Saarela O., Stephens D. A., Moodie E. E., & Klein M. B. (2015). On Bayesian estimation of marginal structural models. *Biometrics*, *71*(2), 279–288. https://doi.org/10.1111/biom.12269
Saville B. R., Connor J. T., Ayers G. D., & Alvarez J. (2014). The utility of Bayesian predictive probabilities for interim monitoring of clinical trials. *Clinical Trials*, *11*(4), 485–493. https://doi.org/10.1177/1740774514531352
Shahbaba B., & Neal R. (2009). Nonlinear models using Dirichlet process mixtures. *Journal of Machine Learning Research*, *10*(63), 1829–1850.
Sklar A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, *8*, 229–231.
Stone M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *36*(2), 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x
Tang Y., Salakhutdinov R., & Hinton G. (2012). Deep mixtures of factor analysers. In *Proceedings of the 29th International Conference on Machine Learning* (pp. 505–512). Omnipress.
Tokdar S. T., Martin R., & Ghosh J. K. (2009). Consistency of a recursive estimate of mixing distributions. *The Annals of Statistics*, *37*(5A), 2502–2522. https://doi.org/10.1214/08-AOS639
Vehtari A., & Lampinen J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, *14*(10), 2439–2468. https://doi.org/10.1162/08997660260293292
Wade S. (2013). *Bayesian nonparametric regression through mixture models* [PhD thesis]. Bocconi University.
Wade S., Walker S. G., & Petrone S. (2014). A predictive study of Dirichlet process mixture models for curve fitting. *Scandinavian Journal of Statistics*, *41*(3), 580–605. https://doi.org/10.1111/sjos.12047
Walker S. G. (2013). Bayesian inference with misspecified models. *Journal of Statistical Planning and Inference*, *143*(10), 1621–1633. https://doi.org/10.1016/j.jspi.2013.05.013
Wang Z., & Scott D. W. (2019). Nonparametric density estimation for high-dimensional data–algorithms and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, *11*(4), e1461.https://doi.org/10.1002/wics.1461
Wasserman L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.
Waudby-Smith I., & Ramdas A. (2023). Estimating means of bounded random variables by betting . *Journal of the Royal Statistical Society Series B: Statistical Methodology*. https://doi.org/10.1093/jrsssb/qkad009
West M. (1991). Kernel density estimation and marginalization consistency. *Biometrika*, *78*(2), 421–425. https://doi.org/10.1093/biomet/78.2.421
Zabell S. L. (1982). WE Johnson's 'sufficientness' postulate. *The Annals of Statistics*, *10*(4), 1090–1099. https://doi.org/10.1214/aos/1176345975

Journal of the Royal Statistical Society Series B: Statistical Methodology, 2023, **85**, 1392–1416

ROYAL STATISTICAL SOCIETY DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society Series B Statistical Methodology

Discussion Paper Contribution

Proposer of the vote of thanks to Fong, Holmes and Walker and contribution to the Discussion of 'Martingale Posterior Distributions'

Philip Dawid

Statistical Laboratory, University of Cambridge, Cambridge, UK

Address for correspondence: Philip Dawid, Statistical Laboratory, University of Cambridge, Cambridge, UK.
Email: apd@statslab.cam.ac.uk

This is a fascinating paper, both wide and deep. I have been intrigued by it, and it has stimulated many comments and questions.

Savage (1961) described Fisher's fiducial argument as 'a bold attempt to make the Bayesian omelet without breaking the Bayesian eggs'. Tonight's paper appears to cook up a posterior omelet without using any eggs at all—no statistical model for the data nor any prior distribution are required. This is a remarkable achievement!

While it is supposed that the data $y_{1:n} = (y_1, \ldots, y_n)$ have arisen from an exchangeable distribution, this property is not used. The only 'modelling' done is the construction of a joint distribution for never-to-be-observed future values $Y_{n+1:\infty}$. This involves an initial choice for the forecast distribution P_n for Y_{n+1}, and a choice of updating method to move from P_i to P_{i+1} $(i \geq n)$, taking into account a value Y_{i+1} simulated from P_i. However, it is not clear how these choices, and the implied 'posterior' inference, depend on the observed data $y_{1:n}$, nor what general principles and pragmatic considerations might inform them.

We are told that these 'predictive models are probabilistic statements on observables'—but the observable (indeed, observed) quantities are $Y_{1:n}$, which are not modelled. In contrast, the modelled $Y_{n+1:\infty}$ are fictions of the authors' imagining. Different specifications of the one-step-ahead predictive distributions (all satisfying the martingale property) will lead to different distributions for $Y_{n+1:\infty}$, and so to different 'posteriors'.

While considerable attention is given to the updating process, less is said about the crucial choice of the initialising P_n. One possibility is to use the empirical distribution of the data, as in the process leading to the Bayesian bootstrap. Another, briefly mentioned in Section 4.5.1, arrives at P_n by starting from an assessed prior predictive P_0 for Y_1, and using the chosen updating formula with the observed data. The sequence $P_0, P_1 \ldots, P_{n-1}$ so generated can be considered as a 'prequential model' (Dawid, 1991) for $Y_1, \ldots, Y_n$. We can then apply various prequential tests of the compatibility of this model with the observed data $y_1, \ldots, y_n$ (Dawid, 1992): for example, we can test whether the successive conditional probability integral transforms $P_0(y_1), \ldots, P_{n-1}(y_n)$ look like a random sample from the uniform distribution on [0, 1]. When such a compatibility test is failed, that is evidence that our updating formula is not a good match to the observed data, and should disqualify its further use. (Note that forecasting using the empirical distribution would fail an obvious compatibility test as soon as a new observation differed from all earlier values.)

Received: August 17, 2023. **Accepted:** August 23, 2023

In Section 4.5.2, the authors maximise the prequential likelihood $p_0(y_1)\ldots p_{n-1}(y_n)$ to estimate the bandwidth parameter ρ. They then fix this value for the future simulations. A fully prequential approach would re-estimate ρ after each new simulated value y_i, and insert that into the update for P_i.

Having settled on P_n, and an update scheme with its c.i.d. structure, we obtain the limiting predictive distribution P_∞, living in the tail σ-field of the sequence $Y_{n+1:\infty}$. This can be regarded as the parameter of an implied statistical model for $Y_{n+1:\infty}$, obtained by conditioning on it. But because the constructed distribution is not exchangeable, this is not a model for i.i.d. variables. It would be good to understand the actual structure of this implied model.

The development in this paper is very closely tied to the exchangeability assumption, and its c.i.d. generalisation which is required for the martingale property. So, how could it be extended to more complex models? Consider for example the Markov AR(1) model with $Y_{i+1} \mid Y_{1:i}, \theta \sim N(\theta Y_i, 1)$, mixed over a smooth positive prior distribution for θ over (–1, 1). The predictive distribution of Y_{i+1} is approximately

$$N\left(y_i\widehat{\theta}_i,\, 1 + \frac{y_i^2}{\sum_{j=1}^{i} y_{j-1}^2}\right),$$

with $\widehat{\theta}_i = (\sum_{j=1}^{i} y_j y_{j-1})/(\sum_{j=1}^{i} y_{j-1}^2)$, and is no longer Markov. Although the joint distribution is not exchangeable or c.i.d., the model parameter θ is recovered as $\lim_{i\to\infty} \widehat{\theta}_i$, and the initial Markov model by conditioning on that. In general, what properties of a joint or sequentially specified distribution for $Y_{1:\infty}$ would be required for consistency with a Markov model, and how might one approximate that with sequentially updated predictive distributions?

Another approach might be to update the conditional distribution of Y_i given Y_{i-1}, *e.g.* using methods like those in Section 4.4. What then would be the implied parameter and statistical model?

The authors have done a great job in developing and presenting this work, and I am delighted to propose a hearty vote of thanks to them.

Conflict of interest: None declared.

References

Dawid A. P. (1991). Fisherian inference in likelihood and prequential frames of reference (with discussion). *Journal of the Royal Statistical Society, Series B*, *53*(1), 79–109. http://dx.doi.org/10.1111/j.2517-6161.1991.tb01810.x

Dawid A. P. (1992). Prequential data analysis. In M. Ghosh & P. K. Pathak (Eds.), *Current Issues in Statistical Inference: Essays in Honor of D. Basu*, IMS Lecture Notes–Monograph Series. (Vol. 17, pp. 113–126). Institute of Mathematical Statistics. https://doi.org/10.1214/lnms/1215458842

Savage L. J. (1961). The foundations of statistics reconsidered. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 575–586). University of California Press.

https://doi.org/10.1093/jrsssb/qkad086
Advance access publication 27 August 2023

Seconder of the vote of thanks to Fong, Holmes and Walker and contribution to the Discussion of 'Martingale Posterior Distributions'

Steffen Lauritzen

Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark

Address for correspondence: Steffen Lauritzen, Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark. Email: lauritzen@math.ku.dk

Let me first congratulate the authors for an interesting, thought provoking and inspiring article, with many fine examples. However, I wonder whether the specific setup has sufficient generality to cover interesting cases, and I should have liked to see the ideas in this paper confronted with some different situations.

Firstly, let me remind everyone of the standard Bayesian setup, where we have *observables* X, Y, a *parameter* Θ of interest that in principle should be added to or be a function of the observables, and a Fisherian model, specifying the conditional distribution $P\{(X, Y) \in A \times B \mid \Theta = \theta\}$. A Bayesian model will also specify a prior distribution π of Θ, hence also a joint distribution $P\{(X, Y, \Theta) \in A \times B \times C\}$. Bayesian inference after observation of $X = x$ will now calculate the posterior distribution $P\{\Theta \in C \mid X = x\}$ and/or the predictive distribution $P\{Y \in B \mid X = x\}$.

Note in particular that this setup is *universal* and applies to almost any thinkable statistical problem, whereas the article specialises to a setting with $X = (X_1, \ldots, X_n)$ and $Y = X_{n+1}, X_{n+2} \ldots$ (asymptotically) exchangeable, so

$$\Theta = \Theta(A) = \lim_{N\to\infty} \frac{\sum_{k=1}^{n} \mathbf{1}_A(X_{n+k})}{N}$$

or some variant thereof. The article then circumvents specifying prior and posterior and goes directly to the predictive.

To highlight some of the issues I am thinking of, let us consider the pure birth process $(X_t, t > 0)$ specified by letting $X_0 = 1$ and for $t > 0$

$$P\{X_{t+h} = j \mid X_t = i, \Lambda = \lambda\} = \begin{cases} i\lambda h + o(h) & (j = i + 1) \\ 1 - i\lambda h + o(h) & (j = i) \\ o(h) & \text{otherwise.} \end{cases}$$

To make a full Bayesian specification, we add a prior exponential distribution for the unknown parameter $\Lambda \sim \exp(1)$.

We now have the following facts, see for example Keiding (1974): Observe X on interval $[0, t]$ and let $S_t = \int_0^t X_u \, du$. Then, almost surely:

$$\lim_{t\to\infty} X_t e^{-\Lambda t} = W; \quad W \mid \Lambda \sim \exp(1) \text{ so } \Lambda \perp\!\!\!\perp W; \quad \lim_{t\to\infty} S_t e^{-\Lambda t} = W/\Lambda.$$

Conditional on $W = w$, the process X_t, $t > 0$ behaves like an inhomogeneous Poisson process (Kendall, 1949, 1966) with intensity $\mu(t) = w\lambda e^{\lambda t}$. Hence, $\log X_t$ grows linearly as

$$\log X_t \sim \log \lambda + \log w + \lambda t$$

so W determines the intercept at 0.

We now have a choice and could either think of Λ or the pair (W, Λ) as the parameter of interest. In both cases, the parameter would be a function of the data for an infinite sample size; the last parameter would give a more detailed description of the behaviour as it will not just give the slope but also the approximate intercept of $(\log X_t, t \to \infty)$. In the first case, the log-likelihood function becomes

$$\ell(\lambda) \sim (x_t - 1)\log\lambda - \lambda s_t,$$

and the MLE is $\hat{\lambda} = (X_t - 1)/S_t$. In the second case, the log-likelihood function becomes

$$\ell(\lambda, w) \sim (x_t - 1)(\log\lambda + \log w) - \lambda(tx_t - s_t) + w(1 - e^{-\lambda t})$$

and the MLE is more complicated and may not exist, for example if the observed growth curve is concave. In both cases, (X_t, S_t) is minimal sufficient.

The predictive distribution for (X_{t+h}, S_{t+h}) given $\Lambda = \lambda$ and X_u, $u \in [0, t]$ has density with respect to product of counting measure and Lebesgue measure (Keiding, 1974):

$$f_{t,t+h}(x, s \mid X_u = x_u, 0 \le u \le t, \lambda) = \binom{x-1}{x_t-1}(\lambda h)^{x-x_t} e^{-\lambda(s-s_t)} g_{x,x_t}(s - s_t),$$

where $g_{x,x_t}(s - s_t)$ is explicit and does not contain unknown parameters. This yields the Bayesian predictive distribution when $\Lambda \sim \exp(1)$ as

$$f_{t,t+h}(x, s \mid X_u = x_u, 0 \le u \le t) = \binom{x-1}{x_t-1} h^{x-x_t} \frac{\Gamma(x - x_t + 1)}{(s - s_t + 1)^{a+1}} g_{x,x_t}(s - s_t).$$

This last predictive distribution defines a 'martingale posterior' using that the process of sufficient statistics (X_u, S_u), $u > t$ is a Markov process. Indeed, as exploited by Doob (1949), the sequence of posterior distributions is always a martingale.

Using the idea of today's article, one could simulate from the predictive distribution and define the estimates via the simulated sample $(X_u, t < u < T)$ by using maximum likelihood on the outcome, hence letting

$$\hat{\lambda}_T = (X_T - 1)/S_T$$

or, ignoring the first part of the sample,

$$\tilde{\lambda}_T = (X_T - X_t)/(S_T - S_t)$$

or, in principle:

$$\breve{\lambda} = \lim_{T\to\infty} \hat{\lambda}_T.$$

However, based on the same predictive distribution, one could also wish to estimate W. Or even extend the model by allowing negative λ values using an inhomogeneous Poisson model with *negative* growth rate λ; this would then accommodate data showing a concave growth curve and give a slightly different estimate $\hat{\lambda}_T$.

But W, Λ are functions of the infinite sample if and only if $\lambda > 0$; Then, the inhomogeneous Poisson model is an *extreme point model* (Lauritzen, 1988), but not otherwise.

In any case, it seems hard to invent the predictive distribution above without going through the standard Bayesian approach so maybe the predictive approach is not so helpful after all?

Is there a potential advantage in using the martingale posterior framework for describing the uncertainty of W by simulation from the predictive rather than the posterior distribution? In any case, it is my absolute pleasure to second the vote of thanks for this interesting article.

Conflict of interest: None declared.

References

Doob J. L. (1949). Application of the theory of martingales. *Colloques Internationaux du Centre National de la Recherche Scientifique, Paris*, *13*, 22–27.

Keiding N. (1974). Estimation in the birth process. *Biometrika*, *61*(1), 71–80. https://doi.org/10.1093/biomet/61.1.71

Kendall D. G. (1949). Stochastic population growth (with discussion). *Journal of the Royal Statistical Society, Series B*, *11*, 230–264.

Kendall D. G. (1966). Branching processes since 1873. *Journal of London Mathematical Society*, *s1-41*(1), 385–406. https://doi.org/10.1112/jlms/s1-41.1.385

Lauritzen S. L.(1988). *Extremal families and systems of sufficient statistics*. (Vol. *49*). Lecture Notes in Statistics. Springer.

The vote of thanks was passed by acclamation.

https://doi.org/10.1093/jrsssb/qkad087
Advance access publication 5 September 2023

Blake Moya's contribution to the Discussion of 'Martingale Posterior Distributions' by Fong, Holmes and Walker

Blake Moya

Department of Statistics and Data Science, University of Texas at Austin, Austin, TX, USA

Address for correspondence: Blake Moya, 105 E 24th St D9800, Welch 5.216, Austin, TX 78705, USA. Email: blakemoya@utexas.edu

I would like to congratulate the authors on an expository and intuitive representation of statistical uncertainty. To assist in the further investigation of predictive resampling techniques, I have developed a software package for R (R Core Team, 2022) which implements some of the algorithms presented here as well as from subsequent work (Moya & Walker, 2023). The CopRe package (Moya, 2022), named for the copula resampling technique described in Section 4, can be installed from the Comprehensive R Archive Network with the command: install.packages('copre').

The copula resampling algorithm is massively parallelisable, and the simplicity of each recursive update makes implementation in very low-level programming languages quite easy. I have developed CopRe's core code in C++ (ISO, 2012), parallelised with OpenMP (Chandra et al., 2001). I have also written core code in CUDA (NVIDIA et al., 2020) for running the algorithm on a GPU that is available upon request. A comparison of the running time for the marginal Dirichlet Process Mixture Model (DPMM) sampler of Escobar and West (1994) and Copula Resampling run in serial or parallelised over a CPU or a GPU is shown in Figure 1. The acceleration of nonparametric Bayesian inference presented by the authors is significant.

By imposing prior information on the mechanics of the data-generating process rather than on its parameters, predictive resampling of martingale posteriors overcomes many of the difficulties involved with the creation of nonparametric priors, and the implementations of Markov chain Monte Carlo samplers for their corresponding posteriors. The current development version of CopRe contains a new Gibbs-type sequence resampling function, SeqRe, which exploits the known predictive update rule of many Gibbs-type priors to sample full random distributions from mixture models like the DPMM without a known representation of the prior or posterior on the random measure. The development version of the package containing new experimental

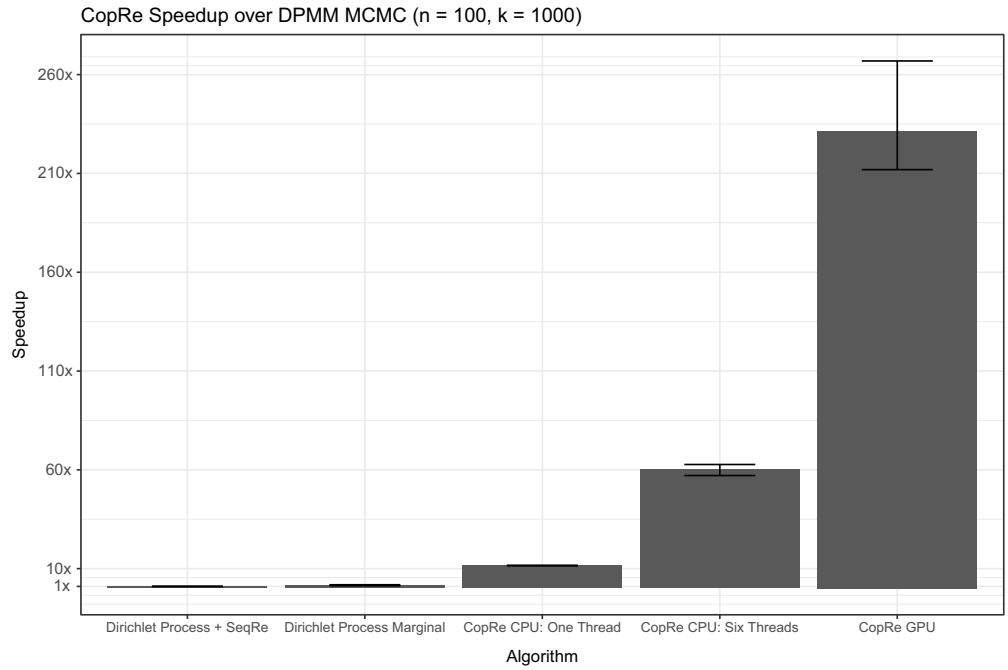

Figure 1. Speedup over the Dirichlet Process Mixture MCMC sampler of Escobar and West (1994) in concert with the sequence resampling approach of Moya and Walker (2023) for the MCMC sampler without resampling extension and three launch configurations for CopRe. The sample size was $n = 100$, $k = 1000$ samples were drawn from each algorithm, and for CopRe $N = 100$ recursive predictive draws were made for each sample. Computations were made with core C++/CUDA code on an Intel Core i5 8600 K clocked to 4.8 GHz and an NVIDIA GTX 1070 Ti.

features can also be installed via the command:

```
devtools::install_github('blakemoya/copre', ref = 'dev')
```

I encourage experimenters to take advantage of this software and hope that it will accelerate further investigation of martingale posteriors.

Conflict of interest: None declared.

References

Chandra R., Dagum L., Kohr D., Menon R., Maydan D., & McDonald J. (2001). *Parallel programming in OpenMP*. Morgan Kaufmann.

Escobar M., & West M. (1994). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*(430), 577–588. https://doi.org/10.1080/01621459.1995.10476550

ISO (2012), *ISO/IEC 14882:2011 Information technology—Programming languages—C++*. International Organization for Standardization, Geneva, Switzerland. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50372.

Moya B. (2022). *CopRe: Tools for nonparametric martingale posterior sampling*. [R package version 0.1.0].

Moya B., & Walker S. G. (2023). Full uncertainty analysis for Bayesian nonparametric mixture models. *Computational Statistics & Data Analysis*, 107838. https://doi.org/10.1016/j.csda.2023.107838

R Core Team (2022), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Vingelmann P., & Fitzek F. H., NVIDIA (2020). Cuda, release: 10.2.89. https://developer.nvidia.com/cuda-toolkit.

https://doi.org/10.1093/jrsssb/qkad088
Advance access publication 28 August 2023

Isadora Antoniano Villalobos's contribution to the Discussion of 'Martingale Posterior Distributions' by Fong, Holmes and Walker

Isadora Antoniano-Villalobos

Department of Environmental Sciences, Informatics of Statistics, Ca' Foscari University of Venice, Via Torino 155, 30172 Venice, VE, Italy

Address for correspondence: Isadora Antoniano-Villalobos, Department of Environmental Sciences, Informatics of Statistics, Ca' Foscari University of Venice, Via Torino 155, 30172 Venice, VE, Italy.
Email: isadora.antoniano@unive.it

I thank the authors for their refreshing ideas, providing many insights and intuitions on the process of statistical learning. For years, Bayesians have considered two levels of uncertainty, the likelihood and the prior. By shifting the focus to the predictive distribution as a single expression of uncertainty about a future observation (given data), the sequential nature of statistical learning is made explicit, together with some of its limitations. If each observation carries a piece of information, it follows that any question requiring more than (at most) a countable number of pieces cannot be answered by this process and must, rather, be assumed. This encapsulates in a single idea the need for smooth link functions and kernels, sparsity priors, and other tools used in different contexts to eliminate by design what cannot be learned from data. At the same time, assumptions such as exchangeability may lose relevance. For example, if one wished to learn the upper bound of a certain quantity, in a standard setup, repeated measurements would be assumed exchangeable. However, if measurements 1, 1.1, 2.3 were observed, how much we learn from the third observation (the change in predictive uncertainty) would arguably depend on the order in which the three values were observed. The symmetry relation between measurements is not preserved by the learning process. This indicates that model definition in terms of the predictive sequence, by asking a different question, may allow inference in the presence of more complex forms of data dependence.

The definition of a coherent predictive model appears challenging and the empirical distribution is the simplest possible choice. However, if only x_1 was available, setting $X_2 = x_1$ almost surely is hardly a reasonable way to model uncertainty. Mixture models with data-dependent weights and/or kernel parameters might be an alternative worth exploring. Rather than simply achieving continuity or greater flexibility, the choice of the kernel could be driven by the need to better extract (or filter) the information that each new observation contains about possible future samples. This idea is wonderfully illustrated by the bivariate copula construction proposed in Section 4. The copula can be interpreted as a measure of the 'speed' at which the predictive is updated (moves) towards its limit (ideally the distribution of interest) as new observations arrive. Studying the relationship between mixtures and copulas could provide a tool to elicit new (perhaps approximate but computationally efficient) predictive models.

Conflict of interest: Isadora Antoniano-Villalobos was a PhD student of Stephen G. Walker, with whom she has coauthored some publications.

https://doi.org/10.1093/jrsssb/qkad089
Advance access publication 31 August 2023

Ramses Mena Chavez's contribution to the Discussion of 'Martingale Posterior Distributions' by Fong, Holmes and Walker

Ramsés H. Mena

IIMAS, Universidad Nacional Autónoma de México, CDMX, Mexico

Address for correspondence: Ramsés H. Mena, IIMAS, Universidad Nacional Autónoma de México, A.P. 20-726, 01000 CDMX, Mexico. Email: ramses@sigma.iimas.unam.mx

I congratulate the authors for introducing a novel approach for Bayesian inference, which avoids directly using the posterior distribution, often not available or computationally expensive. In addition to opening several lines of thinking about the foundations of Bayesian statistics, their proposal unveils many interesting directions to be explored and applied within the world of statistics.

With an infinite population and the epistemic view of uncertainty (see, e.g. Goldstein, 2013), namely when the posterior represents the uncertainty due to missing observations, exchangeability arises naturally. Here, by applying a martingale convergence result by Doob (1949), the authors show that the predictive view of Bayesian inference allows one to obtain the posterior distribution, provided the statistic, θ_n, summarising observations, $y_{1:n}$, is a martingale. This requirement might indeed lead to other types of symmetries, i.e. not necessarily exchangeability.

Within the exchangeability setting, the proposed predictive approach resembles much to that frequently used in Bayesian non-parametrics, e.g. via the Blackwell and MacQueen Pólya-Urn and many of its generalisations used for posterior inference. Hence, with the findings in this paper, one wonders whether the exchangeability requirement is the most practical one. Furthermore, there might be other ways to relax the martingale condition, namely throughout other convergence requirements for θ_n.

Clearly, one of the appealing features of the proposed approach is bypassing the prior to posterior computation, by suitably modelling predictive distributions of conditionally identically distributed (c.i.d.) sequences and θ_n. For some of the examples in the paper, it is clear what the predictive structure should look like, and though the authors propose a fairly general predictive using their copula approach, there are many inference scenarios where one would need to be very creative to achieve the desired inference, for instance Bayesian inference for phenomena typically modelled throughout continuous time Markov processes.

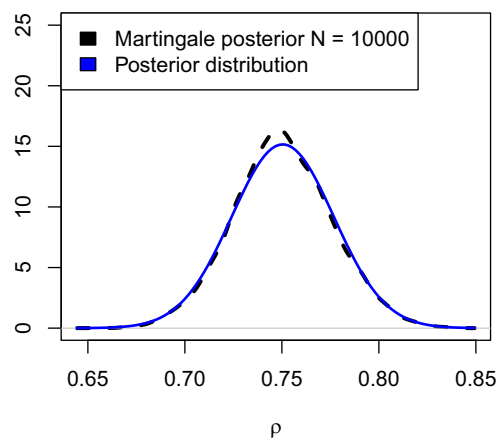

Figure 1. Martingale posterior and posterior distribution for ρ.

For an Autoregressive of order one, AR(1), model with transition density $p(x_i \mid x_{i-1}; \rho) = \mathsf{N}(x_i \mid \rho x_{i-1}, 1)$, and ρ as the parameter of interest, one could use the predictive

$$p(x_{n+1} \mid x_{1:n}) = \mathsf{N}\left(x_{n+1} \mid x_n \frac{a}{b}, \frac{x_n^2 + b}{b}\right),$$

where $a := \sum_{i=2}^{n} x_i x_{i-1}$ and $b := 1 + \sum_{i=2}^{n} x_{i-1}^2$, with functional statistic $\bar{\rho}_N = a/b$. This proposal comes from assuming an $\mathsf{N}(0, 1)$ prior for ρ and computing the predictive. However, without passing throughout such mechanism, a natural question is how could one propose such predictive and the form to be used for $\bar{\rho}_N$. Indeed, the authors proposal works very well, see Figure 1.

Conflict of interest: None declared.

References

Doob J. L. (1949). Application of the theory of martingales. In *Actes du colloque international le calcul des probabilites et ses applications*, Lyon, 28 Juin–3 Juillet 1948 (pp. 23–27).

Goldstein M. (2013). Observables and models: Exchangeability and the inductive argument. In P. Damien, P. Dellaportas, N. G. Polson, & D. A. Stephens (Eds.), *Bayesian theory and applications* (pp. 3–18). Oxford University Press.

The following contributions were received in writing after the meeting.

https://doi.org/10.1093/jrsssb/qkad100
Advance access publication 29 August 2023

Bertrand Clarke's contribution to the Discussion of 'Martingale Posterior Distributions' by Fong, Holmes and Walker

Bertrand Clarke

Department of Statistics, University of Nebraska-Lincoln, Lincoln, NE, USA

Address for correspondence: Bertrand Clarke, Department of Statistics, University of Nebraska-Lincoln, 340 Hardin Hall North, Lincoln, NE 68583-0963, USA. Email: bclarke3@unl.edu

To see the effect of the paper on statistical practice, consider the following hypothetical dialogue between a Quantitative Scientist (QS) and a Predictive Statistician (PS).

QS: I've been thinking about the data generating mechanism and what we can safely assume pre-experimentally.

PS: Actually, I'm more interested in what you think your $n + 1$ data point would be. If you were going to predict your $n + 1$ measurement, how would you do it?

QS: I'd guess what the true model is and take its mean or median. But, I don't know what the true model is.

PS: But you do know what the predictive density $m(y_{n+1} \mid y_{1:n})$ is, right?

QS: If I assume my likelihood and prior are right. But, I don't know that.

PS: Do you believe there is an overall probability model for $Y_{1:n+1}$?

QS: Sure: M.

PS: Then, we can use De Finetti to backform a likelihood and prior if we want. But let's not. Let's think about the martingale model $m(y_{n+1} \mid y_{1:n})$, $m(y_{n+2} \mid y_{1:n+1})$ and so on.

QS: You mean what on average a new random draw would look like given the past data? If I'd seen $y_{1:n+k}$ I'd draw Y_{n+k+1} from some $m(\cdot \mid \bar{y}_{n+k})$. I'd start with a density shaped like this [draws a density] and let it tighten a bit as n increases.

PS: So that's step k in your martingale model. We could use it, but let's not. Instead, I'm going to form one-step ahead predictive DF's using $y_{1:n}$, $y_{1:n+1}$, and so forth. I'm going to draw one data point from each DF to find the next. For each step up to N, I'll find a version of the posterior mean. I'll do this many times and then use the final posterior means to estimate $W(\cdot \mid y_{1:n})$.

QS: You're using repeated generation of data I don't have to find the posterior I do have?

PS: Yes, I'm completing your experiment conceptually by filling out your missing data. Over and over. Then, I can quantify the information you don't have and derive the variability left in data you do have i.e. form the posterior.

QS: And you're ignoring everything we know about the data generator besides the data. Where does P_θ come in?

PS: It doesn't. We're relying on martingale convergence in the mixture M. By changing the probability defining the mode of convergence we get the limits we want for $\bar{\Theta}$ and the DF—that's why we can use them to estimate the posterior.

QS: So, what role do the prior and likelihood—or even conditional mixtures—have anymore? If you're right, why would we bother with them? Some sort of robustness analysis?

PS: We could. But, we let's not. Let's think about the predictive process directly.

QS: What's the magic?

PS: No magic. We just changed the mode from P_θ to M since we're thinking about the whole countably infinite product space $\mathcal{Y}^\infty$.

QS: Is θ still a parameter?

PS: It's a function value. Take $\Theta_n = \Theta_n(Y_{1:n})$ and think of the posterior mean. Or better take $n = \infty$. Each $\Theta = \theta$ has a $V_\theta = \{y_{1:\infty} \mid \theta(y^\infty) = \theta\}$.

QS: Don't θ and a $y_{1:\infty}$ that I might conceptually get have to be compatible in that if $y_{1:\infty} \notin V_\theta$ then θ can't be true?

PS: Yes. If you're looking for magic, it's that at infinity $\Theta = \Theta(Y_{1:\infty}) = E(\Theta \mid Y_{1:\infty})$ pointwise in $\mathcal{Y}^\infty$ and as random variables. Then the $y_{1:\infty}$ you get gives you your θ. So, in the limit, our computational procedure mimics what you would get if you had $y_{1:\infty}$ —but, we don't. What you don't sample gives an assessment of your uncertainty.

QS: But I'm really using a sequence of predictive DF's not posterior means or conditional predictives.

PS: Yes. That way you get objectivity in your predictive thinking.

Conflict of interest: None.

https://doi.org/10.1093/jrsssb/qkad090
Advance access publication 27 August 2023

David Draper and Erdong Guo's contribution to the discussion of 'Martingale posterior distributions', by Fong, Holmes and Walker

David Draper and Erdong Guo

Department of Statistics, Baskin School of Engineering, University of California, Santa Cruz, CA, USA

Address for correspondence: David Draper, Department of Statistics, Baskin School of Engineering, University of California, 1156 High Street, Santa Cruz, CA 95064, USA. Email: draper@ucsc.edu

We have two comments motivated by this interesting paper.

1. The idea, introduced early in the paper, that 'the object of interest is fully defined once all the observations have been viewed' is almost exactly 100 years old: it was a cornerstone of the remarkable paper by Fisher (1922) and has been referred to for many decades as *Fisher consistency*. We are surprised that the authors did not make this connection.
2. The authors make strong distinctions between the frequentist and Bayesian bootstraps. We would like to point out the not-so-widely-known fact that *the frequentist bootstrap is actually an instance of Bayesian nonparametric inference*, as follows. Suppose that the context $\mathbb{C}$ of the problem under study by You (Good, 1950: a person wishing to reason sensibly in the presence of uncertainty) implies that Your uncertainty about real-valued observables $\{Y_1, Y_2, \ldots\}$, which have not yet been observed, is exchangeable. Then, de Finetti's Representation Theorem for real-valued outcomes tells us that this is equivalent to the Bayesian hierarchical model

$$\begin{gathered}(F \mid \mathcal{B}^*) \sim p(F \mid \mathcal{B}^*) \\ \left\{ \begin{array}{c} (Y_i \mid F \mathcal{B}^*) \\ (i = 1, \ldots, n) \end{array} \right\} \overset{\text{IID}}{\sim} F, \end{gathered} \tag{1}$$

in which F is the empirical CDF based on $\{Y_1, Y_2, \ldots\}$, n is a finite positive integer, and $\mathcal{B}^*$ is a finite set of propositions, all rendered true by context $\mathbb{C}$ and exhaustive of all relevant contextual information. As is well known, (a) the conjugate prior for F in this model is the family $DP(\alpha, F_0)$ of Dirichlet processes, where $\alpha > 0$ and F_0 represent the appropriate prior sample size and prior estimate of F, respectively, based on Your information external to the observed data set $\boldsymbol{y} = (y_1, \ldots, y_n)$, and (b) conjugate updating yields the posterior

$$DP\left(\alpha + n, \frac{\alpha F_0 + n\hat{F}_n}{\alpha + n}\right) \tag{2}$$

for F, in which $\hat{F}_n$ is the empirical CDF based on $\boldsymbol{y}$. To create a low-information prior, it is tempting to send $\alpha \downarrow 0$; Terenin and Draper (2017) have shown that this is mathematically meaningful, with the resulting prior, which they call $DP(0)$, yielding the important-for-statistical-science posterior $DP(n, \hat{F}_n)$. A corollary of a result in Terenin et al. (2018) then yields the following theorem, stated informally:

Theorem (Draper & Guo, 2023) Under the conditions detailed above, frequentist bootstrap samples of size n from $(y_1, \ldots, y_n)$ are asymptotically stochastically indistinguishable from stick-breaking samples of the same size from $DP(n, \hat{F}_n)$.

We find empirically that the frequentist bootstrap approximation is good to excellent even for n as small as 25; this has useful implications for high-quality Bayesian data science.

Conflict of interests: None declared.

References

Draper D., & Guo E. (2023). *Optimal Bayesian analysis of A/B test results in big-data settings: The frequentist bootstrap is actually an instance of Bayesian nonparametric inference* [in preparation].

Fisher R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, *222*(594–604), 309–368. https://doi.org/10.1098/rsta.1922.0009

Good I. J. (1950). *Probability and the weighing of evidence*. Griffin.

Terenin A., & Draper D. (2017). A noninformative prior on a space of distribution functions. *Entropy*, *19*(8), 391. https://doi.org/10.3390/e19080391

Terenin A., Magnusson M., Jonsson L., & Draper D. (2018). Pólya Urn Latent Dirichlet Allocation: A doubly sparse massively parallel sampler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(7), 1709–1719. https://doi.org/10.1109/TPAMI.2018.2832641

https://doi.org/10.1093/jrsssb/qkad091
Advance access publication 5 September 2023

Jiaqi Gu and Guosheng Yin's contribution to the Discussion of 'Martingale Posterior Distributions' by Fong, Holmes and Walker

Jiaqi Gu[1] and Guosheng Yin[2]

[1]Department of Neurology and Neurological Sciences, Stanford University, Palo Alto, CA, USA
[2]Department of Mathematics, Imperial College London, London, UK

Address for correspondence: Jiaqi Gu, Department of Neurology and Neurological Sciences, Stanford University, 3180 Porter Drive, Palo Alto, CA 94304, USA. Email: jiaqigu@stanford.edu

We congratulate the authors for their thought-provoking paper, especially in the three aspects as follows.

(a) Instead of imposing a prior $\pi(\theta)$ on the parameter θ of a likelihood function $f_\theta(y)$, where both π and f_θ are elicited subjectively, the authors proposed the framework of *martingale posterior* to directly model the predictive $p(y_{n+1:\infty} \mid y_{1:n})$ for which only the prior predictive $p(y)$ needs to be specified. As a result, Bayesian inference can be conducted on any parameter (or statistical functional) of the true sampling distribution.
(b) As the hyperparameter ρ of the copula is chosen in such a data-driven way that $p(y_{1:n})$ fits the observed data $y_{1:n}$ well, the impact of the initial guess for the prior predictive, $p_0(y)$, on $p(y_{n+1:\infty} \mid y_{1:n})$ can be adjusted properly.
(c) Unlike existing MCMC methods which generate autocorrelated posterior samples, the proposed predictive resampling algorithm is GPU-friendly and parallelisable. Thus, independent posterior samples can be obtained for inference, leading to improvement of both computational and statistical efficiency.

The proposed *martingale posterior* framework, we believe, has the further potentials in several directions.

(a) In multivariate density estimation, the authors considered the Dirichlet process mixture model, which can be represented as Chinese restaurant process, for recursive update of predictives. Analogously, Indian buffet process, another commonly used framework in Bayesian nonparametrics, can also be considered to derive predictive update for factor analysis. One example is the latent feature model (Griffths & Ghahramani, 2011),

$$\mathbf{y}_i = \sum_{k=1}^{\infty} z_{ik}\mathbf{f}_{ik} + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n.$$

As traditional MCMC methods involve cumbersome Gibbs sampling of an infinite sparse binary matrix $\mathbf{Z} = (z_{ik})$, it is expected that incorporating predictive resampling would improve the efficiency on inference of latent features, while the main difficulty lies in deriving the update rule of predictive density.
(b) Because variable selection can generally improve statistical efficiency of regression analysis, it is of interest to investigate how Bayesian variable selection can be conducted with control of the Bayesian false discovery rate under the framework of *martingale posterior*.
(c) Although a default choice is suggested for the initial guess $p_0(y)$, it is possible to leverage some information from the observed data $y_{1:n}$ to initialise $p_0(y)$, similar to empirical Bayes methods. This raises a question on how the *martingale posterior* connects to nonparametric empirical Bayes, which also estimates the prior from the data.

Conflict of interests: The authors declare that they have no conflict of interest.

Reference

Griffths T. L., & Ghahramani Z. (2011). The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, *12*(32), 1185–1224. https://doi.org/10.5555/1953048.2021039

https://doi.org/10.1093/jrsssb/qkad092
Advance access publication 28 August 2023

Filippo Ascolani, Antonio Lijoi, and Igor Prünster's contribution to the Discussion of 'Martingale Posterior Distributions' by Fong, Holmes and Walker

Filippo Ascolani, Antonio Lijoi and Igor Prünster

Bocconi Institute for Data Science and Analytics, Bocconi University, Milan, Italy

Address for correspondence: Antonio Lijoi, Bocconi Institute for Data Science and Analytics, Bocconi University, via Roentgen 1, 20136 Milano, Italy. Email: antonio.lijoi@unibocconi.it

We would like to congratulate the authors on their fine and insightful contribution, which provides an original perspective on Bayesian inference and opens up new exciting research directions. The key point is a novel interpretation of the role of prediction: in a Bayesian framework all the uncertainty lies in the unobserved data $Y_{n+1,\infty}$ and, once they are imputed through the predictive distributions, inference is straightforward. Therefore, prediction rules, besides allowing forecasting and extrapolation, are crucial also to infer parameters of interest.

In Bayesian non-parametrics there, is a large stream of works focussing on the m-step ahead prediction for exchangeable species sampling data (see, e.g. Favaro et al., 2009; Lijoi et al., 2007), with recent contributions also in the partially exchangeable set-up (Camerlenghi et al., 2017). However, the predictive distributions are always determined through an indirect procedure that relies on the specification of a non-parametric prior and derives the prediction rule as a posterior expected value. The authors adopt a different, and more direct, approach by considering conditionally identically distributed sequences (Berti et al., 2023) that are only asymptotically exchangeable: in this case, the predictive distributions are available in closed form, but predictions may depend on the order of the observed data $Y_{1:n}$. This seems in contrast with the assumption of independent and identically distributed data that should imply invariance of inferential results with respect to permutation of the observations. Therefore, one may wonder whether the analysis could be extended so to come up with novel exchangeable predictives, without explicit reference to an underlying prior.

To overcome the lack of invariance with respect to the ordering of the data, the authors suggest to average the predictions over different permutations of $Y_{1:n}$. While being computationally unfeasible, averaging over all the permutations induces a symmetry condition which is reminiscent of exchangeability. It would be interesting to check whether the ensuing prediction mechanism actually identifies an exchangeable sequence. This would boil down to showing invariance of the two-step ahead predictives (Fortini et al., 2000). If this were actually the case, the natural goal would be to identify the underlying prior.

In general, the standard prior-likelihood mechanism may still be a plus when it comes to describing the dependence structure among the observations (i.e. the generative model). For instance, hierarchical models, that distinguish global and group-specific parameters, have proven to be useful in multiple fields. We wonder whether it would be possible to encode these structures directly

into the predictive distributions: in particular, it would be interesting to ascertain whether the neat recursive expressions of Section 4, that allow fast computations, can be retained.

Conflict of interest: None declared.

References

Berti P., Dreassi E., Leisen F, Rigo P., & Pratelli L. (2023). Bayesian predictive inference without a prior. *Statistica Sinica*. forthcoming. https://doi.org/10.5705/ss.202021.0238

Camerlenghi F., Lijoi A., & Prünster I. (2017). Bayesian prediction with multiple-samples information. *Journal of Multivariate Analysis*, *156*, 18–28. https://doi.org/10.1016/j.jmva.2017.01.010

Favaro S., Lijoi A., Mena R. H., & Prünster I. (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *Journal of the Royal Statistical Society Series B*, *71*(5), 993–1008. https://doi.org/10.1111/j.1467-9868.2009.00717.x

Fortini S., Ladelli L., & Regazzini E. (2000). Exchangeability, predictive distributions and parametric models. *Sankhya*, *62*(1), 86–109. https://doi.org/10.2307/25051292

Lijoi A., Mena R. H., & Prünster I. (2007). Bayesian nonparametric estimation of the probability of discovering a new species. *Biometrika*, *94*(4), 769–786. https://doi.org/10.1093/biomet/asm061

https://doi.org/10.1093/jrsssb/qkad093
Advance access publication 27 August 2023

David Huk, Lorenzo Pacchiardi, Ritabrata Dutta and Mark Steel's contribution to the Discussion of 'Martingale posterior distributions' by Fong, Holmes and Walker

David Huk[1], Lorenzo Pacchiardi[2], Ritabrata Dutta[1] and Mark Steel[1]

[1]Department of Statistics, University of Warwick, Coventry, UK
[2]Department of Statistics, University of Oxford, Oxford, UK

Address for correspondence: Ritabrata Dutta, Department of Statistics, University of Warwick, Coventry, UK. Email: ritabrata.dutta@warwick.ac.uk

We congratulate the authors for this thought-provoking paper.

The recursive definition of the predictive distributions through a bivariate copula (Section 4) depends on a hyperparameter ρ, which the authors tune by minimising the prequential log-score $-\sum_{i=1}^{n} \log p_{i-1}(\mathbf{y}_i)$ (Section 4.5.2). The prequential framework nicely connects with the predictive resampling approach used later. However, other strictly proper scoring rules (Gneiting & Raftery, 2007) could be used in place of the log-score, leading to a generic prequential score $\sum_{i=1}^{n} S(P_{i-1}, \mathbf{y}_i)$, where $S(P_{i-1}, \mathbf{y}_i)$ is a scoring rule between the distribution P_{i-1} and data $\mathbf{y}_i$. With exchangeable data, the prequential log-score is the only prequential score invariant to permutations of $\mathbf{y}_{1:n}$ (as it corresponds to the log marginal $-\log p_{1:n}(\mathbf{y}_{1:n})$, Fong & Holmes, 2020) and is thus a natural choice over other scoring rules. Nevertheless, in the present set-up exchangeability is forsaken by defining the predictive distributions directly (as the authors remark in Section 3.2 and address in Section 4.5.3); indeed, computing the prequential log-score on multiple permutations of the data leads to different values, as each p_i is defined iteratively from p_{i-1}. Therefore, there seems to be no theoretical reason to prefer the log-score over other strictly proper scoring rules. We also

believe the connection to cross-validation mentioned by the authors relies on exchangeability of the data through the invariance of the marginal likelihood to data permutations (Fong & Holmes, 2020).

Besides, while the form of predictive distribution used by the authors provides access to the density and thus enables convenient computation of the log-score, extensions of this work could rely on predictive distributions whose density can be computed only up to a normalising constant. Furthermore, one could employ predictive distributions for which simulation is possible but density evaluation is not (relying, for instance, on generative neural networks). In both these cases, the log-score would be inaccessible, while other scoring rules [the Hyvärinen score (Hyvärinen, 2005) in the former case and the energy or kernel score (Gneiting & Raftery, 2007) in the latter] would enable hyperparameter tuning. Interestingly, the kernel score enjoys robustness to outliers in the data in different set-ups (Chérief-Abdellatif & Alquier, 2022; Pacchiardi & Dutta, 2021); although we are unsure if this property translates to the considered framework, this is worth investigating.

As a first test, we tuned ρ for the univariate Gaussian mixture model in Section 5.1.1 with the prequential energy score (estimated using an importance sampling strategy) and obtained comparable values of ρ to the ones reported by the authors with the log-score.

Conflict of interest: None declared.

References

Chérief-Abdellatif B.-E., & Alquier P. (2022). Finite sample properties of parametric MMD estimation: Robustness to misspecification and dependence. *Bernoulli*, *28*(1), 181–213. https://doi.org/10.3150/21-BEJ1338

Fong E., & Holmes C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika*, *107*(2), 489–496. https://doi.org/10.1093/biomet/asz077

Gneiting T., & Raftery A. E (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, *6*(24), 695–709.

Pacchiardi L., & Dutta R. (2021). 'Generalized Bayesian likelihood-free inference using scoring rules estimators', arXiv, arXiv:2104.03889, preprint: not peer reviewed.

https://doi.org/10.1093/jrsssb/qkad094
Advance access publication 14 September 2023

Marta Catalano, Augusto Fasano, and Giovanni Rebaudo's contribution to the discussion of 'Martingale posterior distributions' by Fong, Holmes and Walker

Marta Catalano[1], Augusto Fasano[2] and Giovanni Rebaudo[3]

[1]Department of Statistics, University of Warwick, Coventry, UK
[2]Collegio Carlo Alberto, Torino, Italy
[3]Department of Economics, Social Studies, Applied Mathematics and Statistics, University of Turin, Torino, Italy

Address for correspondence: Giovanni Rebaudo, Corso Unione Sovietica, 218/bis, 10134 Turin, Italy. Email: giovanni.rebaudo@unito.it

We congratulate the authors for an insightful foundational contribution to the statistical literature. This work builds on well-established methodologies, such as predictive inference, Doob's theorem, and conditional independent sequences, providing a unifying framework and a novel understanding

of Bayesian uncertainty. A central role is played by the predictive characterisation of the random parameter in terms of observable (yet unobserved) quantities, which are regarded as the root of all statistical uncertainty. This brings to a new approach to inference, termed martingale posterior, which sheds light on interesting connections between Bayesian and frequentist statistics. Indeed, both leverage on an empirical distribution: the former builds it through the predictive distributions, the latter through independent and identically distributed samples. Importantly, martingale posteriors go beyond some common homogeneity assumptions in the data, such as infinite exchangeability. Moreover, this approach may offer practical advantages both in terms of prior elicitation and computations, as illustrated by the authors in some relevant scenarios.

We believe that the breadth of this contribution will inspire several future research questions. Here, we restrict our attention to two aspects that we found particularly interesting. First, the authors underline that inference and prediction under a martingale posterior might depend on the order of the data, even when a natural order does not exist. Such dependence will vanish as the sample size increases, but can still be relevant for finite samples. To overcome such an issue, the authors suggest using the average of the predictive distribution over M random permutations of the sample (e.g. $M = 10$). We believe that, in such a scenario, it could be useful to define and study predictive rules that go beyond infinite exchangeability yet preserve finite exchangeability for any fixed sample size.

Second, from a theoretical and modelling perspective, it is often relevant to establish frameworks that weaken the homogeneity assumption of infinite exchangeability, while still preserving well-defined limits as the sample size increases and tractable learning updates. The authors rely on an additional principle: the martingale predictive coherence. We believe that it could be of interest to relax this principle to study other classes of converging predictive rules, e.g. those that preserve exchangeability but are not Kolmogorov consistent.

To conclude, we believe the work by E. Fong, C. Holmes, and S. Walker will spur several new theoretical, modelling, and computational research directions. We commend the authors one more time for an outstanding paper.

Conflict of interest: All authors declare that they have no conflicts of interest.

https://doi.org/10.1093/jrsssb/qkad095
Advance access publication 27 August 2023

Pietro Rigo's contribution to the Discussion of 'Martingale Posterior Distributions' by Fong, Holmes and Walker

Pietro Rigo

Dipartimento di Scienze Statistiche 'P. Fortunati', University of Bologna, Bologna, Italy

Address for correspondence: Pietro Rigo, Dipartimento di Scienze Statistiche 'P. Fortunati', University of Bologna, via delle Belle Arti 41, 40126 Bologna, Italy. Email: pietro.rigo@unibo.it

This paper is really interesting and offers a number of intriguing hints. Here, I just make a few isolated remarks without any claim of being exhaustive.

1. I would give more emphasis to the Ionescu-Tulcea theorem (ITT). This theorem is quoted only in passing, but it is the cornerstone of this paper. By ITT, the distribution of $Y_{1:\infty} = (Y_1, Y_2, \dots)$ is completely determined by the assignment of $\{P_n : n \geq 0\}$, where

$P_0(\cdot) = P(Y_1 \in \cdot)$ and $P_n(\cdot) = P(Y_{n+1} \in \cdot \mid Y_{1:n})$. Exploiting ITT has at least two advantages: (i) The first part of the paper can be made shorter and clearer; (ii) Relying on ITT makes transparent that, in general, to introduce the problem investigated in this paper, there is no need of any distributional assumption on $Y_{1:\infty}$. In particular, $Y_{1:\infty}$ needs not be exchangeable. Exchangeability should be assumed if (and only if) the inferrer feels that it is reasonable for the specific data at hand, which is true in some problems but false in others. As regards this paper, the only advantage of exchangeability is that the distribution of $Y_{1:\infty}$ can be assigned in two ways: by the usual likelihood/prior scheme (thanks to de Finetti's theorem) or via ITT selecting $\{P_n : n \geq 0\}$. These two routes are equivalent and both determine the distribution of $Y_{1:\infty}$. Hence, it is obvious that predictive resampling is identical to posterior sampling for exchangeable data. I realise that the existence of these two routes is expository useful. But, I do not see any other general reason for assuming exchangeability from the outset. See e.g. Berti et al. (2021, 2023).

2. Suppose the distribution of $Y_{1:\infty}$ is assigned via ITT, but, for some reason, $Y_{1:\infty}$ is requested to satisfy some distributional assumption. For instance, $Y_{1:\infty}$ is asked to be exchangeable, or c.i.d., or stationary, and so on. This puts some constraints on the predictive distributions P_n. So, the problem arises: Is it possible to characterise a distributional assumption on $Y_{1:\infty}$ in terms of the P_n ? This issue has been addressed in some cases (exchangeability and c.i.d.) but not in others (stationarity, partial exchangeability). See Berti et al. (2021, 2023) and references therein.
3. The information at time n is usually larger than the observed values $y_{1:n}$. This could be modelled by introducing a filtration $\mathcal{G}_n$ such that $\sigma(Y_{1:n}) \subset \mathcal{G}_n$ and defining P_n as $P_n(\cdot) = P(Y_{n+1} \in \cdot \mid \mathcal{G}_n)$. Such a generalisation should have a little cost, as most results on c.i.d. sequences work for an arbitrary filtration $\mathcal{G}_n$.
4. Most probably I miss something, but I have some doubts on Section 2.4.1. It is obviously tempting to assign P_n as the empirical measure. But, it does not work. In fact, if P_n is the empirical measure for every $n \geq 1$, one obtains the trivial sequence $Y_n = Y_1$ a.s. for each n.

Conflict of interest: None declared.

References

Berti P., Dreassi E., Leisen F., Pratelli L., & Rigo P. (2023). Bayesian predictive inference without a prior. *Statistica Sinica*, *33*, 1–25. https://doi.org/10.5705/ss.202021.0238

Berti P., Dreassi E., Pratelli L., & Rigo P. (2021). A class of models for Bayesian predictive inference. *Bernoulli*, *27*(1), 702–726. https://doi.org/10.3150/20-BEJ1255

https://doi.org/10.1093/jrsssb/qkad101
Advance access publication 27 August 2023

David Rossell's contribution to the Discussion of 'Martingale Posterior Distributions' by Fong, Holmes and Walker

David Rossell

Department of Economics & Business, Pompeu Fabra University, Barcelona, Spain

Address for correspondence: David Rossell, Department of Economics & Business, Pompeu Fabra University, Ramon Trias Fargas 25, 08005 Barcelona, Spain. Email: rosselldavid@gmail.com

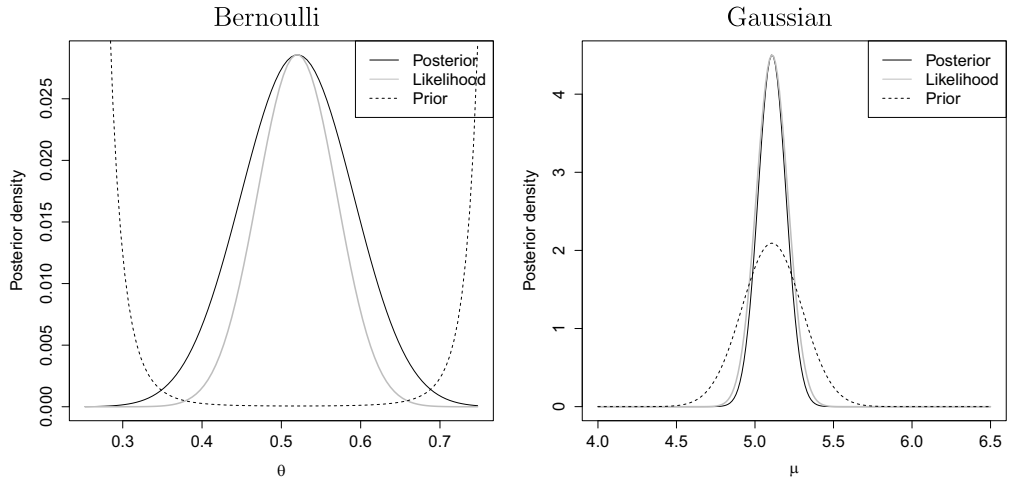

Figure 1. Likelihood, posterior, and prior densities for Bernoulli(0.5) and Normal(5,1) examples with $n = 100$.

Congratulations on a thought-provoking piece. Building Bayesian inference from a (likelihood, predictive) pair, rather than a (likelihood, prior), enriches the paradigm and provides new ways to think about, formulate, and solve problems. A few respectful remarks.

First, although the authors never claim this, it is worth emphasising that the framework is not prior-free. There is a posterior and a likelihood, hence the prior is proportional to their ratio. The key is that said prior is data-dependent, providing an interesting avenue to develop objective Bayes methods, at the cost of losing the coherence property in belief updating. Inspecting the prior can be informative. Figure 1 shows a Bernoulli example where truly $\theta = 0.5$ but the implied prior places little mass around that value, and a Gaussian example where the prior is centred around the sample mean.[1] This apparently erratic prior behaviour might be problematic for model choice via Bayes factors, e.g. returning a very small integrated likelihood in the Bernoulli example.

Second, while sometimes it is easier to elicit a predictive than a prior, in my experience the reverse is often true. For example, in regression, a prior on parameters defines a prior on the R^2 coefficient, an easy-to-interpret quantity, whereas eliciting predictives may be less intuitive for nonstatisticians. Further, note that computational considerations elegantly discussed by the authors severely restrict the range of predictives one may consider in practice, limiting the flexibility of the framework.

Third, I am afraid I disagree on the frameworks' computational convenience. Doing a single optimisation may be faster than sampling, but the framework requires solving many optimisations. This is not cheaper than posterior sampling in a standard (likelihood, prior) construction, also the latter offers fast nonsampling based tools, e.g. Laplace approximations and extensions. It would be interesting to consider analogues for the predictive framework.

Finally, a remark on assuming that at $n = \infty$ there is no uncertainty left. In some settings this is not true, e.g. in high-dimensional regression with $p \gg n$ (one adds higher-order polynomial terms as n grows, say) and a normal prior on the parameters there remains posterior uncertainty even as $n \to \infty$. The proposed framework does not account for such uncertainty, unless suitable adjustments are made.

Conflict of interest: None declared.

https://doi.org/10.1093/jrsssb/qkad096
Advance access publication 1 September 2023

[1] Code at https://github.com/davidrusi/paper_examples/tree/main/2022_Rossell_martingale_posteriors)

Ben Swallow's contribution to the Discussion of 'Martingale Posterior Distributions' by Fong, Holmes and Walker

Ben Swallow

School of Mathematics and Statistics, University of St Andrews, St Andrews, UK

Address for correspondence: Ben Swallow, School of Mathematics and Statistics and Centre for Research into Environmental and Ecological Modelling, University of St Andrews, St Andrews KY16 9LZ, UK.
Email: bts3@st-andrews.ac.uk

Fong et al. (2023) present an interesting and novel approach to conducting Bayesian posterior estimation utilising a joint predictive distribution over missing potential observations and justify how this approach of generating one-step-ahead predictive distributions naturally aligns with the Bayesian likelihood-prior paradigm.

The authors mention time series and hierarchical data as areas of potential future development. However, one of the major topics in Bayesian computational statistics is the problem of selecting a target model from a subset of candidate models or accounting for uncertainty across such models. In their paper, the authors present an approach based on a fixed model structures and conduct posterior inference within those model structures. Given the generality of the approach, I would be interested in hearing the authors' comments on how this approach could be extended to the model uncertainty or model misspecification domain and to what extent the missing data dimension would need to scale with model space size in order to enable feasibility of any such approach.

Conflict of interest: None declared.

Reference

Fong E., Holmes C., & Walker S. G. (2023). Martingale posterior distributions. *The Royal Statistical Society: Series B (Statistical Methodology)*, *85*(5), 1413–1416. https://doi.org/10.1093/jrsssb/qkad135.

https://doi.org/10.1093/jrsssb/qkad097
Advance access publication 27 August 2023

Kolyan Ray and Botond Szabo's contribution to the Discussion of 'Martingale Posterior Distributions' by Fong, Holmes and Walker

Kolyan Ray[1] and Botond Szabó[2]

[1]Imperial College London, South Kensington Campus, London SW7 2AZ, UK
[2]Bocconi Institute for Data Science and Analytics, Bocconi University, Via Roentgen 1, 20136 Milano, Italy

Address for correspondence: Botond Szabó, Bocconi Institute for Data Science and Analytics, Bocconi University, Via Roentgen 1, 20136 Milano, Italy. Email: botond.szabo@unibocconi.it

We congratulate the authors for their thought-provoking article, which includes proposing a copula-based update for the predictive density and establishing its frequentist consistency under relatively mild assumptions. In this contribution, we further explore the *frequentist* properties

of (Bayesian) predictive densities and illustrate through a simple example that, similarly to the posterior distribution, the predictive distribution can also be inconsistent. One therefore requires caution when using martingale posteriors, at least for the frequentist.

Consider a modified version of Example 1 in the present paper, taken from Christensen (2009). Let $Y_1, \ldots, Y_n \sim^{iid} f_\theta$, where

$$f_\theta(y) = \begin{cases} \text{Cauchy}(y \mid \theta) & \theta \in \mathbb{Q}, \\ \mathcal{N}(y \mid \theta, 1) & \theta \in \mathbb{R} \backslash \mathbb{Q}, \end{cases}$$

i.e. for rational parameter values θ, we replace the $\mathcal{N}(\theta, 1)$ Gaussian density by a Cauchy with location parameter θ. As in Example 1, we endow θ with a standard Gaussian prior, i.e. $\pi(\theta) = \mathcal{N}(\theta \mid 0, 1)$. Since the likelihoods in our example and Example 1 are equal almost everywhere under the prior, one can show the corresponding posteriors are also identical, i.e. the posterior is $\mathcal{N}(\theta \mid \bar{\theta}_n, \bar{\sigma}_n^2)$ with $\bar{\theta} = \sum_{i=1}^n y_i/(n+1)$ and $\bar{\sigma}_n^2 = 1/(n+1)$, see Christensen (2009). Similarly, the posterior predictive is $p(y \mid y_{1:n}) = \mathcal{N}(y \mid \bar{\theta}_n, \bar{\sigma}_n^2 + 1)$. By Example 2 in Hahn et al. (2018), the predictive updates can thus be characterised via a Gaussian copula with correlation parameter $\rho_n = (1+n)^{-1}$. However, for any fixed $\theta \in \mathbb{Q}$, which forms a dense set in the parameter space $\mathbb{R}$, the data is Cauchy. The posterior is thus inconsistent for any rational 'true' parameter $\theta \in \mathbb{Q}$, and the predictive density differs substantially from the true Cauchy density, even as $n \to \infty$. Following the notation of the paper, this procedure cannot consistently recover $\theta_\infty = \theta(Y_{1:\infty}) = f_\theta$, even though this parameter is fully defined by the infinite observations $Y_{1:\infty}$. Of course, this example does not contradict Theorem 7, as the assumption $\|f_0/p_0\|_\infty \leq B$ does not hold when f_0 is Cauchy and p_0 is Gaussian.

In this simple example, the discontinuous likelihood function causes the posterior inconsistency, which in turn implies inconsistency of the predictive distribution. The issue is that the above approach only works for parameter values in a set of prior probability one, namely $\mathbb{R} \backslash \mathbb{Q}$. However, prior null sets can be very large if not judged from the prior perspective. This problem becomes more pronounced in nonparametric models, where the set of parameters over which the posterior is consistent can be topologically negligible compared to those where it is inconsistent, see the classical results (Diaconis & Freedman, 1986; Freedman & Diaconis, 1983). This in turn results in predictive distributions not resembling the true data generating distribution.

One must therefore be careful with the choice of predictive distribution, justifying the approach as is done in Theorem 7. We second the authors' view that it is of interest to derive new conditions, analogous to the Kullback–Leibler property in the classical Bayesian setting, which yield consistency and ideally minimax concentration rates.

Acknowledgments

Cofounded by the European Union (ERC, BigBayesUQ, project number: 101041064). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

Conflict of interest: None declared.

References

Christensen R. (2009). Inconsistent Bayesian estimation. *Bayesian Analysis*, *4*(4), 759–762. https://doi.org/10.1214/09-BA428

Diaconis P., & Freedman D. (1986). On inconsistent Bayes estimates of location. *The Annals of Statistics*, *14*(1), 68–87. https://dx.doi.org/10.1214/aos/1176349843

Freedman D., & Diaconis P. (1983). On inconsistent Bayes estimates in the discrete case. *The Annals of Statistics*, *11*(4), 1109–1118. http://dx.doi.org/10.1214/aos/1176346325

Hahn P. R., Martin R., & Walker S. G. (2018). On recursive Bayesian predictive distributions. *Journal of the American Statistical Association*, *113*(523), 1085–1093. https://doi.org/10.1080/01621459.2017.1304219

https://doi.org/10.1093/jrsssb/qkad098
Advance access publication 31 August 2023

Priyantha Wijayatunga's contribution to the Discussion of 'Martingale Posterior Distributions' by Fong, Holmes, and Walker

Priyantha Wijayatunga

Department of Statistics, Umeå University, Umeå, Sweden

Address for correspondence: Priyantha Wijayatunga, Department of Statistics, Umeå University, Umeå, Sweden. Email: priyantha.wijayatunga@umu.se

Statistical uncertainty in (an estimate of) a parameter of a probability distribution is due to missing (unseen) observations (when it is estimated), as authors have noted. We can think that an estimate of the parameter has the maximal uncertainty when no observation is used for it, and no uncertainty when all possible observations are used for it. For example, in Bayesian sense, for a Bernoulli parameter, Beta distribution with parameter values $\alpha = 1$ and $\beta = 1$ represents the full uncertainty. If observation counts are infinite, i.e. α and β are infinite, then there is no uncertainty. The uncertainty of the parameter estimate may be vanished when imputed or really observed data counts used for it are infinite, but the two estimates may converge to different values where the latter is the true value. But unfortunately we often do not have the chance to get it. So, it is not possible to eliminate the uncertainty correctly, i.e. while obtaining the true limiting value for the estimate, by imputing some observations given that some other observations are unknown. This is because

$$p(y'_{n+1:\infty}|y_{1:n}) = \int p(y'_{n+1:\infty}, \theta_{1:n}|y_{1:n})\,\mathrm{d}\theta_{1:n} = \int p(y'_{n+1:\infty}|y_{1:n}, \theta_{1:n})p(\theta_{1:n}|y_{1:n})\,\mathrm{d}\theta_{1:n}$$
$$= \int p(y'_{n+1:\infty}|\theta_{1:n})p(\theta_{1:n}|y_{1:n})\,\mathrm{d}\theta_{1:n}$$

if and only if $Y'_{n+1:\infty}$ is conditionally independent of $Y'_{1:n}$ given the value $\theta_{1:n}$, which is the authors' equation (1.1). That is, imputed observations $y'_{n+1:\infty}$ show only the variation that is dictated by the observations $y_{1:n}$ through $\theta_{1:n}$. However, real observations $y_{n+1:\infty}$ from the random variable Y may show somewhat different variation unless the observations $y_{1:n}$ (through $\theta_{1:n}$) determine the true probability distribution of Y. Generally we may get two different values for $\theta(Y_{1:\infty})$ and $\theta(Y_{1:N} \cup Y'_{N+1:\infty})$ for any finite N, even though both of them are without any uncertainty. The problem is that if there are any mathematical operations or formulae for making these two values the same. Eliminating the uncertainty is not sufficient when we do not obtain the true value of the parameter. Note that, $p(\theta_{1:n}|y_{1:n})$ is true posterior distribution of the parameter in the event of observed data $y_{1:n}$. One should use $p(\theta_0)$ as the prior distribution in the Bayesian updating $p(\theta_{1:n}|y_{1:n}) \propto p(y_{1:n}|\theta_0)p(\theta_0)$ to make the operation clear.

Conflict of interest: None declared.

The authors replied later, in writing, as follows:

https://doi.org/10.1093/jrsssb/qkad099
Advance access publication 29 August 2023

Authors' reply to the Discussion of 'Martingale Posterior Distributions'

Edwin Fong[1], Chris Holmes[2] and Stephen G. Walker[3]

[1]Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong
[2]Department of Statistics, University of Oxford, Oxford, UK
[3]Department of Statistics and Data Sciences, University of Texas, Austin, USA

Address for correspondence: Chris Holmes, Department of Statistics, University of Oxford, 24-29 St Giles', Oxford, OX2 3LB, UK. Email: chris.holmes@stats.ox.ac.uk

We thank all the discussants for their contributions which highlight important aspects of our work and additionally provide many hints for future directions of research. We are pleased that the discussions were generally receptive to expanding beyond the traditional likelihood-prior framework of Bayes, and to shift the focus to the predictive distribution.

We begin our rejoinder by summarizing some core themes that have been raised by the discussants, followed by more detailed discussions. These are as follows:

- Some discussants highlighted the role of the already observed data, $y_{1:n}$. In our view, given the observed information, the predictive $p(y_n+1|y_{1:n})$ is the analysts' best density estimator for the next observation, and this will reflect the underlying structure of the data. There is no explicit need to model the observed $y_{1:n}$. From an intuitive perspective, there is no uncertainty in what has already been seen, i.e. the data $y_{1:n}$ are fully known and we treat it as such. Hence, assigning a probability model to what has been seen could be argued as unnecessary.
- The martingale posterior is not to be regarded as a replacement of traditional Bayesian analysis. If it is deemed that the formal Bayesian model, derived from a prior to posterior, leads to the best predictive, then this remains a part of the framework espoused in the article. We argue however that there are settings where this rigid framework is questionable, such as when tasked with eliciting objective priors in the absence of substantive prior information.
- There is much discussion on the formal structure of the data to be studied. For assumed i.i.d. data, the Bayesian adopts an exchangeable structure from their symmetry of beliefs, not because they actually believe the measured states are physically dependent, but because the predictives, i.e. density estimators for the next observation, evolve as more is seen. It is the predictive which depends on the past, rather than the variable coming from it. For example, when considering a sequence of tosses of a coin, the outcome of one toss does not influence the outcome of any other toss, but an observed toss is certainly informative for estimating the probability of getting heads in the next toss. Exchangeability can be viewed as providing a learning model, as can a c.i.d. structure for $Y_n+1:\infty$, which we adopt as the basic framework for the missing data.

1 The role of observations

The role of the observations $y_{1:n}$ in determining the initial predictive P_n was a common theme in multiple comments. As highlighted by Dawid, the statistical model is the sequence of predictives. While we agree that it is important to link the observed $y_{1:n}$ to the unobserved $Y_{n+1:\infty}$, we argue that the traditional starting point of an assumed structure for $y_{1:n}$ can now play a secondary role to our primary goal, which is to construct the best predictive P_n for Y_{n+1} given $y_{1:n}$. In our view, $y_{1:n}$ has already been observed and there is no longer any uncertainty in their values—only uncertainty in $Y_{n+1:\infty}$. Draper and Guo highlight the asymptotic equivalence between frequentist and Bayesian bootstraps, which could be a useful approximation in some instances. However, we must reiterate that the source of uncertainty is entirely different: the frequentist considers $y_{1:n}$ as random, whereas the Bayesian considers $Y_{n+1:\infty}$ as random given $y_{1:n}$.

The elicitation of the distribution of $Y_{n+1:\infty}$ given $y_{1:n}$ is quite a general problem, of which there are many approaches. A sequential approach would entail a sequence of one-step ahead predictive distributions. Indeed, this becomes the statistical problem. For example, Dawid's prequential approach,

where one specifies P_0, P_1, $\ldots$, P_{n-1} and assesses its fit to the observed data $y_{1:n}$ is a natural approach, and we indeed carry this out to select hyperparameters and learn P_n through our copula updates.

The secondary focus on the data structure $y_{1:n}$ does not preclude us from using knowledge of the data generating mechanism. In the presence of a statistical model, e.g. in the scenario proposed by Clarke, we are free to use the plug-in predictive $p(\cdot|y_{1:n}) = f(\cdot|\widehat{\theta}_n)$, with $(\widehat{\theta}_m)_{m>n}$ as the possible martingale, which was recently proposed by Holmes and Walker (2023). This predictive can be based on a more conventional statistical model and estimator, such as the maximum likelihood estimator. One can then view this as setting P_n to the best density estimate of the data generating distribution given current knowledge. Of course, if we know something about the structure of $Y_{n+1:\infty}$, this may help us elicit the best predictive.

One is also free to incorporate prior knowledge in addition to the observations. For example, Rigo and Antoniano-Villalobos highlight that given a single observation y_1, the empirical measure as the predictive would return us no uncertainty in the martingale posterior. However, we argue that in this setting, it would be natural to incorporate prior information (such as historical data) or smoothing as suggested by Antoniano-Villalobos instead of relying on the empirical distribution for a single datum.

Note that all bootstraps fail with a single observation. When n is large, and in the absence of substantive background knowledge, having to specify a prior will often be a distraction. The martingale posterior frees the Bayesian from this restriction.

When faced with the issue of selecting a predictive, we agree with the discussants that the standard battery of model selection tools can be utilized, such as prequential tests and scores as suggested by Dawid and Huk et al. This can also be applied to select hyperparameters such as the bandwidth ρ or initial guess $p_0(y)$, which has close connections to empirical Bayes as highlighted by Gu and Yin. We agree with Swallow and Gu and Yin that uncertainty in model selection or variable selection is one of the strengths of the Bayesian framework. Future directions of research would be to investigate what the missing future data and the decision problem is under the martingale posterior framework. To date, given a model, we are able to assess the uncertainty via a martingale posterior distribution for the object of interest. We are currently working on decision problems where the idea is to make a choice with associated uncertainty quantification. The idea here is to generate $Y_{n+1:\infty}$ which makes the decision known, whereas each different sequence of future data could present a different decision.

We agree with Ray and Szabo and Wijayatunga on the importance of frequentist consistency of the predictive, which is also connected to the issue of model selection. Ray and Szabo provide an interesting example of posterior inconsistency, which we believe can be alleviated through the above discussion on model selection. If it were known the model was either normal or Cauchy, there are a number of ways this could be determined from $y_{1:n}$ and the appropriate predictive selected. An interesting future direction would be to investigate how good frequentist properties for the predictive impact the posterior.

2 Comparison to traditional Bayes

Lauritzen, Ascolani et al., Rossell, and Mena highlighted that it may sometimes be easier to elicit a likelihood and prior. We agree with this sentiment, and reiterate that the martingale posterior is not aiming to replace Bayes, as it encompasses it. We envision many scenarios, as those suggested by the discussants, where a prior to posterior calculation leads to the best predictive.

However, there may be situations in which we do not have any prior information, such as those considered by objective Bayesians. One of the goals of the article is to highlight that the prior distribution is not needed for Bayesian inference, where the uncertainty arises from the unseen $Y_{n+1:\infty}$. Aside from the Bayesian bootstrap, another illuminating example is the plug-in predictive as discussed in the previous section and Holmes and Walker (2023). Here, our predictive $f(\cdot|\widehat{\theta}_n)$ is based on a statistical model, so we have discarded the prior distribution component of the traditional Bayesian setup. There is now no acknowledged uncertainty in the value of $\widehat{\theta}_n$ from a plug-in density estimator, but statistical uncertainty in θ_∞ is constructed via the sampling from the sequence $Y_{n+1:\infty}$. We thus do not entirely agree with Rossell's comment that we are not prior-free, as there is no explicit prior distribution on θ.

However, if you have prior information, e.g. in the form an initial predictive density p_0, then this can also be evaluated in a data-driven way. Another setting where eliciting a prior could be deemed inconvenient is in Bayesian non-parametrics, where priors can be technical and not necessarily intuitive. As highlighted by Ray and Szabo, predictive distributions elicited through the likelihood-prior construction may be misleading due to large prior null sets, especially in the non-parametric setting. Eliciting and evaluating the predictive directly could be a way to bypass these issues, though other new challenges arise, and we look forward to exploring this direction further.

Another reason to bypass the likelihood-prior construction is due to computational reasons. Depending on the setting, computation with predictives can be noticeably faster than traditional posterior sampling, e.g. the Bayesian bootstrap or predictive resampling as demonstrated by Moya, but we acknowledge that this is not always universal. Rossell highlights that optimization can be costly, but we agree with Gu and Yin and Moya that parallelization is easily carried out and is less straightforward with MCMC methods. Furthermore, predictive resampling as a computational algorithm is still in its infancy. We hope that additional computational methods or approximations can be developed under the martingale posterior framework.

3 Structure in the imputed population

Another common theme among discussants is the structure of the imputed data $Y_{n+1:\infty}$, and going beyond exchangeability. Ascolani et al. and Catalano et al. discuss whether one can identify novel direct predictive updates which are exchangeable (perhaps only finitely), and Mena and Antoniano-Villalobos questions the importance of exchangeability. While we do indeed average over permutations in our construction of P_n to ensure some permutation invariance of the predictive, this is mostly motivated by practical concerns. It may be undesirable for the analysis to be highly sensitive to the order in which we process the data, but we do not view exchangeability as crucial for the martingale posterior. Rigo highlights the importance of the Ionescu-Tulcea theorem in our setting, with which we agree. Indeed, when we are only concerned with the structure of $Y_{n+1:\infty}$, our choice of c.i.d. sequences gives us sufficient asymptotic structure for the i.i.d. setting, but as seen in the plug-in examples above, this is not the only way to guarantee convergence or coherence under the i.i.d. setting.

Closely connected to the above is the question on the extension of our framework beyond the i.i.d. setting. Lauritzen questions if the martingale posterior framework is sufficiently general, and provides a pure birth process example. Similarly, Dawid asks what is necessary under a Markov setting. AntonianoVillalobos also highlights the importance of assumptions such as smoothness and sparsity when our object of interest requires more than a countable number of observations. While the generality of the martingale posterior still requires investigation, we provide some preliminary hints below.

For example, if the data are Markov AR(1) as in Dawid and Mena's example, the datum y_n would play a significant role in predicting y_{n+1}, and the predictive would be of the form $p(y_{n+1}|\widehat{\theta}_n, y_n)$. The corresponding structure for $Y_{n+1:\infty}$ may not be exactly Markov but the outcome would be that it is asymptotically Markov as $\theta_N \to \theta_\infty$ eventually. This line of thinking may also allow us to generalize beyond the martingale condition as asked by Mena. Hence, the structure is to produce the best form of predictive which would arise from the structure of the likelihood function, and a formal characterization of this would be of great interest.

The model proposed by Lauritzen can also be handled using the martingale posterior framework. Suppose we have observed $X(t)$ up to time T from which we have λ_b. We now take the predictive to be, for any $h > 0$, and $t \geq T$,

$$P(Y(t+h) = j+1 | Y(t) = j, \lambda_b(t)) = \lambda_b(t) jh$$

and

$$P(Y(t+h) = j | Y(t) = j, \lambda_b(t)) = 1 - \lambda_b(t) jh.$$

We can discretize this in an obvious way: we use intervals of time $h/Y[N-1]$, for $N = 1, 2, \ldots$ with $Y[0] = X[T]$, for some small h. Hence, jumps can occur at times $t(N) = t(N-1) + h/Y[N-1]$, with $t(0) = T$, and

$$S(N) = \int_0^T X(u)du + \int_T^{t(N)} Y(u)du,$$

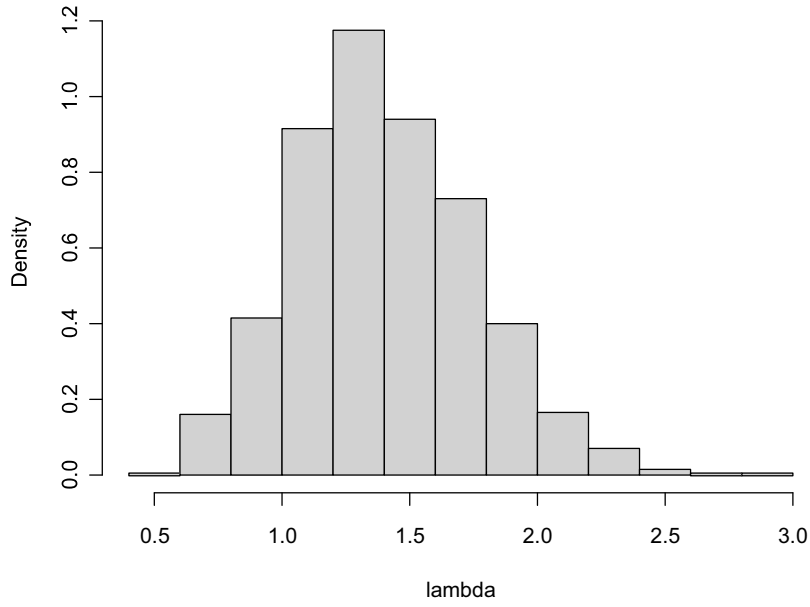

Figure 1. Histogram representation of samples from the martingale posterior for λ.

where the first term on the right is $S(T)$ and the final term on the right is Nh. As we predictively resample $Y(N)$, we update λ_b as we go. Suppose at iteration N we have $Y(N)$ and $\widehat{\lambda}_N$, then

$$Y(N+1) = Y(N) + \text{Bernoulli}\,(\widehat{\lambda}_N h),\ \widehat{\lambda}_{N+1} = \frac{Y(N+1)-1}{S(T)+(N+1)h}.$$

It is easy to see that $(\widehat{\lambda}_N)N_{\geq 0}$ is a martingale.

We generated a process using $\lambda = 1.5$ and observe $X(T) = 15$ with $T = 1.633$ and $S(T) = 10$, with $\lambda_b = 1.4$. We then sample the martingale posterior for $N = 1, \ldots, 10000$, which is far more than is required for the martingale to converge. Repeating this gives us 1,000 samples from the martingale posterior, as shown in Figure 1, where the mean of the samples is 1.40. Lauritzen's second example with both $\{\lambda, w\}$ seems more challenging, and we defer this to future work.

In the above two examples, we have relied on the parametric statistical model to ensure structure in $Y_{n+1:\infty}$, but without the reliance on the full Bayesian machinery (i.e. no prior distribution). An extension to hierarchical models would be interesting, as highlighted by Ascolani et al., and it would be of interest if the structures considered above can be incorporated directly into the predictive. We believe this to be a fruitful direction of future research.

Conflicts of interest: none declared.

References

Holmes, C. C., & Walker, S. G. (2023). Statistical inference with exchangeability and martingales. *Philosophical Transactions of the Royal Society A*, *381*(2247), 20220143. https://doi.org/10.1098/rsta.2022.0143