Biometrika (2026), **113**, 2, asag007 https://doi.org/10.1093/biomet/asag007
Advance Access publication 19 February 2026

Asymptotics for a class of parametric martingale posteriors

BY E. FONG

Department of Statistics and Actuarial Science, University of Hong Kong, Pokfulam, Hong Kong
chefong@hku.hk

AND A. YIU

School of Mathematical Sciences, University of Southampton, University Road, Southampton SO17 1BJ, U.K.
a.h.m.yiu@soton.ac.uk

SUMMARY

The martingale posterior framework replaces the elicitation of the likelihood and prior with that of a sequence of one-step-ahead predictive densities for Bayesian inference. Posterior sampling then involves the imputation of unobserved quantities and can then be carried out in an expedient and parallelizable manner using predictive resampling, without requiring Markov chain Monte Carlo. Recent work has investigated the use of plug-in parametric predictive densities, combined with stochastic gradient descent, to specify a parametric martingale posterior. This paper investigates the asymptotic properties of this class of parametric martingale posteriors. In particular, two central limit theorems based on martingale limit theory are introduced and applied. The first is a predictive central limit theorem, which enables a significant acceleration of the predictive resampling scheme through a hybrid sampling algorithm based on a normal approximation. The second is a Bernstein–von Mises result, which is novel for martingale posteriors, and provides methodological guidance on attaining desirable frequentist properties. We demonstrate the utility of the theoretical results through simulations and a real data example.

Some key words: Bayesian inference; Central limit theorem; Martingale posterior.

1. INTRODUCTION

1.1. *Martingale posteriors*

The predictive view of Bayesian inference highlights the equivalence between posterior uncertainty about the parameter of interest and the predictive uncertainty of yet-to-be seen observables. The predictive Bayesian then leverages this key connection to replace the usual likelihood-and-prior construction of a Bayesian model with the direct elicitation of a sequence of one-step-ahead predictive densities as the statistical model. Given observations $y_{1:n}$, posterior computation now involves the sequential imputation

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

taking $N \to \infty$, where $p(\cdot \mid y_{1:i})$ is the one-step-ahead predictive density conditional on the first i observations. Convergence of the above imputation scheme is then usually guaranteed through the construction of a martingale, which weakens the usual requirement of exchangeability. The parameter of interest is then computed as a function of the limiting imputed population, which we write informally as $\theta_\infty = \theta(y_{1:n}, Y_{n+1:\infty})$, and its distribution is termed the martingale posterior. This distribution of θ_∞ can be interpreted as a posterior, as it is conditioned on $y_{1:n}$, with uncertainty arising from the missing population $Y_{n+1:\infty}$.

One key benefit of the martingale posterior is the ability to work with a much larger class of models for Bayesian inference. These include conditionally identically distributed sequences (Berti et al., 2004; Fong et al., 2023), quantile estimates (Fong & Yiu, 2025) and plug-in parametric predictive densities (Walker, 2022; Holmes & Walker, 2023), the last of which forms the focus of this paper. The sequence of predictive densities may often be elicited without the specification of an explicit prior, although there may be an implicit prior. In addition to gains in modelling flexibility, posterior computation is also drastically different and allows for significant gains in computational speed. The sequential imputation scheme, also known as predictive resampling, relies on recursive simulation and updating of the predictive density with a suitable truncation point N, removing the need for Markov chain Monte Carlo. This sampling scheme can also be entirely parallelized and can utilize modern graphics processing units, and is free of the convergence challenges faced by Markov chain Monte Carlo.

This paper focuses on developing asymptotic theory for the parametric martingale posterior. In particular, we establish two central limit theorems under different asymptotic regimes. The first concerns predictive asymptotics, in which the imputed population size tends to infinity. The second concerns frequentist asymptotics, in which the number of independent and identically distributed observations tends to infinity, as in the Bernstein–von Mises theorem. We show that these theoretical results have practical implications for accelerating posterior computation and for attaining frequentist coverage.

1.2. *Related work*

The predictive view of Bayesian inference has a long tradition, dating back to de Finetti (1937), who emphasized the role of probability statements about observables. The operationalization of this connection into a predictive Bayesian methodology was first outlined by Fortini & Petrone (2020), using the predictive recursion algorithm of Newton & Zhang (1999). The predictive framework was further developed to accommodate a wider class of nonparametric predictive models by later work, including Berti et al. (2021b) and Fong et al. (2023), the latter of which coined the term ‘martingale posterior’. The parametric martingale posterior, which utilizes a parametric predictive for sequential imputation, was then introduced and explored by Walker (2022), Holmes & Walker (2023) and Wang & Holmes (2024), with an extension to mixture models by Cui & Walker (2024). Recently, Fortini & Petrone (2025) also considered the parametric martingale posterior for logistic regression, and Garelli et al. (2024) considered the predictive asymptotics of parametric predictives based on a mean and variance estimate. Our work builds on these proposals in the existing literature, and the theory we develop encompasses a wide class of parametric martingale posteriors.

From a theoretical perspective, leveraging martingale central limit theorems for predictive Bayesian inference was first proposed by Fortini & Petrone (2020), with further developments by Berti et al. (2021a) and Fortini & Petrone (2023). In particular, Fortini & Petrone

(2020) and Favaro & Fortini (2024) utilized a result of Crimaldi (2009) to construct asymptotic credible intervals for the algorithm of Newton & Zhang (1999). Fong & Yiu (2025) also utilized a central limit theorem to carry out approximate posterior sampling from a nonparametric quantile martingale posterior. Closest to our first forthcoming result is the predictive central limit theorem for the logistic regression case provided in Fortini & Petrone (2025), which we extend to more general settings. More generally, the parametric martingale posterior is closely connected to the parametric bootstrap (Efron, 2012), the Bayesian bootstrap (Rubin, 1981) and generalized Bayesian methods (Bissiri et al., 2016; Knoblauch et al., 2022).

2. Methodology

In this section, we introduce a specific form of the parametric martingale posterior distribution, in which the sequence of predictive densities is parametric. Predictive sequences of this type have been investigated by Walker (2022), Holmes & Walker (2023), Garelli et al. (2024) and Fortini & Petrone (2025), and the predictive sequence defined below encompasses them in a certain sense. Let $\{p_\theta(y) : \theta \in \Theta\}$ denote a family of probability density functions on $y \in \mathcal{Y}$, for example the normal family with fixed variance, where $p_\theta(y) = \mathcal{N}(y; \theta, \sigma^2)$. For ease of exposition, we begin with the univariate case, with $\Theta \subseteq \mathbb{R}$. However, all our results generalize fully to the multivariate case, with further details provided in § 5.1 below and § F of the Supplementary Material. For the remainder of the paper, we make the following assumptions on $p_\theta(y)$.

Assumption 1. For each $\theta \in \Theta \subseteq \mathbb{R}$, we require the following, where expectations are understood to be taken over $Y \sim p_\theta$. The score function $s(\theta, y) = \partial \log p_\theta(y)/\partial\theta$ exists, is finite and has mean 0, that is, $E_\theta\{s(\theta, Y)\} = 0$. Furthermore, the score function is square integrable, that is, $\mathcal{I}(\theta) = E_\theta\{s(\theta, Y)^2\} < \infty$, where $\mathcal{I}(\theta)$ is the Fisher information. Finally, we require that $\mathcal{I}(\cdot)$ is continuous at θ and satisfies $\mathcal{I}(\theta) > 0$.

Suppose that we have observed $y_{1:n}$ drawn from p_{θ^*}, where $\theta^* \in \mathbb{R}$ is some true underlying value. We discuss this assumption of model well specification in a later section. Now consider the plug-in predictive density $p_n(y) = p_{\theta_n}(y)$, which acts as our starting point for the martingale posterior. Here, θ_n is an estimate of θ computed from $y_{1:n}$, for example the maximum likelihood estimate. We now introduce the recursive update that allows us to impute the remainder of the population.

Consider the random sequence of predictive densities, $\{p_{\theta_N}(y)\}_{N=n+1,n+2,\ldots}$, where $Y_{N+1} \sim p_{\theta_N}(\cdot)$ and the parameter estimate satisfies

$$\theta_N = \theta_{N-1} + N^{-1}\mathcal{I}(\theta_{N-1})^{-1}s(\theta_{N-1}, Y_N) \tag{1}$$

for all $N > n$. Crucially, the learning rate of N^{-1} ensures that

$$\sum_{N=n+1}^{\infty} N^{-1} = \infty, \qquad \sum_{N=n+1}^{\infty} N^{-2} < \infty, \tag{2}$$

as is standard in stochastic approximation. All of the results in this paper continue to hold if we replace N^{-1} with a general sequence α_N, provided that the above conditions are satisfied.

The update (1) is considered for logistic regression in Fortini & Petrone (2025) without the Fisher information, and in the general case in Walker (2022), Holmes & Walker (2023) and Wang & Holmes (2024).

The above can be interpreted as the online learning rule for the sequence of predictives and has a clear connection to stochastic approximation (Lai, 2003), also referred to as stochastic gradient descent. We can thus leverage tools from the rich stochastic approximation literature to study the implied martingale posterior. The key difference from the traditional stochastic approximation set-up is that the data $(Y_N)_{N>n}$ are generated from the assigned predictive distributions, rather than being independently and identically distributed from p_{θ^*}. The preconditioning of the gradient by the inverse Fisher information is also akin to the Hessian matrix used in Newton's method for optimization, and is also termed the natural gradient in the machine-learning literature (Martens, 2020). This deliberate choice has ramifications for the asymptotic variance later.

Define the filtration $(\mathcal{F}_N)_{N\geqslant n}$, where $\mathcal{F}_N = \sigma(Y_{n+1}, \ldots, Y_N)$, $\mathcal{F}_n$ is the trivial σ-algebra and $y_{1:n}$ is considered fixed for now. It is clear that $(\theta_N)_{N\geqslant n}$ is a martingale with respect to $(\mathcal{F}_N)_{N\geqslant n}$, as we have $E\{s(\theta_N, Y_{N+1}) \mid \mathcal{F}_N\} = 0$ by assumption. Under appropriate assumptions, which will be investigated in detail in the next section, we further have $\sup_N E(\theta_N^2) < \infty$, so that $(\theta_N)_{N\geqslant n}$ is a martingale bounded in L_2. By Doob's martingale convergence theorem (Doob, 1953), there exists a finite $\theta_\infty \in \mathbb{R}$ such that $\theta_N \to \theta_\infty$ almost surely. The distribution of the random variable θ_∞, conditional on $y_{1:n}$, is then termed the parametric martingale posterior.

Up to this point, we have omitted discussion of the initial estimate θ_n. As this work investigates the frequentist properties of the parametric martingale posterior, the results depend on the chosen estimate θ_n. In particular, we could enforce the same update (1) for $N = 1, \ldots, n$, where $\theta_0 \in \mathbb{R}$ is an initial value. This yields a form of coherency, in that the estimation matches predictive resampling. As discussed by Walker (2022), one can then view Bayesian inference as applying update (1) to independent and identically distributed observations $y_{1:n}$ and then continuing the same learning scheme with imputed data $Y_{n+1:\infty}$ drawn from p_{θ_N} once the real observations are exhausted. However, using (1) as the recursive update for observed samples $y_{1:n}$ is accompanied with some disadvantages. First, the estimate θ_n depends on the ordering of the observations $y_{1:n}$. Second, the convergence rate of θ_n to θ^* often requires delicate tuning of the learning rate, which is a well-known issue in stochastic optimization (Lai, 2003). As an alternative, it may be simpler and acceptable to use a well-understood estimator, such as the maximum likelihood estimate, for θ_n, provided that it is asymptotically equivalent to using (1), which is the approach we implement in practice.

3. Predictive central limit theorem

3.1. *Predictive asymptotics*

We begin our study by considering predictive asymptotics, which we distinguish from frequentist asymptotics. In this context, we treat $y_{1:n}$ as fixed and consider the convergence of θ_N to θ_∞ as $N \to \infty$, as also considered by Doob (1953). In other words, we study predictive resampling as the imputed population $Y_{n+1:N}$ grows with $N \to \infty$. In § 4 below, we then consider the frequentist case, in which $Y_{1:n}$ is independently and identically distributed from some P^*, and we take both $n, N \to \infty$.

In this section, we consider the limiting distribution of $(\theta_\infty - \theta_N)$, appropriately scaled, as $N \to \infty$, using martingale limit theory (Hall & Heyde, 1980). The main goal of this section is computational, as understanding the limiting law of $(\theta_\infty - \theta_N)$ allows us to accelerate predictive resampling by truncating early and adding a random correction factor, which we discuss in detail in § 3.4. However, we emphasize that θ_∞ is random, so we are considering the convergence of a random object θ_N to another random object θ_∞. As a result, the limiting law of θ_∞ has random components; this phenomenon is usually encountered when studying the limits of exchangeable sequences, but is less common in frequentist limit theorems.

Predictive asymptotics have previously been studied by Fortini & Petrone (2020, 2025), who provided insightful interpretations of this type of asymptotics. We view our results as a generalization of theirs, tailored to the parametric case. Interestingly, the assumptions are somewhat more restrictive in the parametric setting, as $\theta \in \Theta \subseteq \mathbb{R}$ need not have compact support. This contrasts with the nonparametric case, in which the martingale constructed is usually the cumulative distribution function, which is bounded by 1. Another novelty we demonstrate is that quantification of the above distribution allows us to justify a computational scheme that drastically accelerates posterior sampling, with strong practical benefits.

3.2. *Asymptotic normality*

To find the limiting distribution under predictive resampling, we now utilize the seminal work in martingale central limit theory (Hall & Heyde, 1980), which draws strong analogies between independent and identically distributed variables and martingales. However, while elegant, the assumptions for martingale central limit theorems can become quite technical. This is especially true in our setting, as the estimates θ_N need not lie in a compact set. A discussion of this is also provided in Walker (2022).

For martingale central limit theorems to hold, we require that our martingale is bounded in L_2, that is, $\sup_N E(\theta_N^2) < \infty$, which is equivalent to the bounded-variance condition in traditional central limit theorems. Furthermore, while only boundedness in L_1 is required for the existence of θ_∞, boundedness in L_2 is more convenient to establish in our setting, and we therefore also rely on this property for martingale convergence. However, as we will see shortly, a stronger Lindeberg condition is in fact needed for a martingale central limit theorem to hold. To that end, we introduce the following assumption, which helps control the martingale under predictive resampling.

Assumption 2. Let Z_N denote the natural gradient, that is,

$$Z_N = \mathcal{I}(\theta_{N-1})^{-1} s(\theta_{N-1}, Y_N). \tag{3}$$

The sequence $(Z_N^2)_{N>n}$ is uniformly integrable under predictive resampling with (1) starting from the initial θ_n.

The quantity Z_N is also the efficient influence function for θ under well specification of the parametric model (Van der Vaart, 2000), which we require to be square-uniformly integrable. The above condition is sufficient for a martingale version of Lindeberg's condition to hold. While the condition is easy to state, it may be challenging to verify, which is typical of assumptions required for martingale central limit theorems (Heyde & Johnstone, 1979). This difficulty arises from the fact that we are converging to a random limit θ_∞, which may lie anywhere in $\mathbb{R}$. As a result, we cannot always bound Z_N, whereas boundedness was previously exploited in the results of Fortini & Petrone (2020, 2025). Fortunately, we introduce below an assumption that is easy to verify and that holds for many choices of $p_\theta(y)$, provided that the learning rate satisfies (2). We postpone verification of Assumption 2 until § 3.3.

We now show that Assumption 2 implies another key condition, namely that the martingale is bounded in L_2, which is required for the central limit theorem. In this section, all almost-sure statements are with respect to the measure P^∞ on $Y_{n+1:\infty}$, which is guaranteed to exist by the Ionescu–Tulcea theorem.

Lemma 1. *Under Assumption 2, the sequence $\{\mathcal{I}(\theta_{N-1})^{-1}\}_{N>n}$ is uniformly integrable. Furthermore, $(\theta_N)_{N\geqslant n}$ is a martingale bounded in L_2, and hence $\theta_N \to \theta_\infty$ almost surely.*

We are now equipped to state the main theorem of this section.

Theorem 1. *For $X_N = \theta_N - \theta_{N-1}$, define*

$$V_N^2 = \sum_{i=N+1}^{\infty} E(X_i^2 \mid \mathcal{F}_{i-1}), \qquad s_N^2 = \sum_{i=N+1}^{\infty} E(X_i^2) = E\{(\theta_\infty - \theta_N)^2\}.$$

Under Assumptions 1 and 2, we have

$$\frac{V_N^2}{s_N^2} \to \frac{\mathcal{I}(\theta_\infty)^{-1}}{E\{\mathcal{I}(\theta_\infty)^{-1}\}}$$

almost surely, and

$$V_N^{-1}(\theta_\infty - \theta_N) \to \mathcal{N}(0, 1) \tag{4}$$

in distribution.

Proof. We outline the proof, with details deferred to § D of the Supplementary Material. The proof is an application of a variant of the tail-sum martingale central limit theorem; see, e.g., Hall & Heyde, (1980, Corollary 3.5) or Häusler & Luschgy, (2015, Exercise 6.7). To summarize, Assumption 2 implies a conditional Lindeberg condition and further implies that the normalized variance V_N^2/s_N^2 converges almost surely. These two conditions are then sufficient for the martingale central limit theorem for tail sums to hold. □

Many variants of the martingale central limit theorem exist, which we specialize to our use cases in the Supplementary Material. A key component of the above result is that the scaled limit of the normalizing variance V_N^2 is random and equal to $\mathcal{I}(\theta_\infty)^{-1}$, so the variance of the limiting distribution of $s_N^{-1}(\theta_\infty - \theta_N)$ is also random. Here, s_N^{-1} is the deterministic rate, whereas the ratio V_N^2/s_N^2 plays the role of the random variance. For large N, the heuristic argument is

$$V_N^2 \approx \mathcal{I}(\theta_\infty)^{-1} \sum_{i=N}^{\infty} i^{-2}, \qquad s_N^2 \approx E\{\mathcal{I}(\theta_\infty)^{-1}\} \sum_{i=N}^{\infty} i^{-2},$$

where $\mathcal{I}(\theta_\infty)^{-1}$ is random, but $E\{\mathcal{I}(\theta_\infty)^{-1}\}$ is constant. We then have $s_N^{-1} \propto N^{1/2}$, which is the usual parametric rate.

The random limit of V_N stands in stark contrast to classical central limit theorems, in which the normalizing variance typically converges to a constant. The ability to normalize

$(\theta_\infty - \theta_N)$ by V_N, and thus to express the limiting distribution in (4) independently of θ_∞, is crucial for the practical use of Theorem 1 for approximate posterior sampling, as we will see shortly. The key to this form of (4) is precisely the stable convergence of martingale central limit theorems (Hall & Heyde, 1980, Ch. 3), in the sense of Rényi (1963), where stable convergence is stronger than convergence in distribution and weaker than convergence in probability. The importance of stable convergence is also emphasized in Fortini & Petrone (2020, 2025). Informally, martingale central limit theorems guarantee that $X_N \to \mathcal{N}(0, \eta^2)$ stably, where the limiting variance η is random. Stable convergence then allows the replacement of this limit with $\eta_N^{-1} X_N \to \mathcal{N}(0, 1)$, where $\eta_N \to \eta$ in probability, as in Theorem 1. Crucially, this extension of Slutsky's theorem to the case in which η is random would not necessarily hold if stable convergence were weakened to convergence in distribution. In the interests of space, we provide an extended discussion of stable convergence in the Supplementary Material.

3.3. *Uniform integrability condition*

Assumption 2 is unwieldy to work with in practice, as it imposes a condition on the entire predictive sampling procedure for $N > n$. Fortunately, we have the following helpful result, which allows us to check Assumption 2 by considering only a single time step.

Proposition 1. *Let $Z(\theta, Y) = \mathcal{I}(\theta)^{-1} s(\theta, Y)$, where $Y \sim p_\theta$. Suppose that there exist nonnegative constants $B, C < \infty$ such that the following holds for all $\theta \in \Theta \subseteq \mathbb{R}$:*

$$E_\theta\{Z(\theta, Y)^4\} \leqslant B + C\theta^4.$$

Then Assumption 2 is satisfied.

The intuition is that the above condition, in combination with (2), implies that $\sup_{N>n} E(Z_N^4) < \infty$, which is sufficient for the uniform integrability of $(Z_N^2)_{N>n}$. Although it is likely that bounding $E_\theta\{Z(\theta, Y)^{2+\delta}\}$ for some $\delta > 0$ suffices, choosing $\delta = 2$ is particularly straightforward due to the recursive form of θ_N. The above condition is relatively easy to check and depends only on the predictive model p_θ; we now demonstrate this in the following examples.

Example 1. Here, we verify Assumption 2 in a setting where $Z(\theta, Y)$ is unbounded. Suppose that our parametric model is the normal model with mean zero and unknown variance, $p_\theta(y) = \mathcal{N}(0, \theta)$, where $\theta \in \mathbb{R}^+$. The score function and Fisher information are respectively

$$s(\theta, y) = \frac{y^2 - \theta}{2\theta^2}, \qquad \mathcal{I}(\theta) = \frac{1}{2\theta^2}.$$

Then $\mathcal{I}(\theta)^{-1}$ is continuous in θ, as required. We highlight that $\mathcal{I}(\theta), \mathcal{I}(\theta)^{-1}$ and $Z(\theta, Y)$ are all unbounded, which is the source of our difficulties. We can apply Proposition 1 by computing $E\{Z(\theta, Y)^4\}$. Here, we have $Z(\theta, Y)^4 = (Y^2 - \theta)^4$. Since $Y^2/\theta \sim \chi^2(1)$, which has a fourth central moment equal to 60, we obtain

$$E_\theta\{(Y^2 - \theta)^4\} = 60\,\theta^4$$

for all $\theta \in \mathbb{R}^+$. We can thus apply Proposition 1 with $B = 0$ and $C = 60$.

Example 2. We now verify Assumption 2 for an example that is not in the exponential family and is, in particular, heavy tailed. Consider a location-scale Student-t distribution with location θ, scale τ and degrees of freedom ν, with density

$$p_\theta(y) = \frac{\Gamma\{(\nu+1)/2\}}{\Gamma(\nu/2)\tau\,(\pi\nu)^{1/2}}\left\{1+\frac{1}{\nu}\left(\frac{y-\theta}{\tau}\right)^2\right\}^{-(\nu+1)/2}.$$

For simplicity, we assume that $\tau = 1$. The update term can then be computed as

$$Z(\theta, Y) = \frac{(\nu+3)(Y-\theta)}{\nu+(Y-\theta)^2},$$

which follows from standard results; see, e.g., Lange et al. (1989). We can then compute

$$Z(\theta, Y)^4 = \frac{(\nu+3)^4(Y-\theta)^4}{\{\nu+(Y-\theta)^2\}^4} \leqslant \frac{(\nu+3)^4}{16\nu^2},$$

where we have used the fact that $x^2/(\nu+x^2)^2 \leqslant (4\nu)^{-1}$ for $\nu > 0$. Here, there is no dependence on θ, so we can immediately apply Proposition 1 with $C = 0$.

3.4. *Accelerating predictive resampling*

Predictive resampling with (1) requires the truncation at some finite N in order be computationally feasible, and, while convergence appears relatively quick, the computational cost still grows linearly with N. The key idea of this section is that the form of Theorem 1 allows us to approximate

$$\theta_\infty \approx \theta_N + V_N\varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, 1)$ and the equality is in distribution. This will hopefully allow us to obtain accurate samples for a smaller value of N, and also to approximate the gap between θ_∞ and θ_N.

However, a minor difficulty is that we require a way of approximating the random variance V_N^2, which depends on the future, yet-to-be-sampled values of $Y_{N+1:\infty}$. As the aim is to avoid sampling $Y_{N+1:\infty}$ in order to save computation, we instead require an approximation, as also discussed in Fortini & Petrone (2020). A suitable approximation is

$$\hat{V}_N^2 = \mathcal{I}(\theta_N)^{-1}\sum_{i=N}^{\infty} i^{-2}, \tag{5}$$

where we have assumed that $\mathcal{I}(\theta_N)^{-1}$ is sufficiently close to $\mathcal{I}(\theta_\infty)^{-1}$, which seems reasonable given that $\mathcal{I}(\theta)^{-1}$ is continuous. As $\mathcal{I}(\theta_N)^{-1}$ is already computed as the gradient preconditioner, no additional computation is required. The following corollary justifies this usage.

COROLLARY 1. *Under Assumptions 1 and 2, we have*

$$\hat{V}_N^{-1}(\theta_\infty - \theta_N) \to \mathcal{N}(0, 1)$$

in distribution.

Proof. Since $\mathcal{I}(\theta)^{-1}$ is continuous, we have $\mathcal{I}(\theta_N)^{-1} \to \mathcal{I}(\theta_\infty)^{-1}$ almost surely, and hence $V_N/\hat{V}_N \to 1$ almost surely. Applying Slutsky's theorem to Theorem 1 then yields the result. As the right-hand side of (4) contains no random terms, stable convergence does not play a role here. □

Fortini & Petrone (2020, 2025) and Fong & Yiu (2025) suggested avoiding predictive resampling entirely, and instead directly using the distribution of $\theta_n + \hat{V}_n\varepsilon$ as an approximation to the martingale posterior, where n is the number of real observations. However, we emphasize that the Gaussian approximation is inaccurate if n is too small, and, crucially, the number of real observations n cannot be increased. The key idea underlying our method is that we are free to impute as many simulated data points as desired, subject to a computational budget. We can therefore impute $Y_{n+1:N}$ until we reach a point at which the Gaussian approximation is accurate. We thus recommend a hybrid scheme, in which regular predictive resampling is carried out up to a moderate value of N within the computation budget, and predictive resampling is then completed by adding $\hat{V}_N\varepsilon$. This is illustrated in the following algorithm.

Algorithm 1. Hybrid predictive resampling for a single posterior sample.

Estimate θ_n from $y_{1:n}$
For $i = n+1$ to $i = N$
  Draw $Y_i \sim p_{\theta_{i-1}}$
  $\theta_i \leftarrow \theta_{i-1} + i^{-1}Z_i$
Draw $\varepsilon \sim \mathcal{N}(0,1)$
$\hat{\theta}_\infty \leftarrow \theta_N + \hat{V}_N\varepsilon$
Output $\hat{\theta}_\infty$

The key point is that N can be chosen to be much smaller than would otherwise be required, as the remaining unsampled variation can be approximated by a Gaussian variable with random variance. One potential guideline for choosing N is to set $N + n = Kp$, where p is the dimensionality of θ and K is a chosen constant, for example $K = 100$. We may wish to choose K sufficiently large for the asymptotic approximation to hold well, although this remains an open question. We now demonstrate in a simulation example that the hybrid scheme substantially outperforms truncation alone and the direct Gaussian approximation.

Example 3. Consider the exponential distribution as the predictive $p_\theta(y) = \mathrm{Exp}(\theta)$, where θ is the scale parameter. The estimate θ_N is simply the mean, which admits the recursive update

$$\theta_N = \theta_{N-1} + N^{-1}(Y_N - \theta_{N-1}).$$

We can readily apply Proposition 1 and compute (5); see §H.2 of the Supplementary Material. To demonstrate the benefits of the hybrid scheme, we simulate $n = 10$ data points from $\mathrm{Exp}(\theta = 1)$, and compare the parametric martingale posterior obtained using the hybrid scheme with truncation at $N = 30$ to the ground-truth approximation obtained by truncating at a large value of $N = 20\,000$ (which is essentially exact numerically). As baselines, we also consider the direct Gaussian approximation of Fortini & Petrone (2020, 2025) without predictive resampling, and the truncated parametric martingale posterior at $N = 30$ without the Gaussian correction. Figure 1 shows kernel

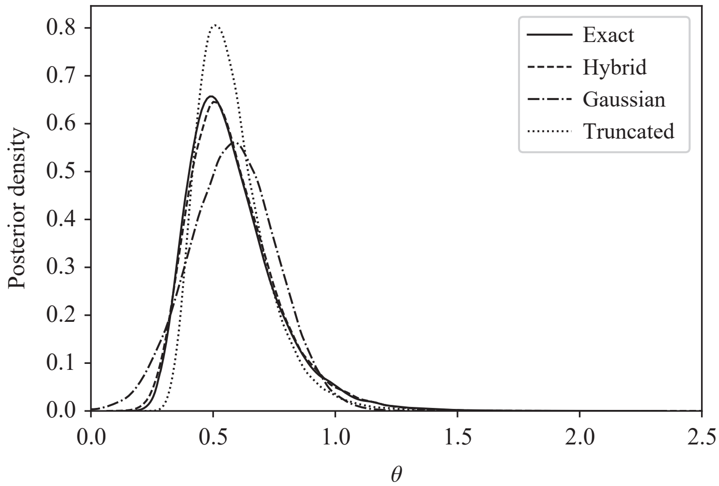

Fig. 1. Kernel density estimate plots of $B = 50\,000$ parametric martingale posterior samples for the exact (solid), hybrid (dashed), truncated (dot) and Gaussian (dash–dot) sampling schemes.

density estimates of $B = 50\,000$ parametric martingale posterior samples for the various sampling schemes. The hybrid and exact scheme are very close, while the truncated scheme is noticeably too narrow, and the Gaussian approximation is inaccurate due to the skewness of the martingale posterior as n is small. In terms of computation time on an Apple M2 Pro chip, the exact sampler required 6.4 s, whereas the truncated and hybrid samplers required 6.3×10^{-4} s, indicating a massive speed-up roughly equal to the ratio of the truncation values. This example illustrates that the first few imputed samples account for most of the non-Gaussianity of the martingale posterior, resulting in an accurate approximation via the hybrid scheme. If n is increased, the direct Gaussian approximation would be more accurate; however, the crucial point is that, for a given dataset, we cannot increase n, whereas N can be increased through predictive resampling.

4. Bernstein–von Mises theorem

4.1. *Frequentist asymptotics*

In this section, we turn to a frequentist study of the parametric martingale posterior. We now have Y_i independently and identically distributed from P^* for $i = 1, \ldots, n$, where we assume that the model is well specified, i.e., there exists θ^* such that $p^* = p_{\theta^*}$. We consider model misspecification in § 5.3 below. We would then like to study the martingale posterior as $n \to \infty$. We begin by illustrating posterior consistency, before establishing a Bernstein–von Mises theorem.

Again, we consider the update (1) for predictive resampling. The notation is slightly more involved due to the two data regimes for $Y_{1:n}$ and $\tilde{Y}_{n+1:\infty}$, where we now use $\tilde{Y}$ and double indexing to distinguish between simulated and observed data. The index n corresponds to the observed samples, whilst i indexes the imputed data. Consider the parametric martingale posterior starting at θ_n with sample size n; that is, for $i \geqslant 1$,

$$\tilde{Y}_{ni} \sim p_{\theta_{n,i-1}}, \qquad \theta_{ni} = \theta_{n,i-1} + (n+i)^{-1}\mathcal{I}(\theta_{n,i-1})^{-1}s(\theta_{n,i-1}, \tilde{Y}_{ni}),$$

where $\theta_{n0} = \theta_n$ and $(\tilde{Y}_{ni})_{i\geqslant 1}$ denotes the simulated population after observing n observations from P^*. As before, we write the update term, or natural gradient, as $Z_{ni} = \mathcal{I}(\theta_{n,i-1})^{-1}s(\theta_{n,i-1}, \tilde{Y}_{ni})$ for shorthand.

As we are interested in the frequentist properties of the martingale posterior, we first consider taking $i \to \infty$, where $\theta_{n\infty}$ corresponds to the almost-sure limit of the martingale for each n, assuming that the appropriate conditions for martingale convergence hold. We can then study the conditional distribution of $\theta_{n\infty}$ as $n \to \infty$. We highlight that, while the notation $\theta_{n\infty}$ may differ from traditional Bayesian notation, our object of interest is the same: $\theta_{n\infty}$ is the random parameter distributed according to the posterior. Our results in this section are also distinct from those in § 3, as the Y_i are now independently and identically distributed from P^*. To that end, we make the following assumption.

Assumption 3. Suppose that $\mathcal{I}(\theta)^{-1}$ is continuously differentiable in a neighbourhood of θ^*. Furthermore, assume that the conditions of Proposition 1 hold.

Intuitively, the first part of Assumption 3 ensures that the variance of the update Z_{ni} does not change too drastically with the initial θ_n. As before, the second part prevents the variance of the martingale posterior from growing too quickly under predictive resampling. As shown in Lemma 1, $\theta_{n\infty}$ exists almost surely, since $(\theta_{ni})_{i\geqslant 1}$ is a martingale in L_2 for each $n \geqslant 1$.

In the proofs of the results to come, it is in fact sufficient for the array $(Z_{ni}^2)_{i\geqslant 1, n\geqslant 1}$ to be uniformly integrable, but this condition is more arduous than Assumption 2, as it also depends on the sequence θ_n. We therefore opt to work with the conditions of Proposition 1, which are simpler to verify and imply the uniform integrability condition when $(\theta_n)_{n\geqslant 1}$ is convergent; see Proposition E.1 in the Supplementary Material for further details. In practice, checking Assumption 3 is the same exercise as for the predictive asymptotic results, as demonstrated earlier.

4.2. *Posterior consistency*

In order to obtain posterior consistency, we merely require that the sequence $(\theta_n)_{n\geqslant 1}$ is a consistent estimator. For example, consistency of the stochastic approximation estimator θ_n via (1) is usually guaranteed under the following assumption.

Assumption 4. The expected score function, $M(\theta) = \int s(\theta, y)\,\mathrm{d}P^*(y)$, exists and has a unique zero at θ^*. Furthermore, $M(\theta) > 0$ if $\theta < \theta^*$, and $M(\theta) < 0$ if $\theta > \theta^*$.

Under Assumption 4 and additional regularity conditions, for θ_n estimated via (1), we can show that $\theta_n \to \theta^*$, $P^{*\infty}$-almost surely, using the almost supermartingale theorem of Robbins & Siegmund (1971). Alternatively, we may take θ_n to be a well-understood point estimator with good properties, such as the maximum likelihood estimator; see § E.5 of the Supplementary Material for more details.

We follow a similar approach to Fong & Yiu (2025) to study the consistency of the parametric martingale posterior, based on Markov's inequality.

Proposition 2. *Let $(\theta_n)_{n\geqslant 1}$ be a sequence of estimators, where θ_n is estimated from $Y_i \overset{\text{i.i.d.}}{\sim} P^*$ for $i = 1, \ldots, n$, and suppose that it is strongly consistent at θ^*. Under Assumptions* 1 *and* 3*, the* *martingale posterior is strongly consistent at θ^*, that is, $\Pi(\theta_{n\infty} \in U^c \mid Y_{1:n}) \to 0$ as $n \to \infty$, $P^{*\infty}$-almost surely, for every neighbourhood U of θ^*.*

Proof. Let $\varepsilon > 0$ denote the size of the neighbourhood U. By Markov's inequality,

$$\Pi(|\theta_{n\infty} - \theta^*| > \varepsilon \mid Y_{1:n}) \leqslant \frac{1}{\varepsilon^2} E\{(\theta_{n\infty} - \theta^*)^2 \mid Y_{1:n}\}.$$

The expectation can be decomposed as

$$E\{(\theta_{n\infty} - \theta^*)^2 \mid Y_{1:n}\} = E\{(\theta_{n\infty} - \theta_n)^2 \mid Y_{1:n}\} + (\theta_n - \theta^*)^2,$$

where the cross-term is zero by the martingale property, and the first term is simply the posterior variance. In Lemma E.1 in the Supplementary Material, we show that the posterior variance converges to 0 as $n \to \infty$. Finally, $\theta_n \to \theta^*$, $P^{*\infty}$-almost surely, which yields the result. □

We see that posterior consistency follows from the posterior variance converging to 0 and consistency of the posterior mean θ_n. The above result can be weakened to convergence in P^{*n}-probability if θ_n is only weakly consistent. Finally, Fong & Yiu (2025) extended these arguments to derive posterior contraction rates in the nonparametric case; however, we omit this for brevity and focus directly on the Bernstein–von Mises result.

4.3. *Asymptotic normality*

We now turn to the Bernstein–von Mises result. To begin, we consider a deterministic sequence $(\theta_n)_{n\geqslant 1}$ that converges to θ^*, before addressing the case in which θ_n is estimated from $Y_i \overset{\text{i.i.d.}}{\sim} P^*$ for $i = 1, \ldots, n$. We then have the following main result.

Theorem 2. *Let $(\theta_n)_{n\geqslant 1}$ be a deterministic sequence that converges to θ^* as $n \to \infty$. Under Assumptions 1 and 3, we have*

$$r_n^{-1}(\theta_{n\infty} - \theta_n) \to \mathcal{N}\{0, \mathcal{I}(\theta^*)^{-1}\}$$

in distribution as $n \to \infty$, where $r_n^2 = \sum_{i=n+1}^{\infty} i^{-2}$.

Proof. The key is to construct an infinite martingale array, where each row is the sequence $(\theta_{ni})_{i\geqslant 1}$ arising from predictive resampling starting at $\theta_{n0} = \theta_n$. In this setting, we have one imputed population for each θ_n. Assumption 3 then implies the required Lindeberg condition and variance convergence result for the martingale central limit theorem for infinite arrays (e.g., Hall & Heyde (1980, Theorem 3.6)). Full details of the proof are provided in § E of the Supplementary Material. □

We now extend the above result to the case where θ_n is estimated from data and is therefore no longer deterministic. In what follows, we require strong consistency of the estimator due to the dependence on a deterministic sequence when applying the martingale central limit theorem; see the beginning of § 4.2 for a discussion of choices of θ_n. We leave the extension to weakly consistent estimators to future work, as we believe it would require a novel proof technique.

Theorem 3. *Let $(\theta_n)_{n\geqslant 1}$ be a sequence of estimators, where θ_n is estimated from $Y_i \overset{\text{i.i.d.}}{\sim} P^*$ for $i = 1, \ldots, n$, and suppose that it is strongly consistent at θ^*. Let $\Pi\{r_n^{-1}(\theta_{n\infty} - \theta_n) \in \cdot \mid Y_{1:n}\}$*

denote the posterior law of $r_n^{-1}(\theta_{n\infty} - \theta_n)$. *Under Assumptions 1 and 3, we have*

$$\Pi\{r_n^{-1}(\theta_{n\infty} - \theta_n) \in \cdot \mid Y_{1:n}\} \to \mathcal{N}\{0, \mathcal{I}(\theta^*)^{-1}\}$$

weakly $P^{*\infty}$*-almost surely, as* $n \to \infty$, *where* $r_n^2 = \sum_{i=n+1}^{\infty} i^{-2}$.

Proof. By strong consistency, there exists an event A with $P^{*\infty}(A) = 1$ such that, for all $y_{1:\infty} \in A$, the sequence $\{\theta_n(y_{1:n})\}_{n\geqslant 1}$ converges to θ^*. Fix any sequence $y_{1:\infty} \in A$. For each n, the distribution $\Pi[r_n^{-1}\{\theta_{n\infty} - \theta_n(y_{1:n})\} \in \cdot \mid y_{1:n}]$ is precisely the law of $r_n^{-1}\{\theta_{n\infty} - \theta_n(y_{1:n})\}$ under predictive resampling starting at $\theta_{n0} = \theta_n(y_{1:n})$. By Theorem 2, it follows that $\Pi[r_n^{-1}\{\theta_{n\infty} - \theta_n(y_{1:n})\} \in \cdot \mid y_{1:n}] \to \mathcal{N}\{0, \mathcal{I}(\theta^*)^{-1}\}$ weakly as $n \to \infty$. The claim therefore holds for $P^{*\infty}$-almost all $y_{1:\infty}$. □

4.4. *Centring and asymptotic variance*

For traditional Bayesian models, Bernstein–von Mises results are centred around an efficient sequence of estimators; that is, any sequence $\hat{\theta}_n$ satisfying

$$n^{1/2}(\hat{\theta}_n - \theta^*) = n^{-1/2} \sum_{i=1}^{n} \mathcal{I}(\theta^*)^{-1} s(\theta^*, Y_i) + o_{P^*}(1).$$

For such a sequence, we have $n^{1/2}(\hat{\theta}_n - \theta^*) \to \mathcal{N}\{0, \mathcal{I}(\theta^*)^{-1}\}$ in distribution, and since the asymptotic variance $\mathcal{I}(\theta^*)^{-1}$ matches that of the posterior, we deduce that central credible sets are also asymptotic confidence sets (e.g., Van der Vaart, 2000). The efficient centring sequence arises automatically from the shape of the likelihood rather than from the choice of prior. In contrast, our results above are always centred on the chosen initial estimate θ_n computed from $Y_{1:n}$. As a result, any guarantees with respect to θ^* depend solely on properties of the estimator θ_n. In this section, we address the implications of this different form of centring and assess the asymptotic variance for coverage.

We continue to assume that the model is well specified. The normalizing rate r_n can be replaced by $n^{-1/2}$, as the two are asymptotically equivalent. Theorem 3 then states that $\pi(\theta_{n\infty} \mid Y_{1:n}) \approx \mathcal{N}\{\theta_n, n^{-1}\mathcal{I}(\theta^*)^{-1}\}$. We therefore obtain correct coverage of posterior credible intervals, as in the usual Bernstein–von Mises result, if $r_n^{-1}(\theta_n - \theta^*) \to \mathcal{N}\{0, \mathcal{I}(\theta^*)^{-1}\}$ in distribution as $n \to \infty$. Fortunately, asymptotic normality of maximum likelihood estimates (Van der Vaart, 2000), as well as estimates obtained via stochastic approximation (Lai, 2003), is well studied. Under correct model specification and additional regularity conditions, we can therefore show that θ_n obtained from (1) satisfies the required central limit theorem. As these results are standard results, we defer further discussion to the Supplementary Material.

Additionally, our result highlights that preconditioning by $\mathcal{I}(\theta)^{-1}$ in (1) determines the magnitude of the asymptotic posterior variance, which is crucial for good frequentist properties of the induced parametric martingale posterior when θ_n is an efficient estimator. Meanwhile, the learning rate N^{-1} in (1) ensures that the posterior variance shrinks at rate n^{-1}, so our recommended choice is again appropriate in this setting.

In summary, for parametric martingale posteriors, there is a noticeable divide between posterior centring and posterior uncertainty, as the sequence θ_n need not necessarily correspond to the update rule (1) for Theorem 3 to hold. This contrasts with traditional Bayesian models, where the point estimate is implied by the full posterior distribution.

Reassuringly, under correct model specification and in regular settings, the frequentist properties of traditional Bayesian posteriors and parametric martingale posteriors do indeed coincide.

5. EXTENSIONS

5.1. *Multivariate parameters*

While we have presented our results for the univariate case for clarity, all results extend to the multivariate case, where $\theta \in \mathbb{R}^p$; we provide the complete treatment in the Supplementary Material, with examples in § 6 below. The multivariate case requires a careful application of the Cramér–Wold device, but is otherwise very similar to the univariate case; that is, $(t^{\mathrm{T}}\theta_N)_{N \geqslant n}$ is now our (scalar-valued) martingale for each $t \in \mathbb{R}^p$, and we leverage the same univariate martingale central limit theorems. The update term Z_N, as in (3), is now a vector that depends on the score vector and the Fisher information matrix, and we work with the norms $\|Z(\theta, Y)\|^4$ and $\|\theta\|^4$ for the analogous version of Proposition 1.

5.2. *Beyond independent and identically distributed*

It is possible to extend the parametric martingale posterior to certain dependent settings, such as regression or time series models (Walker, 2022; Holmes & Walker, 2023; Moya & Walker, 2025). In this section, we outline extensions of the theory and methods to the regression setting. Let $X \in \mathcal{X} \subseteq \mathbb{R}^p$, and $Y \in \mathcal{Y}$, where $\mathcal{Y} \subseteq \mathbb{R}$ or $\mathcal{Y} = \{0, 1\}$ in the continuous and binary cases, respectively, and let $p_\theta(y \mid x)$ denote the conditional model. We assume that $Y_i \mid X_i \sim p_{\theta^*}(y \mid x)$, and that $(X_i)_{i \geqslant 1}$ is a deterministic covariate sequence (i.e., the fixed-design setting).

As discussed in Holmes & Walker (2023), Fong et al. (2023) and Fortini & Petrone (2025), it is most natural to devote modelling efforts to the conditional distribution of $Y \mid X$. For predictive resampling, we then draw $X_{n+1:\infty}$ from the deterministic covariate sequence, or independently and identically from the empirical distribution of the observed covariates $X_{1:n}$ if the full sequence of covariates is not known. In the random-design case, we can use the Bayesian bootstrap to draw $X_{n+1:\infty}$, but we do not consider that here.

Predictive resampling then involves first drawing X_N from the chosen scheme, then subsequently drawing $Y_N \sim p_{\theta_{N-1}}(\cdot \mid X_N)$ and finally computing

$$\theta_N = \theta_{N-1} + N^{-1}\mathcal{I}(\theta_{N-1})^{-1} s(\theta_{N-1}, Y_N; X_N),$$

where $s(\theta, Y; X) = \nabla_\theta \log p_\theta(Y \mid X)$ and

$$\mathcal{I}(\theta; X) = E_\theta\{s(\theta, Y; X)\, s(\theta, Y; X)^{\mathrm{T}} \mid X\}, \qquad \mathcal{I}(\theta) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}(\theta; X_i).$$

One can verify quite easily that the choice of method used to impute $X_{n+1:\infty}$ does not impede the martingale condition. In practice, if $\mathcal{I}(\theta)$ is not known, we can approximate it by

$$\hat{\mathcal{I}}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{I}(\theta; X_i).$$

In this case, $\hat{\mathcal{I}}_n(\theta)$ is estimated once from the observed $X_{1:n}$ and held fixed during predictive resampling. To obtain the central limit theorems, we can simply apply the multivariate extensions of Theorems 1 and 3, where the martingale is with respect to the filtration generated by the pairs $(Y_N, X_N)_{N>n}$. We provide detailed results in § G of the Supplementary Material, along with an additional logistic regression example.

5.3. *Model misspecification*

We highlight that we do not explicitly use the fact that $p^* = p_{\theta^*}$ in the proof of Theorem 2, as outlined in the Supplementary Material. As a result, Theorems 2 and 3 continue to hold as long as the limit θ^* satisfies Assumption 3. In the model misspecification setting, under appropriate regularity conditions, it would be natural to require that the sequence θ_n converges $P^{*\infty}$-almost surely to θ^* that minimises the Kullback–Leibler divergence between p^* and p_θ. For traditional Bayes, under regularity conditions, the asymptotic variance of the regular Bayesian posterior is $J(\theta^*)^{-1}$, where $J(\theta^*) = -E_{P^*}\{\partial s(\theta^*, Y)/\partial\theta\}$ (Kleijn & van der Vaart, 2012), which is not equal to $\mathcal{I}(\theta^*)^{-1}$ in the case of misspecification. On the other hand, as Theorem 3 still holds, the asymptotic variance of the parametric martingale posterior remains of the form $\mathcal{I}(\theta^*)^{-1}$, albeit at a different value of θ^*. We also highlight that the asymptotic variance of the parametric martingale posterior does not depend on P^*, which is unsurprising given the nature of predictive resampling. This provides another concrete demonstration of the separation between the centring and asymptotic variance of the parametric martingale posterior.

Under model misspecification, the asymptotic variance of the initial estimate θ_n will no longer match that of the parametric martingale posterior, so correct coverage will not be attained; this also occurs in the Bayesian case. However, the additional flexibility of the learning rate in (1) offers a potential solution, analogous to a tempered likelihood in the regular Bayesian case (Holmes & Walker, 2017). Specifically, if we replace N^{-1} in (1) with aN^{-1} for $a \in \mathbb{R}^+$, we can appropriately inflate or shrink $\mathcal{I}(\theta^*)$ to match the asymptotic variance of the initial estimate θ_n. As condition (2) will still be satisfied, all of our results continue to hold. This can also be extended to the multivariate case, where a is a chosen positive-definite matrix rather than a scalar. Unlike tempering in regular Bayesian inference, this provides the freedom to adjust the entire covariance structure of the posterior distribution, rather than merely applying a uniform rescaling (as in, e.g., Lyddon et al., 2019).

6. Empirical studies

6.1. *Set-up*

We implement the parametric martingale posterior in the Julia programming language (Bezanson et al., 2017), which offers high efficiency for iterative updates. For baseline comparators requiring Markov chain Monte Carlo, we rely on the efficient No-U-Turn Sampler (Hoffman et al., 2014) implemented in the `Turing.jl` package (Ge et al., 2018) in Julia. All experiments are run on an Apple M2 Pro CPU.

6.2. *Simulations*

This section presents a simulation study that compares the parametric martingale posterior with traditional Bayesian inference in a bivariate normal model. For $y \in \mathbb{R}^2$, our predictive model is $p_\theta(y) = \mathcal{N}(y \mid \mu, \Sigma)$, where $\mu = [\mu_1, \mu_2]^{\mathrm{T}}$ is a bivariate vector and Σ is a 2×2 covariance matrix with entries $\Sigma_{11} = s_1$, $\Sigma_{22} = s_2$ and $\Sigma_{12} = \Sigma_{21} = s_{12}$. We thus consider $\theta = (\mu_1, \mu_2, s_1, s_2, s_{12})$. For the parametric martingale posterior, it is not too challenging to verify that the natural gradient takes the form $Z_{\mu_j}(\theta, Y) = Y_j - \mu_j$ for the means, $Z_{s_j}(\theta, Y) = (Y_j - \mu_j)^2 - s_j$ for the variances and $Z_{s_{12}}(\theta, Y) = (Y_1 - \mu_1)(Y_2 - \mu_2) - s_{12}$ for the covariance. For hybrid predictive resampling, $\mathcal{I}(\theta)^{-1}$ can be easily computed, and the conditions required for the multivariate versions of Theorems 1 and 3 can be verified. We omit these details for brevity, with full derivations provided in the Supplementary Material. For the initial estimate θ_n, we use the sample estimates of μ and Σ, as they are unbiased, easy to implement and negligibly different from those obtained via (1).

For the regular Bayesian model, we opt for weakly informative priors for all parameters (Gelman, 2006; Gelman et al., 2008). In particular, we have

$$\mu_j \sim \mathcal{N}(0, 10^2), \qquad s_j^{1/2} \sim \text{ Half-Cauchy}(0, 5^2), \qquad \rho \sim \text{Un}(-1, 1),$$

where $s_{12} = \rho(s_1 s_2)^{1/2}$. We highlight that it is not straightforward to elicit prior distributions on covariance matrices that are noninformative or weakly informative (Fuglstad et al., 2020; Pinkney, 2024). By contrast, such prior elicitation is not a necessary consideration for the parametric martingale posterior. Furthermore, due to the martingale, the posterior mean of the martingale posterior is always equal to the initial estimate θ_n.

We set the truth to $\theta^* = (-0.5, 1.0, 1.0, 0.5, 0.7)$ and evaluate the coverage and average length of 95% credible intervals for different sample sizes $n = (20, 50, 500)$. For each setting of n, we consider 5000 sampled datasets for the parametric martingale posterior. Owing to the computational cost associated with Markov chain Monte Carlo, we consider only 500 sampled datasets for the Bayesian posterior. We generate $B = 2000$ posterior samples for both methods and use a truncation of $N = n + 50$ for hybrid predictive resampling.

Table 1 illustrates the simulation results for (μ_1, s_1, s_{12}), with results for (μ_2, s_2) reported in the Supplementary Material, as they are practically identical. We see that the parametric martingale posterior slightly undercovers for $n = 20$, but attains accurate coverage as n increases, in agreement with Theorem 3. The Bayesian posterior, on the other hand, attains coverage for all n at the cost of noticeably wider intervals. This discrepancy between the two posteriors disappears as n increases, which is unsurprising given that they are asymptotically equivalent. This agrees with the observation of Moya & Walker (2025) that the Bayesian and martingale posterior variances differ by order n^{-2}, which is likely attributable to the prior distribution. These results highlight the value of a well-specified prior when n is small; however, a poorly specified prior distribution would substantially harm coverage. The parametric martingale posterior may therefore be more suitable when n is of reasonable size or when little prior information is available, and may be viewed as a default or 'frequentist' posterior (Holmes & Walker, 2023).

A much starker contrast appears in computation time, where the parametric martingale posterior requires orders of magnitude less time than traditional Bayes. This highlights the substantial computational gains made possible by sidestepping Markov chain Monte Carlo. Another benefit of predictive resampling is that the samples are independent, unlike those obtained via Markov chain Monte Carlo. Finally, the runtime for the martingale posterior does not increase substantially with n, and in theory we could have chosen N independently of n. The computation time of the martingale posterior is thus bottlenecked by the time required for the initial estimation of θ_n, which is still negligible at the above sample sizes, but may become noticeable for sufficiently large n. We provide further details on computational complexity in the Supplementary Material.

Table 1. *Coverage and average length of 95% credible intervals for the traditional Bayesian posterior (Markov chain Monte Carlo) and the parametric martingale posterior (hybrid predictive resampling) at different sample sizes*

		$n = 20$		$n = 100$		$n = 500$	
Method	Parameter	Coverage	Length	Coverage	Length	Coverage	Length
Bayesian posterior	μ_1	93.6	9.2	92.6	3.9	94.2	1.7
	s_1	94.8	15.3	95.8	5.8	93.0	2.5
	s_{12}	95.2	10.8	95.4	4.1	93.6	1.8
	Max. SE	1.1	0.2	1.1	0.04	1.1	0.008
	Runtime	1.2 s		3.7 s		14.7 s	
Martingale posterior	μ_1	93.0	8.6	93.3	3.9	94.4	1.7
	s_1	91.3	12.0	94.4	5.5	94.8	2.5
	s_{12}	90.0	8.1	94.1	3.8	94.6	1.7
	Max. SE	0.4	0.06	0.4	0.01	0.3	0.002
	Runtime	0.003 s		0.003 s		0.003 s	

SE, standard error. Coverage is in %; length has been multiplied by 10; runtime is reported per sample. Results are over 500 and 5000 repeats for Bayes and the martingale posterior, respectively.

6.3. *Real data example*

We demonstrate the parametric martingale posterior on a robust regression task, using the AIDS Clinical Trials Group Study 175 dataset from the UCI Machine Learning Depository (Asuncion & Newman, 2007; Hammer et al., 1996). The dataset consists of $n = 2139$ patients with HIV who were randomized to four treatment arms: (i) zidovudine + didanosine; (ii) zidovudine + zalcitabine; (iii) didanosine and (iv) zidovudine. Following a similar set-up to Hines et al. (2025) and Li et al. (2023), we consider the continuous outcome to be the CD4 count at 20 ± 5 weeks, which is a measure of immune function after some time on treatment. In addition to the treatment group variable, we consider 12 baseline variables, listed in the Supplementary Material. For pre-processing, we normalize all continuous covariates and the outcome to have mean 1 and standard deviation 1, and introduce three dummy variables for the treatment group variable, with zidovudine as the reference category. With the intercept, our design matrix consists of $p = 16$ variables.

A quick check reveals that the excess kurtosis of CD4 count at baseline is equal to 1.8, indicating that the outcome may also be heavy tailed. We therefore opt to use a heavy-tailed linear Student-t model (Lange et al., 1989; Geweke, 1993). Let $\theta = [\beta, \tau^2]^{\mathrm{T}}$, where $\beta \in \mathbb{R}^p$ denotes the regression coefficients (including an intercept) and $\tau \in \mathbb{R}$ is the scale parameter. Suppose that we observe $(Y_{1:n}, X_{1:n})$, where $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^p$. The model is then

$$Y_i = \beta^{\mathrm{T}} X_i + \tau \varepsilon_i, \qquad \varepsilon_i \sim \text{Student-}t(\nu),$$

where $\nu > 1$ is the fixed degrees of freedom, giving $p_\theta(y \mid x)$ as the location-scale Student-t distribution. A natural plug-in predictive distribution $p_{\theta_n}(y \mid x)$, where θ_n is the maximum likelihood estimate, can then be used for the parametric martingale posterior. We highlight that, in this case, estimation of θ_n requires slightly more care due to the nonconcavity of the loglikelihood: we employ iteratively reweighted least squares (Lange et al., 1989), repeated with multiple initializations, to estimate the global maximum. The asymptotic theory then follows from standard M-estimation theory. Finally, we treat ν as a hyperparameter and set $\nu = 5$. In practice, ν could also be selected by maximizing the prequential loglikelihood

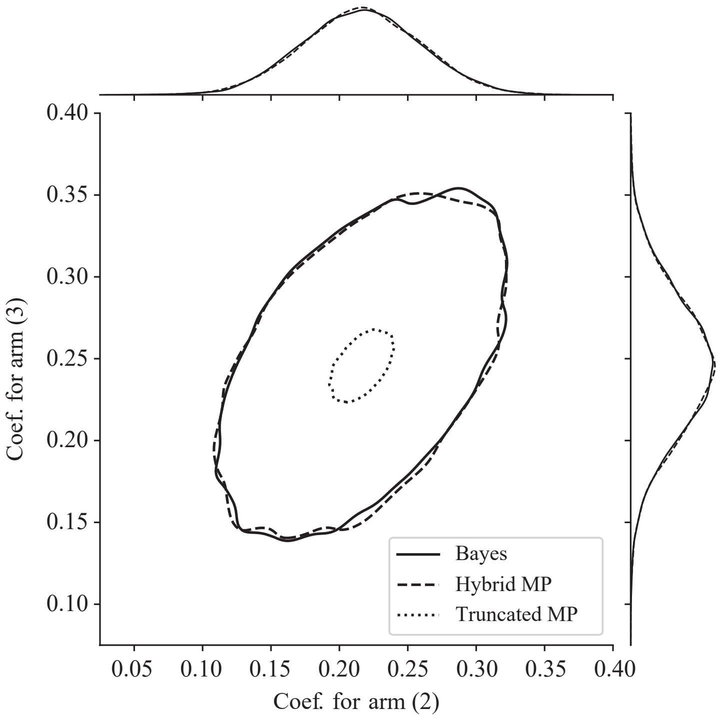

Fig. 2. Posterior 95% probability contours of the coefficients for treatment arms (2) and (3) for the Bayesian posterior (solid), martingale posterior (MP) with hybrid sampling (dashed) and with truncation at $N = n + 100$ (dotted). Marginal kernel density plots are shown above and to the right, with the truncated martingale posterior omitted for clarity.

(Dawid, 1984), $\sum_{i=1}^{n} \log p_{\theta_{i-1}}(Y_i \mid X_i)$, which is closely connected to the marginal likelihood (Gneiting & Raftery, 2007; Fong & Holmes, 2020).

For the parametric martingale posterior, we keep $\nu = 5$ fixed and set the initial estimate θ_n to be the maximum likelihood estimate obtained from 10 repeated optimizations. We truncate at $N = n + 50\,000$ and $N = n + 100$ for exact and hybrid predictive resampling, respectively. One can show that the natural gradient is $Z_n(\theta, Y; X) = [Z_\beta(\theta, Y; X)^{\mathrm{T}},\ Z_{\tau^2}(\theta, Y; X)]^{\mathrm{T}}$, where

$$Z_\beta(\theta, Y; X) = \left\{\frac{\tau(\nu+3)R}{\nu+R^2}\right\}\Sigma_{n,x}^{-1}X, \qquad Z_{\tau^2}(\theta, Y; X) = \frac{\tau^2(\nu+3)(R^2-1)}{\nu+R^2},$$

for $R = (Y - \beta^T X)/\tau$ and $\Sigma_{n,x} = n^{-1}\sum_{i=1}^{n} X_i X_i^{\mathrm{T}}$. We show in the Supplementary Material that the above update satisfies all conditions for the regression extension of Theorem 1, provided that $\Sigma_{n,x}$ is positive definite. Under weak assumptions on the sequence of design points $X_{1:n}$, together with additional regularity conditions ensuring consistency of the maximum likelihood estimator, we also obtain the regression extension of Theorem 3. For the Bayesian model, we set the same value of $\nu = 5$ and again elicit the weakly informative priors,

$$\beta_j \sim \mathcal{N}(0, 10^2), \qquad \tau \sim \text{Half-Cauchy}(0, 5^2).$$

We generate $B = 10\,000$ posterior samples in both cases.

Table 2. *Runtimes on the ACTG 175 dataset*

Method	*Runtime*(*s*)
Bayes	132.4
Martingale posterior ($N = n + 50\,000$)	14.4
Martingale posterior (hybrid, $N = n + 100$)	0.03

Figure 2 illustrates the posterior 95% contours for (β_j, β_k) corresponding to treatment arms (2) and (3) under the Bayes posterior, the martingale posterior with truncation $N = n+100$ and the hybrid scheme. We exclude the exact scheme ($N = n+50\,000$), as it is visually indistinguishable from the hybrid scheme; this comparison is provided in the Supplementary Material. We see that the martingale posterior truncated at $N = n + 100$ substantially underestimates the uncertainty, but this is corrected by the addition of the Gaussian correction in the hybrid scheme. The martingale posterior with hybrid predictive resampling is very similar to the Bayes posterior, which is unsurprising given the weakly informative prior. Both the Bayesian and martingale posteriors suggest that treatment arms (2) and (3) are more effective than the reference arm (4), in agreement with Hammer et al. (1996).

Table 2 illustrates the runtimes. Regular Bayes with Markov chain Monte Carlo required 132.4 s for posterior sampling. By contrast, the martingale posterior with $N = n + 50\,000$ and $N = n + 100$ required 14.4 and 0.03 s, respectively, with hybrid predictive resampling requiring a negligible additional amount of time compared with the latter. As before, the runtime for predictive resampling is roughly proportional to N, and the hybrid scheme is by a very large margin the fastest. In § I.3 of the Supplementary Material, we show that the computational complexity of the parametric martingale posterior in this example (assuming that $n \gg p$) is $O\{np^2 + B(N - n)p\}$. We also discuss the general case and compare this complexity with that of Markov chain Monte Carlo.

7. Discussion

In this work, we have leveraged martingale central limit theorems to accelerate predictive resampling and to provide frequentist guarantees for the parametric martingale posterior. The experiments demonstrate that the parametric martingale posterior can be significantly faster than traditional Bayes, with results that match that of a posterior arising from a noninformative prior. We now outline several interesting directions for future research. First, it would be of interest to consider how the parametric martingale posterior might be used when informative prior information is available, which could open up predictive resampling as a posterior approximation scheme. While our work focuses on the case where $n \gg p$, it would be of considerable interest to investigate both the practical and theoretical properties of the parametric martingale posterior in high-dimensional settings, where prior information or regularization is likely to play a key role. It would also be of interest to extend the above methodology to models with additional structure, such as hierarchical models, where Bayesian methods arguably have the most application. Finally, it would be interesting to examine whether the above results and the hybrid predictive resampling scheme can be applied to general nonparametric settings beyond those considered in Fong & Yiu (2025).

Acknowledgement

A.Y. received funding from Novo Nordisk. E.F. received funding from the Research Grants Council of Hong Kong through the Early Career Scheme (27304424) and the General Research Fund (17306925). We thank Chris Holmes, Ziyu Wang and Neil Chada for helpful discussions. We also thank the associate editor and two reviewers for their constructive feedback, which has substantially improved the paper.

Supplementary material

The Supplementary Material includes technical proofs and additional derivations.

References

Asuncion, A. & Newman, D. (2007). UCI machine learning repository. https://archive.ics.uci.edu [last accessed 22 October 2024].

Berti, P., Pratelli, L. & Rigo, P. (2004). Limit theorems for a class of identically distributed random variables. *Ann. Prob*. **32**, 2029–52.

Berti, P., Pratelli, L. & Rigo, P. (2021a). A central limit theorem for predictive distributions. *Mathematics* **9**, 3211.

Berti, P., Dreassi, E., Pratelli, L. & Rigo, P. (2021b). A class of models for Bayesian predictive inference. *Bernoulli* **27**, 702–26.

Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. (2017). Julia: a fresh approach to numerical computing. *SIAM Rev.* **59**, 65–98.

Bissiri, P. G., Holmes, C. C. & Walker, S. G. (2016). A general framework for updating belief distributions. *J. R. Statist. Soc. B* **78**, 1103–30.

Crimaldi, I. (2009). An almost sure conditional convergence result and an application to a generalized Pólya urn. *Int. Math. Forum* **4**, 1139–56.

Cui, F. & Walker, S. G. (2024). A Bayesian bootstrap for mixture models. *Bayesian Anal.* **1**, 1–28.

Dawid, A. P. (1984). Statistical theory: the prequential approach. *J. R. Statist. Soc. A* **147**, 278–90.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Ann. Inst. H. Poincaré* **7**, 1–68.

Doob, J. (1953). *Stochastic Processes*. New York: John Wiley.

Efron, B. (2012). Bayesian inference and the parametric bootstrap. *Ann. Appl. Statist*. **6**, 1971–97.

Favaro, S. & Fortini, S. (2024). Quasi-Bayesian sequential deconvolution. *arXiv:* 2408.14402v2.

Fong, E. & Holmes, C. C. (2020). On the marginal likelihood and cross-validation. *Biometrika* **107**, 489–96.

Fong, E., Holmes, C. & Walker, S. G. (2023). Martingale posterior distributions. *J. R. Statist. Soc. B* **85**, 1357–91.

Fong, E. & Yiu, A. (2025). Bayesian quantile estimation and regression with martingale posteriors. *J. R. Statist. Soc. B*, doi: 10.1093/jrsssb/qkaf080.

Fortini, S. & Petrone, S. (2020). Quasi-Bayes properties of a procedure for sequential learning in mixture models. *J. R. Statist. Soc. B* **82**, 1087–114.

Fortini, S. & Petrone, S. (2023). Prediction-based uncertainty quantification for exchangeable sequences. *Phil. Trans. R. Soc. A* **381**, 20220142.

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

Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H. & Riebler, A. (2020). Intuitive joint priors for variance parameters. *Bayesian Anal.* **15**, 1109–37.

Garelli, S., Leisen, F., Pratelli, L. & Rigo, P. (2024). Asymptotics of predictive distributions driven by sample means and variances. *arXiv:* 2403.16828v3.

Ge, H., Xu, K. & Ghahramani, Z. (2018). Turing: a language for flexible probabilistic inference. In *Proc. 21st Int. Conf. Artif. Intel. Statist*. (Proc. Mach. Learn. Res. **84**), pp. 1682–90. Brookline, Massachusetts, United States: JMLR.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1**, 515–34.

Gelman, A., Jakulin, A., Pittau, M. G. & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Statist.* **2**, 1360–83.

Geweke, J. (1993). Bayesian treatment of the independent Student-t linear model. *J. Appl. Economet*. **8**, S19–40.

Gneiting, T. & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Assoc*. **102**, 359–78.

Hall, P. & Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*. New York: Academic Press.

Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M. et al. (1996). A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New Engl. J. Med*. **335**, 1081–90.

Häusler, E. & Luschgy, H. (2015). *Stable Convergence and Stable Limit Theorems (Prob. Theory Stoch. Mod. 74)*. Cham: Springer.

Heyde, C. & Johnstone, I. (1979). On asymptotic posterior normality for stochastic processes. *J. R. Statist. Soc. B* **41**, 184–9.

Hines, O., Diaz-Ordaz, K. & Vansteelandt, S. (2025). Variable importance measures for heterogeneous treatment effects. *Biometrics* **81**, ujaf140.

Hoffman, M. D. & Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res*. **15**, 1593–623.

Holmes, C. C. & Walker, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika* **104**, 497–503.

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

Kleijn, B. & van der Vaart, A. (2012). The Bernstein-von Mises theorem under misspecification. *Electron. J. Statist.* **6**, 354–81.

Knoblauch, J., Jewson, J. & Damoulas, T. (2022). An optimization-centric view on Bayes' rule: reviewing and generalizing variational inference. *J. Mach. Learn. Res*. **23**, 1–109.

Lai, T. L. (2003). Stochastic approximation. *Ann. Statist.* **31**, 391–406.

Lange, K. L., Little, R. J. & Taylor, J. M. (1989). Robust statistical modeling using the *t* distribution. *J. Am. Statist. Assoc*. **84**, 881–96.

Li, H., Hubbard, A. & van der Laan, M. (2023). Targeted learning on variable importance measure for heterogeneous treatment effect. *arXiv:* 2309.13324v1.

Lyddon, S. P., Holmes, C. & Walker, S. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika* **106**, 465–78.

Martens, J. (2020). New insights and perspectives on the natural gradient method. *J. Mach. Learn. Res*. **21**, 1–76.

Moya, B. & Walker, S. G. (2025). Martingale posterior distributions for time-series models. *Statist. Sci.* **40**, 68–80.

Newton, M. A. & Zhang, Y. (1999). A recursive algorithm for nonparametric analysis with missing data. *Biometrika* **86**, 15–26.

Pinkney, S. (2024). A short note on a flexible Cholesky parameterization of correlation matrices. *arXiv:* 2405.07286v1.

Rényi, A. (1963). On stable sequences of events. *Sankhya* **25**, 293–302.

Robbins, H. & Siegmund, D. (1971). A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, Ed. J. S. Rustagi, pp. 233–57. New York: Academic Press.

Rubin, D. B. (1981). The Bayesian bootstrap. *Ann. Statist.* **9**, 130–4.

Van der Vaart, A. W. (2000). *Asymptotic Statistics (Camb. Ser. Statist. Prob. Math. 3)*. Cambridge: Cambridge University Press.

Walker, S. G. (2022). A new look at Bayesian uncertainty. In *Handbook of Statistics*, vol. **47**, Ed. A. S. R. S. Rao, G. A. Young and C. R. Rao, pp. 83–101. Amsterdam: Elsevier.

Wang, Z. & Holmes, C. C. (2024). On uncertainty quantification for near-Bayes optimal algorithms. *arXiv*: 2403.19381v2.

[*Received on* 22 *October* 2024. *Accepted on* 15 *December* 2025]