# Bayesian Inference Under Model Misspecification: A Modern Overview

Alika Grigorova

September 2025

## Table of Contents

# 1 Introduction

Bayesian statistics became widely used in many modern applications. The Bayesian approach allows one to take into account prior knowledge about the parameter of interest, to quantify uncertainty by returning a distribution rather than a single point estimate, and to fit complex models.

However, in complex modern settings, along with the benefits of the Bayesian framework, there also exist serious challenges. As Knoblauch et al. 2022 point out, when calculating the Bayesian posterior in a traditional setting, we make three very strong assumptions. (1) There exists some prior knowledge about the parameters of interest, which can be expressed mathematically as a probability distribution. (2) We assume that the model (likelihood) is well-specified, meaning that for some parameter $\theta^*$, $p(x \mid \theta^*)$ corresponds exactly to the true data distribution $p_{true}(x)$. (3) We assume that sufficient computational resources are available. These assumptions are often violated in complex real-world tasks.

The studies we focus on in this Literature Review describe the ways to address these challenges, particularly emphasizing model misspecifications. Model misspecification means that the assumed likelihood does not correspond to the true data-generating distribution for any value of the parameter. In such cases, the Bayesian posterior does not concentrate around a "true" parameter, but instead around a *pseudo-true parameter* $\theta_0$. The pseudo-true parameter is the value that makes the model distribution as close as possible to the true data-generating distribution, even if the model is misspecified and no parameter reproduces the truth exactly. It is defined as

$$\theta_0 = \arg\min_{\theta \in \Theta} \mathrm{KL}(p_{true}(x) \,\|\, p(x \mid \theta)) \,,$$

meaning that $\theta_0$ is the parameter value for which the likelihood $p(x \mid \theta)$ — the probability of observing the data given the value of the parameter — yields the smallest Kullback–Leibler divergence from the true data-generating distribution $p_{true}(x)$.

The resulting posterior is sometimes referred to as a *pseudo-posterior*. As Knoblauch et al. 2022 remark, “whenever the likelihood model is severely misspecified, inference outcomes suffer dramatically” (§3.4). This means that Bayesian procedures, when based on a misspecified model, can yield unreliable predictions.

Another challenge closely connected to model misspecification comes from the Bernstein–von Mises theorem, as stated by Lai and Yao 2024. The Bernstein–von Mises theorem states that when the sample size $n \to \infty$, the posterior distribution concentrates to a point mass around the maximum likelihood estimator (MLE). Such asymptotic behavior causes the posterior uncertainty about the parameters to vanish, which may lead to concentration around an incorrect point estimate, in case of a model being misspecified, or reduce the variability of predictions.

The following sections will provide a more detailed overview of the reviewed studies. At the end of some sections, potential research directions will be highlighted.

# 2 A General Framework for Updating Belief Distributions

In their study, Bissiri et al. 2016 underline that the traditional Bayesian approach to updating beliefs about the parameters of interest might be unreasonable and “cumbersome.”

The authors mention that it is often the case that the true data distribution $p_{true}(x)$ is simply unknown, or not contained in the model family $p(x \mid \theta)$ (which can be called an *M-open case*). In this situation, no choice of $\theta$ makes the model exactly correct, and inference can at best find a pseudo-true parameter that minimizes the discrepancy to $p_{true}(x)$. Even in cases where the model family is technically correct, the parameter $\theta$ may be ultra–high-dimensional, consisting mainly of nuisance components, while in fact only the small subset of parameters might be scientifically relevant. In both situations, the standard Bayesian update may be an unreasonable way to update beliefs about the parameters of interest, since it will either concentrate on a pseudo-true parameter that does not represent the data-generating process or overfit irrelevant components, in both cases leading to potentially incorrect predictions. By contrast, only in the *M-closed* case — when there exists some $\theta^*$ such that $p(x \mid \theta^*) = p_{true}(x)$ — is the classical Bayesian update reliable for making predictions. The authors therefore propose a different paradigm of Bayesian inference. Bissiri et al. 2016 begin by postulating that posterior updating has the form of some function $\psi$, of the prior distribution $\pi$ and a loss function $\ell(\theta, x)$,

$$q(\theta \mid x) \;=\; \psi\{\ell(\theta, x), \pi(\theta)\},$$

where $\theta$ are the parameters of interest, which come from a parameter space $\Theta$. The authors make several key assumptions about $\psi$, as follows:

1. Updating the posterior based on data $(x_1, x_2)$ jointly should give the same result as updating it with $x_1$ and then $x_2$ sequentially.

2. Updating the posterior with the prior restricted to a subset $A \subset \Theta$ gives the same result as restricting the posterior, obtained from the full prior, to $A$.

3. If the loss for some $\theta$ is strictly larger under data $x$ than under data $y$, then under the same prior $\pi$, the posterior probabilities on $\theta$ should be smaller given $x$ than given $y$.

4. If $\ell(\theta, x)$ is a constant, then the posterior should be equal to the prior.

5. Adding a constant $c$ to the loss, $\ell(\theta, x) + c$, should not change the posterior.

Under these assumptions, Bissiri et al. 2016 prove that the update of the posterior uniquely takes the form

$$q(\theta \mid x) \;\propto\; \exp\{-\ell(\theta, x)\}\, \pi(\theta).$$

They denote that the standard Bayesian update arises as the special case, when $\ell(\theta, x) = -\log p(x \mid \theta)$.

The authors further demonstrate that this posterior update can be equivalently formulated as a solution to an optimization problem. Bissiri et al. 2016 define $q^*$ as the minimizer of the loss function $L(q, \pi, x)$. They show that this function is a cumulative loss composed of two independent parts: a data-dependent loss over the parameter of interest and a prior-based loss. Here, the optimization is taken over $\mathcal{P}(\Theta)$ – the set of all probability measures on the parameter space $\Theta$, $\ell(\theta, x)$ is a user-specified loss function, and $\mathrm{KL}(q \,\|\, \pi)$ is the Kullback–Leibler divergence between the posterior and the prior. Formally, $q^*$ is defined as the solution to

$$q^* \;=\; \arg\min_{q \in \mathcal{P}(\Theta)} \left\{ \int \sum_{i=1}^{n} \ell(\theta, x_i)\, q(\theta)\, d\theta \;+\; \mathrm{KL}(q \,\|\, \pi) \right\}.$$

In this way, Bissiri et al. 2016 define the *generalized Bayesian posterior*, which does not require a full probabilistic model to be specified. The authors also show that standard Bayesian posterior can be recovered as a special case of the generalized Bayesian posterior, when the loss function $\ell$ is chosen as the negative log-likelihood. The authors point out that the main benefit of their framework is that one does not need a defined model for every data point, but only needs to construct a loss function considering only the parameters of interest.

**Choice of loss functions.** Bissiri et al. 2016, as already discussed above, note that in the M-closed case the traditional Bayesian update is rational, which corresponds to choosing $\ell(\theta, x) = -\log p(x \mid \theta)$ in the definition of the generalized Bayesian posterior. In the M-open case, the authors point out that one may still use the same loss $\ell = -\log p(x \mid \theta)$, but posterior will concentrate around a pseudo-true parameter (the member of the model family closest in Kullback–Leibler divergence to the true distribution).

Bissiri et al. 2016 also suggest using different types of loss functions, for example Hüber's $\rho$ loss, when the robustness to misspecifications and outliers is desired. However, when a choosing loss function other than negative log-likelihood, the problem of different scales between the loss component and the divergence component arises. The authors suggest a few ways to deal with it. Once a weighting parameter $w > 0$ is introduced, the posterior takes the form

$$q^* \;\propto\; \pi(\theta)\, \exp\Bigg( - w \sum_{i=1}^{n} \ell(\theta, x_i) \Bigg)\, d\theta.$$

Different calibration strategies correspond to different ways of selecting the value of $w$.

**Conclusion.** This paper introduces a concept of *General Bayesian Posterior* which allows to make credible predictions of posteriors without the model being well-specified.

## 3 Generalized Variational Inference: Three Arguments for Deriving New Posteriors

Knoblauch et al. 2022 propose an optimization-centric generalization of Bayesian inference, called the *Rule of Three* (RoT). They show that the solution found by a classical Bayes rule approach can be equivalently formulated as the solution of an infinite-dimensional optimization problem. Formally, the generalized Bayesian posterior is defined as

$$q^* \;=\; \arg\min_{q \in \mathcal{P}(\Theta)} \left\{ \mathbb{E}_{q(\theta)} \left[ \sum_{i=1}^{n} \ell(\theta, x_i) \right] + \mathrm{KL}(q \,\|\, \pi) \right\}. \tag{1}$$

This formulation, first emphasized by Bissiri et al. 2016 (the previous study considered by this Review), shows that Bayesian updating can be understood as an optimization problem that balances data fit, measured through a loss function, with regularization through closeness to the prior.

The classical Bayesian posterior arises as a special case of (1) when the loss function is chosen as the negative log-likelihood, $\ell(\theta, x_i) = -\log p(x_i \mid \theta)$. In this case, the posterior distribution $q_B^*$ is characterized as

$$q_B^* \;=\; \arg\min_{q \in \mathcal{P}(\Theta)} \left\{ \mathbb{E}_{q(\theta)} \left[ -\sum_{i=1}^{n} \log p(x_i \mid \theta) \right] + \mathrm{KL}(q \,\|\, \pi) \right\}. \tag{2}$$

Knoblauch et al. 2022 claim that: "any commitment to a Bayesian posterior is always a commitment to an optimization objective", meaning that they treat the optimization centric approach in (1) and (2) and its objective in the same way as the standard Bayesian posterior.

**Optimality of Variational Inference and Suboptimality of Alternative Methods.** Knoblauch et al. 2022 show that the approximation of a posterior by vanilla Variational Inference (VI) can be viewed as the solution to an optimization problem. Specifically, for a variational family $Q \subset \mathcal{P}(\Theta)$, the VI posterior is given by

$$q^*_{\mathrm{VI}} \;=\; \arg\min_{q \in Q} \left\{ \mathbb{E}_{q(\theta)} \Big[ \sum_{i=1}^{n} \ell(\theta, x_i) \Big] + \mathrm{KL}(q \,\|\, \pi) \right\}. \tag{3}$$

The authors prove that

> "Relative to the infinite-dimensional optimization problem over $\mathcal{P}(\Theta)$ characterizing Bayesian inference and a fixed finite-dimensional variational family $Q$, standard VI produces the optimal posterior belief in $Q$."

(Theorem 3 in Knoblauch et al. 2022).

This theorem has two main implications. First, standard VI and the optimization-centric form of classical Bayes (eqs. (1) and (2)) share the same objective. Once we accept the Bayesian posterior $q^*_B$ as desirable, we are also accepting that the objective in eq. (3) encodes the properties we want our posterior to satisfy. Restricting to $Q$, VI then computes the best possible solution within that variational family. Second, from this theoretical point of view, alternative approximation methods such as Expectation Propagation, $\alpha$-divergence VI, or Discrepancy VI are suboptimal relative to eqs. (1)–(3), since they have different objectives. However, in practice, some of these alternative methods can outperform VI. The authors explain that it can happen on one of the following cases:

(i) the original Bayesian objective is misspecified (the model or prior are wrong) and does not reflect the belief distribution we wish to compute, or (ii) the approximating family $Q$ is too restrictive to capture the true posterior structure.

Thus, some alternative *(theoretically sub-optimal)* methods can sometimes produce empirically better approximations.

**Motivation for the Rule of Three (RoT).** The empirical success of alternatives to standard VI methods underlines the necessity for a more flexible approach which allows to control, specify, and design the objective we want to optimize. As the pivotal point of their paper, the authors suggest a way to do it. They axiomatically build the RoT – a tool for Bayesian inference, which postulates that any posterior belief distribution should be depend on three arguments $P(\ell, D, \Pi)$: (1) a loss $\ell$, (2) a divergence $D$, and (3) a feasible set of probability measures $\Pi \subseteq \mathcal{P}(\Theta)$. The authors define the general RoT posterior to be

$$q^*_{RoT} = \arg\min_{q \in \Pi} \left\{ \mathbb{E}_{q(\theta)} \left[ \sum_{i=1}^{n} \ell(\theta, x_i) \right] + D(q \| \pi) \right\}. \tag{4}$$

The authors also show the modularity of RoT. Formally, hold $n$, $\pi(\theta)$ and $\Pi$ fixed and take $q^*_{RoT,1} \in \Pi$ as a posterior computed via $P(\ell, D, \Pi)$. If one wishes to derive an alternative posterior $q^*_{RoT,2} \in \Pi$ (in order to address some particular issues of the previous one) through the RoT, this can be done in three modular ways: 1) change $\ell$ to make the posterior more robust to model misspecification, 2) change $D$ to make the posterior robust to prior misspecification without changing the parameter of interest, 3) change $D$ to affect uncertainty quantification without changing the parameter of interest.

**Generalized Variational Inference (GVI).** A tractable special case of RoT is obtained when $\pi(\theta)$ is restricted to a variational family $Q$. The authors call this *Generalized Variational Inference* (GVI), with objective

$$q^*_{\mathrm{GVI}} = \arg\min_{q \in Q} \left\{ \mathbb{E}_{q(\theta)} \left[ \sum_{i=1}^{n} \ell(\theta, x_i) \right] + D(q \| \pi) \right\}. \tag{5}$$

Knoblauch et al. 2022 highlight the benefits of RoT and GVI: while having a strong axiomatical foundation, these methods suggest a structured way to design an appropriate objective which suits better to calculate a posterior in each particular case. The modularity explained above contributes to the design process.

Many existing methods, including Standard Bayes, Standard VI and others can be recovered as special cases of RoT or GVI. Some of the methods which are described above as *sub-optimal* alternatives to Standard VI cannot be recovered from RoT for the same reason as explained above – they do not share the same optimization objective as the classical Bayes, and thus, as RoT and GVI. .

**Empirical results.** Knoblauch et al. 2022 finally conduct experiments to compare GVI (using Rényi's $\alpha$-divergence) with Standard VI and Disrepancy VI (with divergence being $\alpha$-divergence or Rényi's $\alpha$-divergence). They choose to address $D$ in particular as a result of the modularity of RoT, because the suggested model for experiments — a Bayesian Neural Network — suffers from a badly specified prior. The results show that GVI's performance depends heavily on the choice of parameter $\alpha$, with some values of $\alpha$ showing consistently better results than Standard VI. Disrepancy VI, in comparison to VI, shows some improved performance, but not as consistent as GVI.

**Conclusion.** The authors conclude that GVI is a theoretically motivated tool that allows to design more robust posteriors in a structured way. Empirical findings confirm this.

**Potential research questions.** The authors mention this as one of the future research directions as well: how to choose hyperparameters occurring in the loss or uncertainty quantifier (divergence). Does the choice of these parameters depend on the type of problem we aim to solve? Another question worth exploring arises from Section 3 of Bissiri et al. 2016. They point out that for loss functions other than the negative log-likelihood, one needs to *calibrate* the relative weight of the loss and divergence terms, since they can have different scales. It may be an interesting direction to investigate how such calibration procedures could be systematically incorporated into the RoT and GVI methodology.

# 4 Prediction-Centric Uncertainty Quantification via MMD

In their paper, Shen et al. 2025 highlight the disadvantages of a standard Bayesian framework, especially when applied to deterministic models. As discussed above, model misspecification creates fundamental challenges for Bayesian inference. These challenges become particularly important in deterministic settings. The reason is that a deterministic model assumes no variance in outcomes: once the parameters are fixed, predictions are point values or trajectories without a random spread. By the Bernstein–von Mises theorem, as the sample size grows, the Bayesian posterior concentrates around a single (pseudo-true, in case of misspecified models) parameter value. In deterministic models, this concentration implies that the posterior distribution degenerates to a Dirac measure, vanishing uncertainty in parameters. This is especially bad for predictions in cases when the original model is misspecified, because they might loose variability and become incorrectly overconfident.

It is also worth noting that GVI methodology described above exhibits the same issue: producing wrong and overconfident predictions in case of misspecified deterministic models.

Shen et al. 2025 introduce a different approach to produce more robust predictions in cases of deterministic and misspecified models – *Prediction-Centric Uncertainty Quantification (PCUQ)*. Instead of relying on a wrong likelihood $p(x \mid \theta)$, which measures how well the data is explained by each parameter value, they propose to evaluate how well the *predictive distribution* explains the data. The authors define the mixture predictive distribution as $p_q(x) = \int p(x \mid \theta) q(\theta)\, d\theta$. Note that if $q(\theta) = \delta_\theta$, the original likelihood $p(x \mid \theta)$ is recovered. In this setting, $p_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$ denotes the empirical distribution of the observed data $x_{1:n}$. The fit between the predictive distribution $p_q$ and the empirical distribution $p_n$ is measured by the squared Maximum Mean Discrepancy $\mathrm{MMD}^2(p_n, p_q)$. Regularisation is achieved by the Kullback–Leibler divergence, weighted by a positive constant $\lambda$ controlling the trade-off between predictive fit and regularisation. With this notation in place, the PCUQ objective is defined as

$$q_n \;=\; \arg\min_{q \in \mathcal{P}(\Theta)} \Big\{ \, \mathrm{MMD}^2(p_n, p_q) \;+\; \lambda \, \mathrm{KL}(q \,\|\, \pi) \, \Big\}.$$

The outcome of PCUQ is the belief distribution $q_n$ over parameters. Unlike standard Bayes or generalised Bayes, which collapse to a single parameter value as $n \to \infty$, PCUQ sometimes keeps uncertainty. This ensures that even when the original model $p(x \mid \theta)$ is deterministic and misspecified, the predictive distribution, induced by the PCUQ posterior $q_n$, keeps a meaningful spread, avoiding the problem of overconfident but wrong predictions.

**Empirical results.** The experiments Shen et al. 2025 have conducted confirm their theoretical conclusion. With misspecified deterministic models, both standard Bayesian inference and Generalized Bayes produce bad results – the posterior concentrates around a single parameter configuration. Results produced by PCUQ avoid over confidence.

**Potential research questions.** Authors highlight this research direction themselves: systematic and theoretically grounded way might exist to better select the value of of $\lambda_n$ – the weight parameter of the regularization term. Moreover, we could explore the posteriors produced by the same objective with $MMD^2$ being replaced by other loss functions.

**Questions.** In the theoretical foundations of their methodShen et al. 2025 do not seem to be using the fact that all $p(x \mid \theta)$ are deterministic. They explain why the challenges PCUQ aims to solve are especially important in case of deterministic models. However, could PCUQ also be applied to non-deterministic (stochastic) models? Would this approach bring any benefit?

# 5 Predictive Variational Inference: Learn the Predictively Optimal Posterior Distribution

In their work, Lai and Yao 2024 question the advantages of standard Bayesian inference and variational inference (VI) in the presence of model misspecification. As discussed in previous sections, when the sample size is large enough, both VI and the Bayesian posterior concentrate around a single point estimate, vanishing the uncertainty about the parameters of interest. Lai and Yao 2024 claim that under model misspecification, predictive performance of both standard VI and standard Bayes are even less optimal than the ones produced by frequentist approach. To address this challenge, the authors develop a new framework for constructing more reliable and optimal posteriors: *Predictive Variational Inference (PVI).* In PVI, the goal is to find the variational parameter $\phi \in \Phi$ such that the variational distribution $q_\phi$ induces a predictive distribution that performs best under the chosen scoring rule. Here, they define the *posterior predictive distribution* of the next unseen data point as $p_{q_\phi}(\tilde{x}) = \int \; p(\tilde{x} \mid \theta)\, q_\phi(\theta)\, d\theta$. The authors evaluate how well the predictive distribution explains the data point $x_i$ using a proper scoring rule $S(p_{q_\phi}(\tilde{x}), x_i)$. Regularization is achieved by choosing a regularizer $r(\phi)$ and its positive weight $\lambda$. The objective specified by Lai and Yao 2024 is defined as:

$$\hat{\phi} = \arg\max_{\phi \in \Phi} \left[ \sum_{i=1}^{n} S\big(p_{q_\phi}(\tilde{x}), x_i\big) \; - \; \lambda\, r(\phi) \right]. \tag{6}$$

Lai and Yao 2024 also prove the following theoretical properties. If the model is well-specified, PVI behaves like regular Bayesian methods: the variational distribution $q_\phi(\theta)$ concentrates on the true parameter value as the data sample becomes large enough. However, if the model is misspecified in the sense that the true data-generating process corresponds to a distribution of some parameter rather than a single fixed value, then PVI does not collapse to a single point estimate. Instead, it converges to the true distribution of the parameter of interest. Formally, if there exists a parameter $\phi^*$ such that

$$p_{\text{true}}(y) \;=\; \int p(y \mid \theta)\, q_{\phi^*}(\theta)\, d\theta,$$

then the PVI solution will asymptotically converge to $\phi^*$, with the parameter variational posterior also converging to the true distribution, but not a single point mass. Thus, unlike standard Bayes or VI, which concentrate on a single pseudo-true parameter even under misspecification, PVI is able to recover the full distribution $q_{\phi^*}(\theta)$.

The part of this theoretical conclusion about the case with misspecified models has a very important implication: one can use it for *heterogeneity detection.* If the PVI posterior keeps uncertainty instead of collapsing to a single point estimate, the model one assumed is misspecified. In addition, it gives a clearer picture about which parameters within the population vary.

**Empirical results.** The results of experiments Lai and Yao 2024 have conducted agree with their theoretical findings. PVI shows better predictive performance compared to standard VI, keeping uncertainty about the parameters of interest. In some cases, where performance of two methods was relatively similar, the model was well specified.

**Potential research questions.** One of potential research directions could investigate the fine tuning of the objective – for example, trying out different scoring rules and values of parameter $\lambda$ (weight of the regularisation term in PVI objective). Another open question is – extending PVI framework to work with dependently distributed data as well (the current framework always assumed $x_{1:n}$ are IID). Research conducted by Lai and Yao 2024 has a few parallels with the previous work considered in this Review – Prediction-Centric Uncertainty Quantification via MMD (Shen et al. 2025). These two approaches could be compared on the same datasets. Moreover, the generalized framework for making reliable predictions could be developed. It is very noticeable that the objectives Lai and Yao 2024 and Shen et al. 2025 introduce are very similar. They both construct a predictive distribution and compare it to an empirical distribution. We could construct a general framework for predictive inference, similar to RoT and GVI described by Knoblauch et al. 2022, and test it with different loss or similarity functions, divergences, and hyperparameters.

# 6 Reproducible Parameter Inference Using Bagged Posterior

In their work Huggins and Miller 2024 address another challenge which appears in Bayesian inference under model misspecifications – lack of posterior reproducibility. It means that when the model is misspecified, the posterior distributions obtained from two independent data sets from the same data-generating process, can place nearly all of their mass on disjoint sets. The authors not only suggest a different framework, but also introduce a criterion for reproducible uncertainty quantification.

**Overlap Criterion for Reproducible Uncertainty Quantification.** To have a threshold for posterior reproducibility, Huggins and Miller 2024 introduce an *overlap criterion*. Here, $X$ and $Y$ are independent, both sampled from $p_{true}$, and $A_X$ and $B_Y$ are the posterior credible sets corresponding to $X$ and $Y$, which contain the functional of interest with probabilities greater than or equal to $1-\alpha$ and $1-\alpha'$, respectively. In this setting, to satisfy the overlap criterion, the probability of their intersection must satisfy

$$\mathbb{P}\big(A_X \cap B_Y\big) \;\geq\; (1-\alpha)(1-\alpha'). \tag{7}$$

Failure to meet this bound indicates the lack of reproducibility of posteriors, given different datasets from the same true distribution, which makes the inference results unreliable.

**Introducing BayesBag.** Huggins and Miller 2024 introduce a new posterior inference framework, which produces more reliable (with respect to the overlap criterion) results under model misspecifications. The idea is based on taking bootstrap samples of size $M$ from the original dataset of size $N$ with replacement, and then calculating the average of posteriors for each of the bootstrap copies. The authors refer to this technique as *BayesBag*. More specifically, they define *bagged posterior* as

$$q^*(\theta \mid x) = \frac{1}{N^M}\sum_{x^*}(q(\theta \mid x^*)), \tag{8}$$

where $x^* := (x_1^*, ..., x_M^*)$ is a bootstrap copy of the original dataset, $q(\theta \mid x^*)$ is a standard posterior given the data $x^*$, and $N^M$ refers to the number of all possible bootstrap datasets of $M$ samples, drawn from the original dataset with replacement. In practice, the authors recommend approximating the bagged posterior by generating $B$ independent bootstrap datasets $x^{*(1)}, \ldots, x^{*(B)}$, where each $x^{*(b)}$ consists of $M$ samples drawn with replacement from the original dataset $x$. The bagged posterior is then approximated as

$$q^*(\theta \mid x) \;\approx\; \frac{1}{B} \sum_{b=1}^{B} q(\theta \mid x^{*(b)}). \tag{9}$$

Since the bagged posterior is simply the average of standard Bayesian posteriors, any known algorithm for posterior inference can be applied to each bootstrap dataset, and the results subsequently averaged. Huggins and Miller 2024 point out that $B$ can be chosen appropriately based on the error of the Monte Carlo approximation described above. In their experiments, they show that setting $B = 50$ or $B = 100$ is a reasonable choice for many problems.

**Theoretical Justification of the Bagged Posterior.** The authors demonstrate that the bagged posterior can be motivated theoretically, which explains its good reproducibility properties under model misspecification. The authors point out that "for reproducibility, one needs to represent uncertainty across datasets from the true distribution." A natural framework for doing so is *Jeffrey conditionalization.* The approach of Jeffrey conditionalization to quantifying uncertainty about the parameter is to use the true data distribution $p_{true}(x)$. Formally,

$$q(\theta) = \int q(\theta \mid x_{1:N})\, p_{true}(x_{1:N})\, dx_{1:N}. \tag{10}$$

However, even when we are not given the true data distribution $p_{true}(x_{1:N})$, but instead observe $X_1, \ldots, X_N \sim p_{true}$, we can use the empirical distribution $P_N = \frac{1}{N} \sum_{n=1}^{N} \delta_{X_n}$ as a consistent estimator of $p_{true}$. The authors show that by replacing the true distribution with the empirical distribution in equation (10), we obtain exactly the bagged posterior

$q^*(\theta \mid x)$ from equation (8), when $M = N$. In this setting, the bagged posterior represents uncertainty in $\theta$ by integrating over datasets drawn from an approximation to the true distribution, while remaining close to the standard posterior $q(\theta \mid x)$. This allows bagged posterior to combine Bayesian model-based uncertainty with frequentist sampling variability. Thus, it is expected that bagged posterior will improve reproducibility across datasets.

**Empirical results.** On multiple examples Huggins and Miller 2024 show that under model misspecifications, bagged posterior typically satisfies the overlap criterion (equation (7)), while standard Bayesian posterior does not. In cases when the model is well specified, both Bayesian and bagged posteriors usually satisfy the criterion.

**Conclusion.** The authors conclude by pointing out the upsides and downsides of using BayesBag. The tool is very straightforward and general, applicable to parameter inference, making predictions, and model selection (the next section of this Literature Review will discuss this aspect in a more detail). In addition to that, the only parameters used are $B$ and $M$, which are straightforward to set. With showing the equivalence of bagged posterior for $M = N$ to Jeffrey conditionalization, Huggins and Miller 2024 show that $M = N$ is a natural choice. Another benefit of bagged posterior stated above, is that it incorporates frequentist and Bayesian uncertainties, without sacrificing the core benefits of using Bayes. On the negative side, using BayesBag is computationally expensive.

**Potential research questions.** Future work could extend BayesBag to use dependent data.

# 7 Reproducible Model Selection Using Bagged Posterior

In this paper Huggins and Miller 2023 address another serious challenge which arises in Bayesian statistics – model selection. All the previous studies were exploring parameter inference or making predictions, having an assumed model in place. However, the question of selecting the best available model is not less important. The authors state that the regular way of quantifying uncertainty in the choice of model is the posterior distribution over the available models. They define the posterior probability of the model $m$ as

$$q(m \mid X_{1:N}) \ \propto \ p(X_{1:N} \mid m)\, \pi(m),$$

where $\pi(m)$ is the prior belief over the model $m$, and $p(X_{1:N} \mid m)$ is the likelihood of seeing the observed data undet model $m$. However, so as in case of parameter inference, this approach has a very strong assumption, which is almost always violated – one of the candidate models is exactly correct. In cases when this assumption does not hold, the model selection procedure might yield unreliable and non-reproducible results. For example, the authors prove that when there are two models $m_1$ and $m_2$, which explain the data equally well, but both are misspecified, when the dataset size is sufficiently large, the posterior probability of each model will oscillate between values close to 0 and 1. It means that one dataset drawn from the true data-generating process may lead the posterior to collapse to $m_1$, while another dataset may lead it to collapse to $m_2$. This behavior illustrates the instability and lack of reproducibility in Bayesian model selection under misspecification.

**Problems of Bayesian model selection.** Huggins and Miller 2023 demonstrate that in cases when the true model is not one of the candidate models, Bayesian model selection can become unstable. They first consider a simplified example of two candidate models $\mathcal{M} = (m_1, m_2)$, where each contains exactly one parameter, and the observed data $x_1, \ldots, x_N$ drawn as realizations of i.i.d. random variables $X_1, \ldots, X_N$. To analyze the asymptotic behavior, they define the model log-likelihood ratio

$$Z_N := \log p(X_{1:N} \mid m_1) - \log p(X_{1:N} \mid m_2),$$

and log-likelihood ratio per-observation

$$Z_{Nn} := \log p_N(X_n \mid m_1) - \log p_N(X_n \mid m_2), \qquad n = 1, \dots, N,$$

where $p_N$ is the observation model, which depends on $N$. To be able to compare the mean and the standard deviation of $Z_N$ on the asymptotic scale, the authors assume that the following limits exist and are finite:

$$\mu_\infty := \lim_{N\to\infty} N^{1/2}\,\mathbb{E}[Z_{N1}], \qquad \sigma^2_\infty := \lim_{N\to\infty} \operatorname{Var}(Z_{N1}), \qquad \sigma^2_\infty \in (0, \infty). \tag{i}$$

The normalization of $\mu_\infty$ by $N^{1/2}$ ensures that the mean and the standard deviation of $Z_N$, which are $\mathbb{E}[Z_N] \approx N^{1/2}\,\mu_\infty$ and $\operatorname{Std}(Z_N) \approx N^{1/2}\,\sigma_\infty$, grow at the same rate, so that neither term dominates in the limit. Finally, they define the asymptotic effect size

$$\eta_\infty := \frac{\mu_\infty}{\sigma_\infty},$$

which quantifies the amount of evidence in favor of $m_1$ relative to $m_2$ under the true data distribution. If $\eta_\infty > 0$, $m_1$ is favored, and when $\eta_\infty < 0$, $m_2$ is favored. Theorem 3.1 in Huggins and Miller 2023 proves that the standard Bayesian model posterior converges to a Bernoulli distribution. Except (i), the authors make another assumption,

$$\limsup_{N\to\infty} \mathbb{E}\left[|Z_{N1}|^6\right] < \infty. \tag{ii}$$

The result of the theorem 3.1 states that, under assumptions (i) and (ii),

$$q(m_1 \mid X_{1:N}) \ \xrightarrow{D}\ \operatorname{Bernoulli}(\Phi(\eta_\infty))\,,$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal $\mathcal{N}(0,1)$. For example, if $\eta_\infty = 0$, which means that $m_1$ and $m_2$ are asymptotically equally misspecified, then both posteriors converge in distribution to Bernoulli$(1/2)$, meaning that with probability $1/2$ the posterior mass collapses to model $m_1$ and with probability $1/2$ it collapses to model $m_2$. This result illustrates that even when two models explain the data equally well, the Bayesian posterior does not remain balanced between them but instead oscillates randomly between 0 and 1 across different data samples, leading to instability and lack of reproducibility in model selection. In more realistic scenarios, models do not provide exactly the same fit, meaning that $\eta_\infty$ is not exactly 0. However, the lack of reliability and reproducibility is still present. For effect sizes $\eta_\infty > 1$, which should strongly favor model $m_1$, the Bayes posterior still assigns non-negligible probability to collapsing to model $m_2$. This being the motivation, Huggins and Miller 2023 propose an alternative approach to model selection.

**BayesBag for Model Selection.** The authors propose using the already defined in this review technique not only for parameter posterior inference, but also for model posterior. Huggins and Miller 2023 define the bagged posterior for model selection as

$$q^*(m \mid x) := \frac{1}{N^M} \sum_{x^*} q(m \mid x^*),$$

where $M$ is the size of each bootstrap dataset $x^*$. In Theorem 3.1, by making two additional assumptions, (iii) that the bootstrap sample size $M$ grows sufficiently fast so that $\lim_{N\to\infty} M/N^{1/2} = \infty$, and (iv) that the ratio $c := \lim_{N\to\infty} M/N \in [0, \infty)$ exists, the authors show that the bagged posterior of model $m_1$ converges to a continuous random variable on $[0, 1]$ whose distribution depends on both $\eta_\infty$ and $c$. In the special case $\eta_\infty = 0$ (corresponding to two models providing equally good fits), the bagged posterior converges to $\mathrm{Unif}(0, 1)$ when $M = N$ $(c = 1)$. This means that, instead of collapsing randomly to 0 or 1 as in the standard Bayesian posterior, the bagged posterior remains stable and spreads probability continuously across $[0, 1]$. Thus, the bagged posterior provides more reliable and reproducible results.

**Choice of $M$ and Recommended Workflow.** Compared to the study about applying BayesBag to parameter posterior inference, in the current setting the choice of $M$ – size of bootstrap datasets drawn from the original dataset with replacement, is not as straightforward. Huggins and Miller 2023 introduce a mismatch index $\mathcal{I}(f)$, which measures the amount of misspecification between a chosen *reference model* and the data. The reference model can either be selected as a well-specified model (if such a candidate exists), or it can be constructed as a disjoint set of models that share interpretable parameters. Let $v_N$ and $v_M^*$ denote, respectively, the Bayes and BayesBag posterior variances of the functional of interest $f(\theta)$ under the reference model. If the reference model is well-specified, then asymptotically the following relation holds:

$$Mv_M^* \ = \ 2Nv_N.$$

Based on this, the asymptotic version of the mismatch index is defined as

$$\mathcal{I}(f) \ = \ \begin{cases} 1 - \dfrac{2Nv_N}{Mv_M^*}, & \text{if } Mv_M^* > Nv_N, \\ \text{NA}, & \text{otherwise.} \end{cases}$$

The mismatch index $\mathcal{I}(f)$ measures how strong the variance captured by the BayesBag posterior deviates from that of the standard posterior under the reference model. When $\mathcal{I}(f) \approx 0$, which means that there is no misspecification, the two variances are in agreement. Positive values of mismatch index indicate that Bayes posterior is overconfident (and underconfident in case of negative values). The mismatch index allows to calibrate the value of $M$. Huggins and Miller 2023 suggest the following workflow for choosing the value of $M$: start from choosing $M = N$, and calculate the mismatch index, to understand the true variability between $v_N$ and $v_M^*$. In case the mismatch index is less than the mismatch cutoff (0.3 by default), they recommend recomputing bagged posterior with $M = N^{0.95}$ and oppositely, with $M = N^{0.75}$ if the misspecification seems to be very likely. The authors denote that these values are just rough guidelines, and even in their experiments try other powers of N as well, depending on the amount of misspecification.

**Empirical results and Conclusion.** Huggins and Miller 2023 illustrate the application of BayesBag for model selection on multiple examples. They come to the conclusion that bagged posterior seems to work well in cases when the significant misspecification is present, by reliably putting significant posterior mass on most optimal models. At the same time the authors denote that bagged posterior is more conservative, putting the posterior probabilities further away from extremes of 0 and 1. This conservativity might be not desired in cases when one of the models is correct, though.

**Potential research questions.** The authors denote that using BayesBag for model selection is computationally expensive. The further research might look into ways of optimizing the framework to become more computationally efficient.

**Questions.** After conducting a model selection procedure, what are our next steps? When we have a posterior over the models, do we choose one model which appears to be the optimal, or do we somehow average several models with relative weights?

# 8 On the Stability of General Bayesian Inference

In the current study, Jewson et al. 2024 consider the following problem: in modern statistical settings, it is often difficult to express one's beliefs about the likelihood model and the true data-generating process (DGP) with complete accuracy. For this reason, decision makers may want to use *canonical* models — well-known and computationally convenient families of distributions — in order to approximate their true beliefs about the likelihood and the DGP. When such a *functional model* is chosen for approximation, it certainly captures some of the main aspects of the decision maker's beliefs. However, it interpolates between those aspects in a convenient but arbitrary way, which does not necessarily reflect the decision maker's beliefs.

A trustful Bayesian analysis should not be overly sensitive to these arbitrary aspects of the functional model. Inference should be *stable* with respect to those parts of the chosen functional model that do not reflect the actual beliefs. In this study, the authors propose a framework for stable inference. Their idea is based on the general Bayesian perspective of inference (Bissiri et al. 2016), described in Section 2 of this review. Jewson et al. 2024 propose to measure stability with respect to the *posterior predictive distribution*, defined as

$$m_f^D(x_{\text{new}} \mid x) \;=\; \int f(x_{\text{new}};\theta)\, q^{\ell}(\theta \mid x)\, d\theta,$$

where $f(\cdot;\theta)$ denotes the functional likelihood model, $q^{\ell}(\theta \mid x)$ is the general Bayesian posterior with respect to a chosen loss function $\ell$, which corresponds to the divergence $D$, and $x$ denotes the observed data.

The goal of stable inference is to ensure that the posterior predictive distribution derived under an approximate functional model $f$ is close to the one obtained under the decision maker's true belief model $h$, that is

$$m_f^D(x_{\text{new}} \mid x) \;\approx\; m_h^D(x_{\text{new}} \mid x).$$

Jewson et al. 2024 illustrate that, in the case of standard Bayesian inference, stability of the posterior predictive distribution is practically very difficult to guarantee. Standard Bayesian inference corresponds to the choice of the loss function as the negative log-likelihood, which in turn corresponds to the Kullback–Leibler divergence (KLD). The authors refer to this setting as *KLD-Bayes.* Their main proposal is to use the $\beta$-divergence and the corresponding loss function instead, yielding a generalized Bayesian update that they refer to as *$\beta D$-Bayes.* They show that under $\beta D$-Bayes, posterior predictive stability can be achieved under much weaker and more interpretable for the user conditions than in the KLD-Bayes case.

**Stability to the specification of the Likelihood Function.** As their theoretical contributions, Jewson et al. 2024 derive and compare the necessary requirements on the models $f$ and $h$ to produce stable posterior predictive distributions in case of KLD-Bayes and $\beta$D-Bayes. They make an assumption about $h$ and $f$. Let $\Theta$ and $\mathcal{A}$ be the parameter spaces of $f$ and $h$, respectively, and $\theta \in \Theta$ and $\nu \in \mathcal{A}$ be the arbitrary parameters. The necessary condition (i) requires the mappings $I_f : \Theta \to \mathcal{A}$ and $I_h : \mathcal{A} \to \Theta$ to exist, such that

$$\text{if } D(p_{true} \,\|\, h(\cdot; I_f(\theta))) < D(p_{true} \,\|\, h(\cdot; \nu)) \,,$$

then $q^D(\nu \mid x)$ vanishes exponentially fast, and symmetrically, if

$$D(p_{true} \,\|\, f(\cdot; I_h(\nu))) < D(p_{true} \,\|\, f(\cdot; \theta)) \,,$$

then $q^D(\theta \mid x)$ vanishes exponentially fast. Jewson et al. 2024 prove that under the assumption stated above, for the KLD posterior predictive induced by $h$ and by $f$ to produce similar approximations to the DGP, it is necessary that for any parameter $\nu$,

$$\big| \log h(\cdot; \nu) - \log f(\cdot; I_h(\nu)) \big|$$

is small everywhere. The authors conclude that practically it is nearly impossible to comply with this. The nature of the log scale requires no mismatches in small probability values for the expression above to be small. It means that the user needs to choose $f$ and $h$ which agree on their tail probabilities, which correspond to outliers. This is not a reasonable requirement.

The solution proposed by Jewson et al. 2024 is to use the $\beta$-divergence, which imposes a much easier requirement for stability to satisfy. They suggest a metric by which the user is able to approximate their beliefs more easily. They define the *total variation divergence* (TVD) as

$$\text{TVD}(f(\cdot;\theta), h(\cdot;\nu)) := \sup_{X \in \mathcal{X}} \big| f(X;\theta) - h(X;\nu) \big| = \tfrac{1}{2} \int \big| f(x;\theta) - h(x;\nu) \big| \, dx.$$

They also define the *TVD neighborhood of likelihood models* as follows: likelihood models $f(\cdot;\theta)$ and $h(\cdot;\nu)$ are in the neighborhood $\mathcal{N}_{\text{TVD}}(\varepsilon)$ of size $\varepsilon$ if

$$\forall \theta \in \Theta, \, \exists \nu \in \mathcal{A} \;\text{ s.t. }\; \text{TVD}(f(\cdot;\theta), h(\cdot;\nu)) \leq \varepsilon \text{ and,}$$

$$\forall \nu \in \mathcal{A}, \, \exists \theta \in \Theta \;\text{ s.t. }\; \text{TVD}(f(\cdot;\theta), h(\cdot;\nu)) \leq \varepsilon.$$

Being in each others $\mathcal{N}_{\mathrm{TVD}}(\varepsilon)$ neighborhood, $f$ and $h$ are guaranteed to have a difference of at most $\epsilon$ between the probabilities of any events. Jewson et al. 2024 prove (with $\beta \in (1, 2]$ and assumption (i) in place) that as long as there exists an $\varepsilon$ such that $f$ and $h$ are in each other's TVD neighborhood, the resulting $\beta D$-Bayes posterior predictive distributions are stable. Moreover, not only do $f$ and $h$ yield stable posterior predictives with respect to each other, but they also yield stable approximations to the true DGP. This is a much easier and interpretable requirement to comply with for a user.

**Stability to the Data Generating Process.** In their study Jewson et al. 2024 answer not only the question of how to choose a good functional approximation to the true belief about the likelihood, but also how to choose an approximation to the user's belief about the data-generating process, for the posterior predictive to be stable and do not diverge for the belief distribution and its approximation. Let $g_1(x)$ denote the idealized user's belief about the DGP, and $g_2(x)$ a convenient approximation to it. The authors prove that similarly to the results of the previous subsection, in KLD-Bayes case, $g_1$ and $g_2$ must be close in their tails in order to produce stable posterior predictive distributions. As discussed above, in practice is it almost impossible to achieve. Jewson et al. 2024 define the TVD neighborhood of data-generating processes as follows: $g_1$ and $g_2$ are in the DGP neighborhood of size $\epsilon$, if $\mathrm{TVD}(g_1, g_2) \leq \epsilon$. The theoretical results, similarly to the previous subsection, show that in case of $\beta D$-Bayes, any $g_1$ and $g_2$ close in TVD sense, will produce stable results – posterior predictives close to one another, and to the true DGP.

**Choice of $\beta$.** The authors denote that the value of $\beta$ can be figured out individually in each setting in multiple ways. The general rule is that setting $\beta$ closer to 1, means being less robust to outliers, but more computationally efficient, with $\beta = 1$ $\beta$D-Bayes being the same as KLD-Bayes. Thus, the less confident the user is about their approximation to the beliefs about the likelihood or about the DGP, the larger value of $\beta$ they need to set.

**Empirical results.** The empirical results prove the theoretical findings – $\beta$D-Bayes inference is stable across likelihoods and DGPs which are TVD close.

**Potential research questions.** Jewson et al. 2024 emphasize that the $\beta$-divergence is not the only divergence that provides robustness to outliers. Future research could explore how different divergences affect stability, rather than focusing only on the comparison between KLD and $\beta$-divergence. Moreover, this paper further illustrates that the RoT and GVI framework of Knoblauch et al. 2022 could be developed further into a comprehensive guide, providing recommendations on which loss functions or divergences are most appropriate for different inferential goals and settings.

# 9 Martingale Posterior Distributions

In this work, Fong et al. 2023 argue for the necessity of a different approach to inference from the traditional likelihood–prior–posterior framework. The authors claim and illustrate that uncertainty in the parameter of interest comes from missing observations. In case the entire population were given, one would know the true value of the parameter. Thus, they suggest a different framework, and instead of relying on the traditional Bayesian model, focus on recovering the missing observations, and then compute the value of the parameter. The authors define the predictive density, given the observations $X_{1:n} \sim p_{true}$, as $p(x_{n+1:\infty} \mid x_{1:n})$, where the observations $X_{n+1:\infty}$ are to be recovered. The parameter induced from the predictive density is defined as $\theta_\infty = \theta(X_{1:\infty})$, meaning that $\theta_\infty$ is the true value of the parameter, when all data is observed. Thus, $\theta_\infty$ is the quantity of interest in this paper, and its distribution is referred to as *martingale posterior*. The authors' motivations to using such a framework are the following: (i) predictive models are probabilistic statements about quantities observed, and exclude the need of eliciting subjective statements about parameters (such as defining the likelihood and the prior), which might have no real world interpretations, (ii) this framework acknowledges that uncertainty in $\theta$ comes from missing observations, (iii) and delineates a stronger connection between Bayesian and frequentist uncertainties. To illustrate the difference of their framework from the traditional one, the authors use the following diagram,

$$\begin{tikzcd}
p(x \mid \theta),\ \pi(\theta) \arrow[r, "\text{Bayes' rule}"] & q(\theta \mid x_{1:n}) \arrow[r, "\substack{\text{posterior predictive,} \\ \int p(x|\theta)\, q(\theta|x_{1:n})\, d\theta}"] & p(x \mid x_{1:n})
\end{tikzcd}$$

$$\begin{tikzcd}
q(\theta \mid x_{1:n}) & p(x \mid x_{1:n}) \arrow[l, "\substack{\text{Doob's theorem,} \\ X_{n+1:\infty} \sim p(\cdot|x_{1:n})}"'] & p(x) \arrow[l, "\text{predictive update}"']
\end{tikzcd}$$

which indicates that traditional Bayesian approach requires to start with the prior and the likelihood, compute the posterior using Bayes rule, and from there compute the predictive distribution $p(x \; |x_{1:n})$. Oppositely, the framework suggested by Fong et al. 2023 starts from the observed data, uses the predictive update (further specified in more details) to get the predictive density, and from there computes the parameter posterior, relying on results of *Doob's theorem.*

**Theoretical justification** Doob's theorem provides a theoretical foundation for the martingale posterior by showing that Bayesian uncertainty about the parameter of interest can be understood as uncertainty about the unobserved data.

Formally, suppose all the data is yet to be observed, and $X_{1:\infty}$ are i.i.d. from a sampling density $p_\Theta$, where $\Theta$ is drawn from the prior distribution $\Pi$. An appropriate Bayesian point estimate of $\Theta$ is the posterior mean, defined as

$$\bar{\theta}_N = \mathbb{E}[\Theta \mid X_{1:N}],$$

where $X_{1:N}$ are the first $N$ observations. The sequence $\{\bar{\theta}_N\}_{N=1}^{\infty}$ forms a *martingale*, meaning that

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

Thus, by Doob's martingale convergence theorem

$$\bar{\theta}_N \; \to \; \Theta \quad \text{almost surely.}$$

In other words, as more data are simulated, the posterior mean converges almost surely to the true value of the parameter $\Theta$ that generated the data. Since $\Theta$ itself was originally drawn from the prior distribution $\Pi$, this result demonstrates the equivalence of two ways of sampling $\Theta$ from the prior before seeing any data.

The first is to draw directly $\Theta \sim \Pi$. The second, which this paper advocates for, is to construct the sequence of unseen observations $X_1, X_2, \ldots$ from the sequence of predictive densities,

$$X_1 \sim p(\cdot), \quad X_2 \sim p(\cdot \mid x_1), \quad X_3 \sim p(\cdot \mid x_1, x_2), \ \ldots \ .$$

Given this random infinite sequence, the posterior mean computed on the entire dataset is

$$\bar{\theta}_\infty = \lim_{N \to \infty} \bar{\theta}_N,$$

and by Doob's theorem, it is equivalent to

$$\bar{\theta}_\infty = \Theta \quad \text{almost surely.}$$

Thus, the posterior mean computed on the infinite (simulated) dataset—whose randomness comes from the uncertainty in the unobserved data—has the same distribution as the prior $\Pi$. Thus, prior uncertainty in $\Theta$ is equivalent to predictive uncertainty in the unseen data $X_{1:\infty}$.

Fong et al. 2023 point out that the same interpretation holds *posteriori*—after observing $X_{1:n} = x_{1:n}$. Sampling $\Theta$ from the Bayesian posterior is equivalent to sampling the unobserved observations $X_{n+1:\infty}$ conditional on $x_{1:n}$ and then computing $\theta_\infty$ from the completed infinite dataset.

After establishing this equivalence and justifying the proposed framework, the authors introduce ways to 1) sample the missing data and 2) recover the quantity of interest $\theta_\infty$.

**Sampling the Missing Data** As already stated earlier, the standard way to compute predictive distribution of the next unseen data, given the observed data, is to rely on the Bayesian posterior, computer from prior and likelihood. As the authors of this paper base their approach on stepping aside from the traditional prior-likelihood-posterior framework, they propose to compute the distribution of unobserved data given the observed data, using the infinite sequence of 1-step ahead predictive densities $\{p(\cdot \mid x_{1:N})\}_{N\geq n}$. Formally, the predictive distribution then takes the form

$$p(x_{n+1:N} \mid x_{1:n}) = \prod_{i=n+1}^{N} p(x_i \mid x_{1:i-1}).$$

Fong et al. 2023 discuss practical ways of constructing these 1 step ahead predictive distributions, and establish the theoretical properties of such updates necessary to guarantee the existence of the limit $\theta_\infty$. The authors conclude that these conditions are satisfied when $X_{n+1}, X_{n+2}, ...$ are conditionally (on observed data) identically distributed, and therefore in this case the limit $\theta_\infty$ exists.

**Computing the Quantity of Interest** After the unobserved data was sampled as described above, the authors propose to construct $F_\infty(x)$ — the random limiting empirical distribution, which includes both observed and sampled unobserved data. Formally, it is defined as

$$F_\infty(x) = \lim_{N\to\infty} \frac{1}{N}\left(\sum_{i=1}^{n} \mathbf{1}\{x_i \leq x\} + \sum_{i=n+1}^{N} \mathbf{1}\{X_i \leq x\}\right).$$

Having this distribution in place, the authors refer to the generalized Bayesian framework described in Section 2 of this Review, and define $\theta_\infty$ as the minimizer of the expected loss with respect to $F_\infty$:

$$\theta_\infty = \arg\min_\theta \int l(\theta, x)\, dF_\infty(x).$$

Fong et al. 2023 also suggest an algorithm for computing a quantity of interest – a Monte Carlo approximation to the martingale posterior. Before providing the algorithm itself, the authors define the predictive probability distribution function as,

$$\mathbb{P}_i(x) := \mathbb{P}(X_{i+1} \leq x \mid x_{1:i}),$$

where the corresponding density functions are denoted as $p_i(x)$, and $i$ denotes the length of the conditioning sequence.

The algorithm can be summarized as follows:

1) Compute $\mathbb{P}_n$ from the observed data $x_{1:n}$.

2) Choose a large $N > n$.

3) For $i = n+1, \ldots, N$, sequentially sample $X_i \sim \mathbb{P}_{i-1}$ from the one-step ahead predictive distribution, and update the predictive $\mathbb{P}_i$ to incorporate the new simulated value $X_i$.

4) After simulating $\{X_{n+1}, \ldots, X_N\}$, construct the empirical distribution $F_N$ from the combined dataset $\{x_{1:n}, X_{n+1:N}\}$.

5) Compute the approximation of $\theta_\infty$ either as $\theta_N = \theta(F_N)$, or as $\theta(P_N)$ if $\theta$ is supposed to denote a continuous quantity.

6) Repeat steps 3–5 $B$ times to obtain $\{\theta_N^{(1)}, \ldots, \theta_N^{(B)}\}$, which provide a Monte Carlo approximation to the martingale posterior distribution of $\theta_\infty$.

**Empirical results** Fong et al. 2023 show that in practice a good update rule for 1 step ahead predictives to use is the bivariate copula method — a bivariate probability density function $c : [0,1]^2 \to \mathbb{R}$, which updates the predictive distribution of the next step as

$$p_{i+1}(x) = c_{i+1}\{\mathbb{P}_i(x), \mathbb{P}_i(x_{i+1})\}p_i(x),$$

where the choice of $c$ depends on the exact data structure and task. In most experimental cases the results of martingale posterior agree with those of standard Bayesian posterior, while sometimes the process of computation being more efficient for martingale posterior.

# References

Bissiri, Pier Giovanni, Chris C. Holmes, and Stephen G. Walker (2016). "A general framework for updating belief distributions". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.5, pp. 1103–1130. DOI: 10.1111/rssb.12158.

Fong, Edwin, Chris Holmes, and Stephen G. Walker (2023). “Martingale posterior distributions”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 85.5. Read before The Royal Statistical Society at a meeting organized by the Research Section on 12 December 2022., pp. 1357–1391. DOI: 10.1093/jrsssb/qkad005. URL: https://doi.org/10.1093/jrsssb/qkad005.

Huggins, Jonathan H. and Jeffrey W. Miller (2023). “Reproducible Model Selection Using Bagged Posteriors”. In: *Bayesian Analysis* 18.1, pp. 79–104. DOI: 10.1214/21-BA1301. URL: https://doi.org/10.1214/21-BA1301.

— (2024). “Reproducible parameter inference using bagged posteriors”. In: *Electronic Journal of Statistics* 18.1, pp. 1549–1585. DOI: 10.1214/24-EJS2237. URL: https://doi.org/10.1214/24-EJS2237.

Jewson, Jack, Jim Q. Smith, and Chris Holmes (2024). “On the Stability of General Bayesian Inference”. In: *Bayesian Analysis* TBA.TBA, pp. 1–31. DOI: 10.1214/24-BA1502. URL: https://doi.org/10.1214/24-BA1502.

Knoblauch, Jeremias, Jack Jewson, and Theodoros Damoulas (2022). “An Optimization-centric View on Bayes’ Rule: Reviewing and Generalizing Variational Inference”. In: *Journal of Machine Learning Research* 23. Ed. by Frank Wood, pp. 1–109. URL: https://jmlr.org/papers/volume23/19-1047/19-1047.pdf.

Lai, Jinlin and Yuling Yao (2024). “Predictive variational inference: Learn the predictively optimal posterior distribution”. In: arXiv: 2410.14843 [stat.ML]. URL: https://arxiv.org/abs/2410.14843.

Shen, Zhen et al. (2025). “Prediction-centric uncertainty quantification via MMD”. In: *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*. Vol. 258. Proceedings of Machine Learning Research. PMLR. URL: https://arxiv.org/abs/2410.11637.