Bayesian Analysis (2023) **18**, Number 1, pp. 79–104

Reproducible Model Selection Using Bagged Posteriors*

Jonathan H. Huggins† and Jeffrey W. Miller‡

Abstract. Bayesian model selection is premised on the assumption that the data are generated from one of the postulated models. However, in many applications, all of these models are incorrect (that is, there is misspecification). When the models are misspecified, two or more models can provide a nearly equally good fit to the data, in which case Bayesian model selection can be highly unstable, potentially leading to self-contradictory findings. To remedy this instability, we propose to use bagging on the posterior distribution ("BayesBag") – that is, to average the posterior model probabilities over many bootstrapped datasets. We provide theoretical results characterizing the asymptotic behavior of the posterior and the bagged posterior in the (misspecified) model selection setting. We empirically assess the BayesBag approach on synthetic and real-world data in (i) feature selection for linear regression and (ii) phylogenetic tree reconstruction. Our theory and experiments show that, when all models are misspecified, BayesBag (a) provides greater reproducibility and (b) places posterior mass on optimal models more reliably, compared to the usual Bayesian posterior; on the other hand, under correct specification, BayesBag is slightly more conservative than the usual posterior, in the sense that BayesBag posterior probabilities tend to be slightly farther from the extremes of zero and one. Overall, our results demonstrate that BayesBag provides an easy-to-use and widely applicable approach that improves upon Bayesian model selection by making it more stable and reproducible.

Keywords: asymptotics, bagging, Bayesian model averaging, bootstrap, model misspecification, stability.

1 Introduction

In Bayesian statistics, the usual method of quantifying uncertainty in the choice of model is simply to use the posterior distribution over models. An implicit assumption of this approach is that one of the assumed models is exactly correct. But it is widely recognized that in practice, this assumption is typically unrealistic. When all of the models are incorrect (that is, they are *misspecified*), the posterior concentrates on the model that provides the best fit in terms of Kullback-Leibler divergence (Berk, 1966). However, when two or more models can explain the data almost equally well, the posterior becomes unstable and can yield contradictory results when seemingly inconsequential changes are

*J.H.H. was supported in part by the National Institute of General Medical Sciences of the National Institutes of Health under grant number R01GM144963 as part of the Joint NSF/NIGMS Mathematical Biology Program. J.W.M. was supported by the National Cancer Institute of the National Institutes of Health under grant number R01CA240299. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

†Department of Mathematics & Statistics, Boston University, huggins@bu.edu

‡Department of Biostatistics, Harvard University, jwmiller@hsph.harvard.edu

 https://doi.org/10.1214/21-BA1301

made to the models or to the data (Meng and Dunson, 2020; Oelrich et al., 2020; Yang and Zhu, 2018). For instance, as the size of the data set grows, the posterior probability of a given model may oscillate between values very close to 1 and very close to 0, *ad infinitum*. In short, Bayesian model selection can be unreliable and non-reproducible.

This article develops the theory and practice of *BayesBag*, a simple and widely applicable approach to stabilizing Bayesian model selection. Originally suggested by Waddell, Kishino and Ota (2002) and Douady et al. (2003) in the context of phylogenetic inference and then independently proposed by Bühlmann (2014) (who coined the name), the idea of BayesBag is to apply bagging (Breiman, 1996) to the Bayesian posterior. Let $Q(\mathfrak{m} \,|\, x) \propto p(x \,|\, \mathfrak{m}) Q_0(\mathfrak{m})$ denote the posterior probability of model $\mathfrak{m} \in \mathfrak{M}$ given data x, where $\mathfrak{M}$ is a finite or countably infinite set of models, $p(x \,|\, \mathfrak{m})$ is the marginal likelihood, and $Q_0(\mathfrak{m})$ is the prior probability. We define the *bagged posterior* $Q^*(\mathfrak{m} \,|\, x)$ by taking bootstrapped copies $x^* := (x_1^*, \ldots, x_M^*)$ of the original dataset $x := (x_1, \ldots, x_N)$ and averaging over the posteriors obtained by treating each bootstrap dataset as the observed data – that is,

$$Q^*(\mathfrak{m} \,|\, x) := \frac{1}{N^M} \sum_{x^*} Q(\mathfrak{m} \,|\, x^*), \tag{1}$$

where the sum is over all possible N^M bootstrap datasets of M samples drawn with replacement from the original dataset. The BayesBag approach is to use $Q^*(\mathfrak{m} \,|\, x)$ to quantify uncertainty in the model $\mathfrak{m}$. In practice, we can approximate $Q^*(\mathfrak{m} \,|\, x)$ by generating B bootstrap datasets $x^*_{(1)}, \ldots, x^*_{(B)}$, where each $x^*_{(b)}$ consists of M samples drawn with replacement from x, yielding the approximation

$$Q^*(\mathfrak{m} \,|\, x) \approx \frac{1}{B} \sum_{b=1}^{B} Q(\mathfrak{m} \,|\, x^*_{(b)}). \tag{2}$$

Hence, BayesBag is easy to use since the bagged posterior model probability is simply an average over Bayesian model probabilities. No additional algorithmic tools are needed beyond what a data analyst would normally use for posterior inference. Implementing BayesBag via (2) does require more computation since one must approximate B posteriors (one for each bootstrap dataset), where typically $B \approx 100$. However, this drawback is minimized by the fact that each posterior can be approximated in parallel, which is ideal for modern cluster-based high-performance computing environments.

Despite its attractive features, there has been limited methodological or theoretical work on BayesBag prior to the present paper. Bühlmann (2014) and Huggins and Miller (2019) consider BayesBag in the parameter inference and prediction setting. In this paper, we focus on the use of BayesBag for model selection, which has been explored empirically in an application to phylogenetic tree reconstruction (Douady et al., 2003; Waddell, Kishino and Ota, 2002). Building off this previous work, our primary contributions are:

1. We develop a rigorous asymptotic theory showing that, when all models are misspecified and two or more models have similar predictive accuracy, Bayesian model

selection is unstable, while BayesBag model selection remains stable. Our analysis quantifies the effects of the relevant factors such as the mean and variance of the log-likelihood ratios and the correlation structure of the log-likelihoods.

2. We provide concrete guidance on selecting the bootstrap dataset size M and, via our theory, we clarify the effect of M on the stability of BayesBag model selection.

3. We verify through numerical experiments on synthetic and real data that, when all of the models are misspecified, BayesBag model selection leads to more stable inferences across datasets and small model changes, while Bayesian model selection is unstable. When one of the models is correctly specified, BayesBag is slightly more conservative than Bayesian model selection, in the sense that the bagged posterior probabilities tend to be slightly farther from zero and one.

In short, we find that in the presence of misspecification, model selection with the bagged posterior has appealing statistical properties while also being easy to use and computationally tractable on practical problems.

The paper is organized as follows. Section 2 provides an overview of our theory, methodology, and experiments, and how they relate to previous work. In Section 3, we present our theoretical results, illustrate the theory graphically, discuss the use of BayesBag for model criticism, and outline our recommended workflow. Section 4 contains a simulation study using BayesBag for feature selection in linear regression. In Section 5, we evaluate BayesBag on real-world data in applications involving (i) feature selection for linear regression and (ii) phylogenetic tree reconstruction. We conclude in Section 6 with a discussion of current limitations and future directions. All data and code for the results in this paper are available at https://github.com/TARPS-group/bayesbag-model-selection-code.

2 Summary of results

2.1 Theory

It has long been known that when the best fit to the data distribution is attained by more than one model, the posterior typically does not converge on a single model (Berk, 1966). In Theorems 3.1 and 3.2, we characterize the asymptotic distribution of the posterior on models in this setting, for both the usual posterior ("Bayes") and the bagged posterior ("BayesBag"). More generally, our theory covers the case of multiple misspecified models with approximately equally good fit.

Suppose the observed data $x_1, \ldots, x_N$ are realizations of independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_N \in \mathbb{X}$, and denote $X_{1:N} = (X_1, \ldots, X_N)$. First, consider the special case of two distinct models, $\mathfrak{M} = \{\mathfrak{m}_1, \mathfrak{m}_2\}$. Assume these models are asymptotically equally misspecified in the sense that

$$\lim_{N\to\infty} N^{-1/2}\mathbb{E}\{\log p(X_{1:N} \,|\, \mathfrak{m}_1) - \log p(X_{1:N} \,|\, \mathfrak{m}_2)\} = 0.$$

Then under mild conditions, Theorem 3.1 (part 1) shows that the Bayes posterior mass on model $\mathfrak{m}_1$ converges in distribution to a $\mathsf{Bern}(1/2)$ random variable:

$$Q(\mathfrak{m}_1 \,|\, X_{1:N}) \xrightarrow[N\to\infty]{\mathcal{D}} \mathsf{Bern}(1/2). \tag{3}$$

In other words, when N is large, with probability $1/2$ model $\mathfrak{m}_1$ has posterior probability ≈ 1 and otherwise it has posterior probability ≈ 0. Since, asymptotically, both models provide equally good fit to the true data-generating distribution, one might hope that $Q(\mathfrak{m}_1 \,|\, X_{1:N}) \to 1/2$. However, (3) describes the opposite behavior: a single arbitrary model has posterior probability 1.

We show that BayesBag model selection does not suffer from this pathological behavior (Theorem 3.1, part 2). In the special case above (two models with asymptotically equally good fit), when $M = N$, the bagged posterior probability of model $\mathfrak{m}_1$ converges in distribution to a uniform random variable on the interval from 0 to 1:

$$Q^*(\mathfrak{m}_1 \,|\, X_{1:N}) \xrightarrow[N\to\infty]{\mathcal{D}} \mathsf{Unif}(0,1).$$

Alternatively, if we choose M such that $M/N \to 0$ and $M/N^{1/2} \to \infty$, then the bagged posterior mass on model $\mathfrak{m}_1$ has the appealing behavior of converging to $1/2$:

$$Q^*(\mathfrak{m}_1 \,|\, X_{1:N}) \xrightarrow[N\to\infty]{P} 1/2.$$

This is not simply due to the bagged posterior reverting to the prior; this result holds for any prior giving positive mass to both models. Theorem 3.2 extends Theorem 3.1 to the case of more than two models, in which case the asymptotic distribution depends on the covariance structure of the log marginal likelihoods of the models. Corollary 3.3 extends Theorem 3.1 to the case of models with non-trivial parameter spaces.

In practice, it is unlikely that two models would fit the true data-generating distribution *exactly* equally well. However, even if, say, model $\mathfrak{m}_1$ has posterior probability tending to 1 asymptotically, for a finite sample size it may be that $N^{-1/2}\mathbb{E}\{\log p(X_{1:N} \,|\, \mathfrak{m}_1) - \log p(X_{1:N} \,|\, \mathfrak{m}_2)\} \approx 0$, such that with probability $\approx 1/2$, model $\mathfrak{m}_2$ has posterior probability near 1. Indeed, the analysis of Yang and Zhu (2018) was motivated by observations of this phenomenon in Bayesian phylogenetic tree reconstruction (Alfaro, Zoller and Lutzoni, 2003; Douady et al., 2003; Wilcox et al., 2002), though it occurs more generally (Meng and Dunson, 2020), such as in neuroscience and economic modeling (Oelrich et al., 2020).

To understand this kind of finite-sample behavior via an asymptotic analysis, Theorems 3.1 and 3.2 are formulated for sequences of models for $N = 1, 2, \ldots$ that are not exactly equally good, but are asymptotically comparable in the sense that the expected log-likelihood ratios between models are $O(N^{1/2})$. In this way, our results provide insight into cases where the models are not dependent on N but the sample size is not yet large enough for the posterior to concentrate at the best fitting model(s).

2.2 Methodology

BayesBag requires the choice of a bootstrap dataset size M and the number of bootstrap datasets B. The choice of B controls the accuracy of the Monte Carlo approximation to the bagged posterior; see (1) and (2). It is straightforward to empirically estimate the error using the standard formula for the variance of a Monte Carlo approximation (Huggins and Miller, 2019). If $M = N$, we have found $B = 100$ to be sufficient in all of the applications we have considered. For $M < N$, the following result provides a natural lower bound on B to ensure all available data are used with high probability.

Proposition 2.1. *For $N > 1$, if $B \geq (N - 1/2)\log(N/\delta)/M$ then the probability that all observations are included in at least one bootstrap sample is greater than $1 - \delta$.*

While Proposition 2.1 offers a minimum value for B, we still recommend checking that the Monte Carlo standard error is sufficiently small for the application at hand.

For the choice of M, our theoretical and empirical results indicate that $M = N^{0.95}$ is a good default choice that will behave fairly well for model selection, both in cases where one model is correctly specified and, at the opposite extreme, when multiple misspecified models explain the data-generating distribution equally well. If significant misspecification is likely and there is a sufficient amount of data, a more aggressive choice such as $M = N^{0.75}$ could be appropriate. A recommended workflow is in Section 3.3.

2.3 Experiments

We validate our theory and proposed methods through simulations on feature selection for linear regression, and we evaluate the performance of BayesBag on real-data applications involving feature selection and phylogenetic tree reconstruction. Overall, our empirical results demonstrate that in the presence of significant misspecification, the bagged posterior produces more stable inferences and puts significant mass on optimal models more reliably than the usual Bayes posterior; on the other hand, when one of the models is correctly specified, the bagged posterior with $N^{0.95} \leq M \leq N$ is slightly more conservative than the posterior. Thus, BayesBag leads to more stable model selection results that are robust to minor changes in the model or representation of the data.

2.4 Related work

First, we discuss previous work in the parameter inference and prediction setting, with a model smoothly parameterized by $\theta \in \Theta \subseteq \mathbb{R}^D$. In a short discussion paper, Bühlmann (2014) introduced the name "BayesBag" to refer to bagging the posterior, and he presented a few simulation results in a simple Gaussian location model. However, Bühlmann (2014) employed a parametric bootstrap, which does not provide much benefit in a misspecified setting. In contrast, in recent work (Huggins and Miller, 2019), we found that using the nonparametric bootstrap to implement BayesBag yielded significant benefits for parameter inference and prediction under misspecification. In that work, we developed asymptotic theory for uncertainty quantification of the Kullback-Leibler optimal

parameter, providing insight into how to choose the bootstrap dataset size ($M = 2N$ if the model is correctly specified, and $M = N$ if the model is misspecified). Neither paper considers model selection, which raises fundamentally different issues because it involves a discrete space where smoothness does not play a role. Notably, our recommendation in this paper to take $M = o(N)$ for model selection is very different from our recommendations for parameter inference and prediction.

The previous work most closely related to the present work is a mix of empirical investigation (Douady et al., 2003; Oelrich et al., 2020; Waddell, Kishino and Ota, 2002) and theoretical work (Bühlmann and Yu, 2002; Oelrich et al., 2020; Yang and Zhu, 2018). The purely empirical papers undertake limited investigations in the setting of phylogenetic tree inference: Waddell, Kishino and Ota (2002) focus primarily on speeding up model selection and Douady et al. (2003) mainly aim to compare Bayesian inference to the bootstrap. Our Theorem 3.1 is similar in spirit to the bagging result of Bühlmann and Yu (2002, Proposition 2.1). However, the Bühlmann and Yu (2002) result is not applicable in the model selection setting since it would require assigning probability 1 to whichever model has the larger marginal likelihood—which does not correspond to Bayesian model selection—and then applying bagging to this selection procedure. Our other results (Theorem 3.2 and Corollary 3.3) go well beyond the scope of the Bühlmann and Yu (2002) result, covering three or more models as well as non-trivial parameter spaces.

Regarding the behavior of Bayesian model selection under the usual posterior, Yang and Zhu (2018) prove a result similar to (3) but more limited than our general versions in part 1 of Theorems 3.1 and 3.2. Finally, Oelrich et al. (2020) provide complementary results to our own: they study additional real-world examples of overconfident model selection and, in the feature selection setting, analyze the mean and variance of the log marginal likelihood ratio for a particular type of linear regression model with known variance, offering a more precise characterization of the posterior in that particular setting. However, they do not analyze or consider using the bagged posterior.

3 Theory and methodology

In this section, we present our theoretical results, illustrate the theory with plots comparing the asymptotics of BayesBag versus Bayes (Section 3.1), discuss the use of BayesBag for model criticism (Section 3.2), and provide a recommended workflow (Section 3.3).

3.1 Asymptotic analysis

In Bayesian model selection, we have a countable set of models $\mathfrak{M}$. Assume that model $\mathfrak{m} \in \mathfrak{M}$ has prior probability $Q_0(\mathfrak{m}) > 0$ and marginal likelihood

$$p(X_{1:N} \,|\, \mathfrak{m}) = \int \left\{ \prod_{n=1}^{N} p_{\theta_\mathfrak{m}}(X_n \,|\, \mathfrak{m}) \right\} \Pi_0(\mathrm{d}\theta_\mathfrak{m} \,|\, \mathfrak{m}),$$

where $\theta_\mathfrak{m} \in \Theta_\mathfrak{m}$ is an element of a model-specific parameter space with prior distribution $\Pi_0(\mathrm{d}\theta_\mathfrak{m} \,|\, \mathfrak{m})$. Assume $X_1, X_2, \ldots$ are i.i.d. from some unknown distribution $P_\circ$. Further,

for each $\mathfrak{m} \in \mathfrak{M}$, assume there is a unique parameter

$$\theta_{\mathfrak{m}\circ} := \operatorname*{arg\,min}_{\theta_{\mathfrak{m}} \in \Theta_{\mathfrak{m}}} -\mathbb{E}\{\log p_{\theta_{\mathfrak{m}}}(X_1 \,|\, \mathfrak{m})\}$$

that minimizes the Kullback-Leibler divergence from $P_\circ$ to the model. We say that *model* $\mathfrak{m}$ *is misspecified* if $P_{\theta_{\mathfrak{m}\circ}} \neq P_\circ$.

The posterior probability of $\mathfrak{m} \in \mathfrak{M}$ is $Q(\mathfrak{m} \,|\, X_{1:N}) \propto p(X_{1:N} \,|\, \mathfrak{m}) Q_0(\mathfrak{m})$. Let $X^*_{1:M}$ denote a bootstrapped copy of $X_{1:N}$ with M observations; that is, each observation X_n is replicated K_n times in $X^*_{1:M}$, where $(K_1, \ldots, K_N) \sim \mathsf{Multi}(M, 1/N)$ is a multinomial-distributed count vector. The bagged posterior probability of model $\mathfrak{m} \in \mathfrak{M}$ is then

$$Q^*(\mathfrak{m} \,|\, X_{1:N}) := \mathbb{E}\{Q(\mathfrak{m} \,|\, X^*_{1:M}) \,|\, X_{1:N}\},$$

which is equivalent to the informal definition in (1).

Two models with degenerate parameter spaces We first state our asymptotic theory in the case of two misspecified models, $\mathfrak{M} = \{\mathfrak{m}_1, \mathfrak{m}_2\}$, since the results are more intuitive in this special case. For the moment, we also assume that each model contains a single parameter value (that is, $|\Theta_{\mathfrak{m}}| = 1$). On the other hand, we allow the observation model $p_N(X_n \,|\, \mathfrak{m})$ to depend on the number of observations N, so that $p(X_{1:N} \,|\, \mathfrak{m}) = \prod_{n=1}^N p_N(X_n \,|\, \mathfrak{m})$. Let $Z_N := \log p(X_{1:N} \,|\, \mathfrak{m}_1) - \log p(X_{1:N} \,|\, \mathfrak{m}_2)$ denote the model log-likelihood ratio and, for $n = 1, \ldots, N$, let $Z_{Nn} := \log p_N(X_n \,|\, \mathfrak{m}_1) - \log p_N(X_n \,|\, \mathfrak{m}_2)$ denote the log-likelihood ratio for each observation.

To perform an asymptotic analysis that captures the behavior of the nonasymptotic regime in which the mean of Z_N is comparable to its standard deviation, we assume that $\mu_\infty := \lim_{N\to\infty} N^{1/2}\mathbb{E}(Z_{N1})$ and $\sigma^2_\infty := \lim_{N\to\infty} \mathrm{Var}(Z_{N1})$ exist. Thus, when N is large, $\mathbb{E}(Z_N) \approx N^{1/2}\mu_\infty$ and $\mathrm{Std}(Z_N) \approx N^{1/2}\sigma_\infty$. Consequently, $\mathbb{E}(Z_N)$ does not overwhelm $\mathrm{Std}(Z_N)$, even in the asymptotic regime. The asymptotic effect size $\eta_\infty := \mu_\infty/\sigma_\infty$ quantifies the amount of evidence in favor of $\mathfrak{m}_1$ under the true distribution $P_\circ$. If $\eta_\infty > 0$, then $\mathfrak{m}_1$ is favored, whereas $\mathfrak{m}_2$ is favored if $\eta_\infty < 0$.

Our first result shows that (1) the posterior probability of model $\mathfrak{m}_1$ converges to a Bernoulli random variable with parameter depending on η_∞, and (2) the bagged posterior probability of model $\mathfrak{m}_1$ converges to a continuous random variable on $[0, 1]$ with a distribution that depends on η_∞ and $\lim_{N\to\infty} M/N$. For $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, let $\mathcal{N}(\mu, \sigma^2)$ denote the normal distribution with mean μ and variance σ^2, and let $\Phi(t)$ denote the cumulative distribution function of the standard normal distribution $\mathcal{N}(0, 1)$.

Theorem 3.1. *Let* $X_1, X_2, \ldots$ *i.i.d.* $\sim P_\circ$ *for some distribution* $P_\circ$ *and assume*

(i) $\mu_\infty := \lim_{N\to\infty} N^{1/2}\mathbb{E}(Z_{N1}) \in \mathbb{R}$ *exists,*

(ii) $\sigma^2_\infty := \lim_{N\to\infty} \mathrm{Var}(Z_{N1}) \in (0, \infty)$ *exists,*

(iii) $\limsup_{N\to\infty} \mathbb{E}(|Z_{N1}|^6) < \infty$*,*

(iv) $M = M(N)$ *satisfies* $\lim_{N\to\infty} M/N^{1/2} = \infty$*, and*

(v) $c := \lim_{N\to\infty} M/N \in [0, \infty)$*.*

Then

1. *for the usual posterior,* $Q(\mathfrak{m}_1 \,|\, X_{1:N}) \xrightarrow{\mathcal{D}} U \sim \mathsf{Bern}(\Phi(\eta_\infty))$*, where* $\eta_\infty = \mu_\infty/\sigma_\infty$*;*
2. *for the bagged posterior,* $Q^*(\mathfrak{m}_1 \,|\, X_{1:N}) \xrightarrow{\mathcal{D}} \Phi(c^{1/2}W^*)$*, where* $W^* \sim \mathcal{N}(\eta_\infty, 1)$*.*

In particular, for the usual posterior, if $\eta_\infty = 0$ *then* $Q(\mathfrak{m}_1 \,|\, X_{1:N}) \xrightarrow{\mathcal{D}} \mathsf{Bern}(1/2)$*. Meanwhile, for the bagged posterior, if* $\eta_\infty = 0$ *then* $Q^*(\mathfrak{m}_1 \,|\, X_{1:N}) \xrightarrow{\mathcal{D}} \mathsf{Unif}(0,1)$ *when* $c = 1$ *and* $Q^*(\mathfrak{m}_1 \,|\, X_{1:N}) \xrightarrow{P} 1/2$ *when* $c = 0$*.*

Theorem 3.1 will follow as an immediate corollary of Theorem 3.2 below. Note that in part 2, when $c > 0$, the cumulative distribution function of the random variable $U^* := \Phi(c^{1/2}W_*)$ is given by $u \mapsto \Phi(c^{-1/2}\Phi^{-1}(u) - \eta_\infty)$ for $u \in (0,1)$. Thus, by differentiating this function, we find that the density of U^* is, for $u \in (0,1)$,

$$f(u) = \Phi'\big(c^{-1/2}\Phi^{-1}(u) - \eta_\infty\big)c^{-1/2}/\Phi'(\Phi^{-1}(u)).$$

Figure 1 illustrates how Theorem 3.1 establishes the greater stability of BayesBag versus Bayes for model selection. Even for effect sizes $\eta_\infty > 1$, which should strongly favor model $\mathfrak{m}_1$, the Bayes posterior overwhelmingly favors model $\mathfrak{m}_2$ with non-negligible probability – that is, $\mathbb{P}\{Q(\mathfrak{m}_1 \,|\, X_{1:N}) \approx 0\} \not\approx 0$. On the other hand, the probability that the BayesBag posterior strongly favors model $\mathfrak{m}_2$ goes to zero more rapidly as η_∞ increases – that is, $\mathbb{P}\{Q^*(\mathfrak{m}_1 \,|\, X_{1:N}) \approx 0\} \to 0$ more rapidly as η_∞ grows. For example, when $\eta_\infty = 2$ and $c = 1$, $\mathbb{P}(U = 0) > 0.02$ whereas $\mathbb{P}(U^* < 0.1) < 7 \times 10^{-5}$. Thus, in this example, Bayes will overwhelmingly favor the "wrong" model in approximately 1 out of 50 experiments, whereas BayesBag will strongly favor the wrong model in only approximately 7 out of 100,000 experiments.

Extension to three or more models In the case of three or more models, the behavior of the posteriors is more complicated because there is dependence on both the correlation structure and the relative variances of the log-likelihood ratios between each pair of models. Consider the case of $K < \infty$ models and enumerate them from 1 to K, so that $\mathfrak{M} = \{\mathfrak{m}_1, \ldots, \mathfrak{m}_K\}$. For $k = 1, \ldots, K$, define the individual model log-likelihood terms $Y_{Nn,k} := \log p_N(X_n \,|\, \mathfrak{m}_k)$, and let $Y_{Nn} := (Y_{Nn,1}, \ldots, Y_{Nn,K})^\top \in \mathbb{R}^K$. For $t, \mu \in \mathbb{R}^{K-1}$ and $\Sigma \in \mathbb{R}^{(K-1)\times(K-1)}$ positive definite, let $\Phi_{\mu,\Sigma}(t)$ denote the cumulative distribution function of the $(K-1)$-dimensional normal distribution $\mathcal{N}(\mu, \Sigma)$.

Theorem 3.2. *Let* $X_1, X_2, \ldots$ *i.i.d.* $\sim P_\circ$ *for some distribution* $P_\circ$*. Defining* $\mu'_N := N^{1/2}\mathbb{E}(Y_{N1})$ *and* $\Sigma'_N := \mathrm{Cov}(Y_{N1})$*, assume*

(i) $\mu'_\infty := \lim_{N\to\infty} \mu'_N \in \mathbb{R}^K$*,*

(ii) $\Sigma'_\infty := \lim_{N\to\infty} \Sigma'_N$ *positive definite,*

(iii) $\limsup_{N\to\infty} \mathbb{E}(\|Y_{N1}\|_2^6) < \infty$*,*

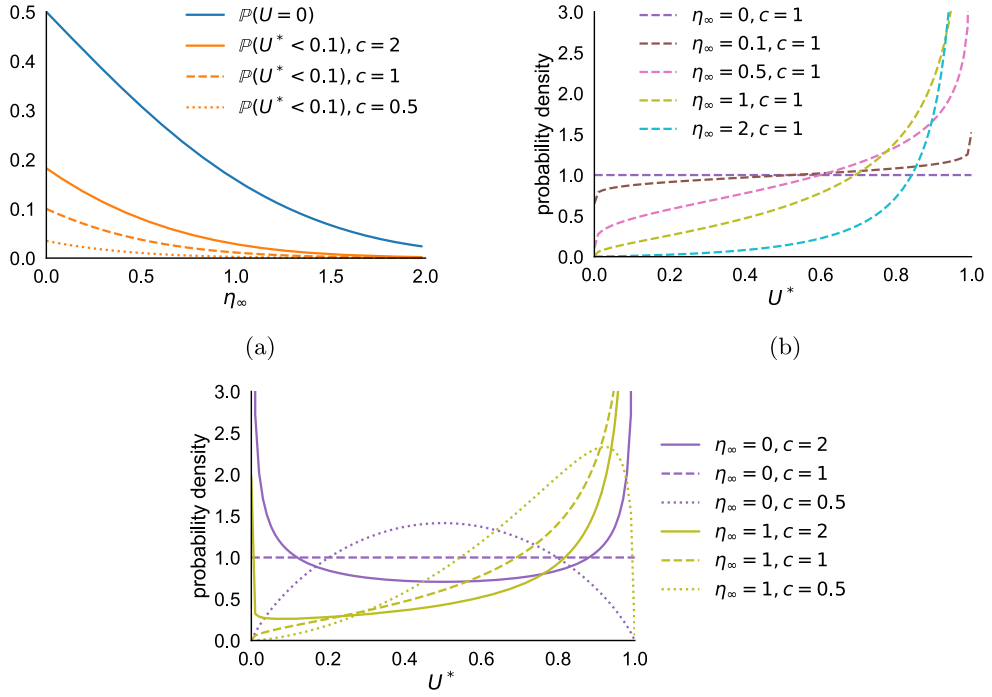


Figure 1: The bagged posterior (BayesBag) is far less likely than the usual posterior (Bayes) to strongly favor the wrong model (or an arbitrary equally good model). When there are two models, the asymptotic posterior probability of model $\mathfrak{m}_1$ is a random variable U (for Bayes) or U^* (for BayesBag), where $U \sim \mathsf{Bern}(\Phi(\eta_\infty))$, $U^* = \Phi(c^{1/2}W^*)$, $W^* \sim \mathcal{N}(\eta_\infty, 1)$, and η_∞ is the asymptotic effect size in favor of $\mathfrak{m}_1$ (see Theorem 3.1). **(a)** $U = 0$ represents the event that the Bayes posterior overwhelmingly favors the wrong model (or equally good model, if $\eta_\infty = 0$) – that is, the model with lower (or equal) expected log-likelihood under the true distribution. Likewise, $U^* < 0.1$ is the event that the BayesBag posterior strongly favors the wrong (or equivalent) model. **(b)** U^* is a continuous random variable on $[0, 1]$. The density of U^* is shown for a range of η_∞ values, with $c = \lim M/N$ fixed at $c = 1$, where N is the dataset size and M is the bootstrap dataset size (see Theorem 3.1). **(c)** Densities of U^* as both η_∞ and c vary.

(iv) $M = M(N)$ *satisfies* $\lim_{N\to\infty} M/N^{1/2} = \infty$, *and*

(v) $c := \lim_{N\to\infty} M/N \in [0, \infty)$.

Without loss of generality, consider the probability of $\mathfrak{m}_1$. *Define* $\mu_{\infty,k} := \mu'_{\infty,1} - \mu'_{\infty,k+1}$ *and* $\Sigma_{\infty,k,\ell} := \Sigma'_{\infty,1,1} + \Sigma'_{\infty,k+1,\ell+1} - \Sigma'_{\infty,1,k+1} - \Sigma'_{\infty,1,\ell+1}$ *for* $k, \ell \in \{1, \ldots, K-1\}$.

Then

1. *for the usual posterior,* $Q(\mathfrak{m}_1 \mid X_{1:N}) \xrightarrow{\mathcal{D}} U \sim \mathsf{Bern}(\Phi_{-\mu_\infty,\Sigma_\infty}(0))$*;*
2. *for the bagged posterior,* $Q^*(\mathfrak{m}_1 \mid X_{1:N}) \xrightarrow{\mathcal{D}} \Phi_{0,\Sigma_\infty}(c^{1/2}W^*)$*, where* $W^* \sim \mathcal{N}(\mu_\infty, \Sigma_\infty)$*.*

The proof is in Section S.3.2 of the Supplementary Material (Huggins and Miller, 2022). Figure 2 shows how Theorem 3.2 establishes that across a range of mean and covariance structures of the log-likelihoods, BayesBag is more stable than Bayes. Indeed, both methods behave fairly consistently as the covariance structure varies.

Extension to non-degenerate parameter spaces We now extend Theorem 3.1 to non-degenerate parameter spaces $\Theta_1 \subset \mathbb{R}^{D_1}$ and $\Theta_2 \subset \mathbb{R}^{D_2}$ and we integrate over $\theta_\mathfrak{m} \in \Theta_\mathfrak{m}$ for each model $\mathfrak{m}$. To avoid tedious arguments, we only consider the case where $\mu_\infty = 0$. For $\mathfrak{m} \in \{\mathfrak{m}_1, \mathfrak{m}_2\}$, let $\ell_{\mathfrak{m},\theta_\mathfrak{m}}(X_n) := \log p_{\theta_\mathfrak{m}}(X_n \mid \mathfrak{m})$ and recall that the optimal parameter is $\theta_{\mathfrak{m}\circ} = \arg\min_{\theta_\mathfrak{m} \in \Theta_\mathfrak{m}} -\mathbb{E}\{\ell_{\mathfrak{m},\theta_\mathfrak{m}}(X_1)\}$. Let $\Lambda_{X_{1:N}} := \log p(X_{1:N} \mid \mathfrak{m}_1)Q_0(\mathfrak{m}_1) - \log p(X_{1:N} \mid \mathfrak{m}_2)Q_0(\mathfrak{m}_2)$, where $p(X_{1:N} \mid \mathfrak{m}) = \int \{\prod_{n=1}^N p_{\theta_\mathfrak{m}}(X_n \mid \mathfrak{m})\}\Pi_0(\mathrm{d}\theta_\mathfrak{m} \mid \mathfrak{m})$ denotes the marginal likelihood. Let $X_{1:\infty}$ denote the infinite sequence $(X_1, X_2, \dots)$. We will assume that conditionally on $X_{1:\infty}$, for almost every $X_{1:\infty}$,

$$\Lambda_{X^*_{1:M}} = \frac{1}{2}(D_2 - D_1)\log N + \sum_{m=1}^{M} \log \frac{p_{\theta_{1\circ}}(X^*_m \mid \mathfrak{m}_1)}{p_{\theta_{2\circ}}(X^*_m \mid \mathfrak{m}_2)} + O_{P^+}(1), \tag{4}$$

where $X^*_{1:M}$ is bootstrapped from $X_{1:N}$ and $O_{P^+}(1)$ denotes a random quantity which is bounded in (outer) probability. It is well known that (4) holds with $X_{1:N}$ in place of $X^*_{1:M}$, under standard regularity assumptions (Clarke and Barron, 1990). Thus, we expect (4) to hold under similar but slightly stronger conditions, since we must consider a triangular array rather than a sequence of random variables.

The posterior distribution given $X_{1:N}$ and $\mathfrak{m}$ is

$$\Pi(\mathrm{d}\theta_\mathfrak{m} \mid X_{1:N}, \mathfrak{m}) := \frac{\prod_{n=1}^N p_{\theta_\mathfrak{m}}(X_n \mid \mathfrak{m})}{p(X_{1:N} \mid \mathfrak{m})}\Pi_0(\mathrm{d}\theta_\mathfrak{m} \mid \mathfrak{m}).$$

The *bagged posterior* $\Pi^*(\cdot \mid X_{1:N}, \mathfrak{m})$ given $X_{1:N}$ and $\mathfrak{m}$ is defined such that

$$\Pi^*(A \mid X_{1:N}, \mathfrak{m}) := \mathbb{E}\{\Pi(A \mid X^*_{1:M}, \mathfrak{m}) \mid X_{1:N}\}$$

for all measurable $A \subseteq \Theta$. Let $J_{\theta_\mathfrak{m}} := -\mathbb{E}\{\nabla^2_{\theta_\mathfrak{m}} \ell_{\mathfrak{m},\theta_\mathfrak{m}}(X_1)\}$ denote the Fisher information matrix. Finally, for a measure ν and function f, we use the shorthand notation $\nu(f) := \int f \mathrm{d}\nu$.

Corollary 3.3. *Let* $X_1, X_2, \dots$ *i.i.d.* $\sim P_\circ$ *and for* $\mathfrak{m} \in \{\mathfrak{m}_1, \mathfrak{m}_2\}$*, assume that*

(i) $\theta_\mathfrak{m} \mapsto \ell_{\theta_\mathfrak{m}}(X_1)$ *is differentiable at* $\theta_{\mathfrak{m}\circ}$ *in probability;*

(ii) there is an open neighborhood U *of* $\theta_{\mathfrak{m}\circ}$ *and a function* $m_{\theta_{\mathfrak{m}\circ}} : \mathbb{X} \to \mathbb{R}$ *such that* $P_\circ(m^3_{\theta_{\mathfrak{m}\circ}}) < \infty$ *and for all* $\theta_\mathfrak{m}, \theta'_\mathfrak{m} \in U$*,* $|\ell_{\theta_\mathfrak{m}} - \ell_{\theta'_\mathfrak{m}}| \le m_{\theta_{\mathfrak{m}\circ}} \|\theta_\mathfrak{m} - \theta'_\mathfrak{m}\|_2$ *a.s.* $[P_\circ]$*;*

(iii) $-P_\circ(\ell_{\theta_\mathfrak{m}} - \ell_{\theta_{\mathfrak{m}\circ}}) = \frac{1}{2}(\theta_\mathfrak{m} - \theta_{\mathfrak{m}\circ})^\top J_{\theta_{\mathfrak{m}\circ}}(\theta_\mathfrak{m} - \theta_{\mathfrak{m}\circ}) + o(\|\theta_\mathfrak{m} - \theta_{\mathfrak{m}\circ}\|_2^2)$ *as* $\theta_\mathfrak{m} \to \theta_{\mathfrak{m}\circ}$*;*

(iv) $J_{\theta_{\mathfrak{m}\circ}}$ *is an invertible matrix; and*

(v) letting $\vartheta^*_{\mathfrak{m}} \sim \Pi^*(\cdot \,|\, X_{1:N}, \mathfrak{m})$*, it holds that conditionally on* $X_{1:\infty}$*, for almost every* $X_{1:\infty}$*, for every sequence of constants* $C_N \to \infty$*,*

$$\mathbb{E}\Big\{\Pi(\|\vartheta^*_{\mathfrak{m}} - \theta_{\mathfrak{m}\circ}\|_2 > C_N/M^{1/2} \,|\, X^*_{1:M}, \mathfrak{m}) \;\Big|\; X_{1:N}\Big\} \to 0.$$

Further, assume that (4) *holds,* $\lim_{N\to\infty} M/N^{1/2} = \infty$*,* $c := \lim_{N\to\infty} M/N \in [0, \infty)$*,* $\mathbb{E}\{\ell_{\mathfrak{m}_1,\theta_{1\circ}}(X_1) - \ell_{\mathfrak{m}_2,\theta_{2\circ}}(X_1)\} = 0$*, and* $\mathbb{E}[\{\ell_{\mathfrak{m}_1,\theta_{1\circ}}(X_1) - \ell_{\mathfrak{m}_2,\theta_{2\circ}}(X_1)\}^3] \in (0, \infty)$*. Then the conclusions of Theorem 3.1 apply in the case of* $\eta_\infty = 0$*.*

The proof is in Section S.3.3 of the Supplementary Material.

Extension to dependent data A further extension, which we will not pursue in detail, is to non-independent data such as those encountered in time-series and spatial data analysis. In principle the generalization to, for example, time-series using the block bootstrap (or another nonparametric estimator such as a Gaussian process) is straightforward. However, the accompanying theory is much less straightforward since we must (A) determine the asymptotic distribution of rescaled log marginal likelihoods $N^{-\kappa} \log p(\mathfrak{m} \,|\, X_{1:N})$ and (B) show that a nonparametric estimator has the same asymptotic distribution. More concretely, consider the two-model scenario and define $W(X_{1:N}) := N^{-\kappa}\{\log p(\mathfrak{m}_1 \,|\, X_{1:N}) - \log p(\mathfrak{m}_2 \,|\, X_{1:N})\}$. Then we must determine an appropriate κ such that $W(X_{1:N}) \xrightarrow{\mathcal{D}} W_\infty$, where W_∞ is a non-degenerate distribution. Moreover, for (A) we must show that $\lim_{N\to\infty} d_{\mathcal{C}}(\mathcal{L}\{W(X_{1:N})\}, \mathcal{L}(W_\infty)) = 0$, where $\mathcal{L}(\xi)$ denotes the law of a random variable ξ and the metric $d_{\mathcal{C}}$ is defined in Section S.3.2 of the Supplementary Material. Then, for (B) we must show that for data $X^*_{1:M}$ distributed according to the nonparametric estimator,

$$d_{\mathcal{C}}(\mathcal{L}\{W(X^*_{1:M}) - (N/M)^\kappa W(X_{1:N}) \mid X_{1:N}\}, W_\infty - \mathbb{E}\{W_\infty\}) \xrightarrow{P} 0.$$

We leave a thorough investigation of models for dependent data to future work.

3.2 Model criticism with BayesBag

In the setting of parameter inference and prediction, Huggins and Miller (2019) develop a measure quantifying the amount of misspecification, referred to as the *model-data mismatch index*, based on comparing the BayesBag posterior to the Bayes posterior. To define a mismatch index in the setting of model selection, we perform parameter inference in a special designated model that we refer to as the *reference model*. Suppose $f(\theta_\circ)$ is a selected quantity of inferential interest, where $f : \Theta \to \mathbb{R}$ and Θ is the parameter space of the reference model. Let v_N and v^*_M denote, respectively, the Bayes and BayesBag posterior variances of $f(\theta)$ under the reference model. If the reference model is well-specified, then asymptotically, $Mv^*_M = 2Nv_N$ (Huggins and Miller, 2019). We define the asymptotic version of the mismatch index as

$$\mathcal{I}(f) := \begin{cases} 1 - 2Nv_N/(Mv^*_M) & \text{if } Mv^*_M > Nv_N, \\ \mathsf{NA} & \text{otherwise,} \end{cases}$$

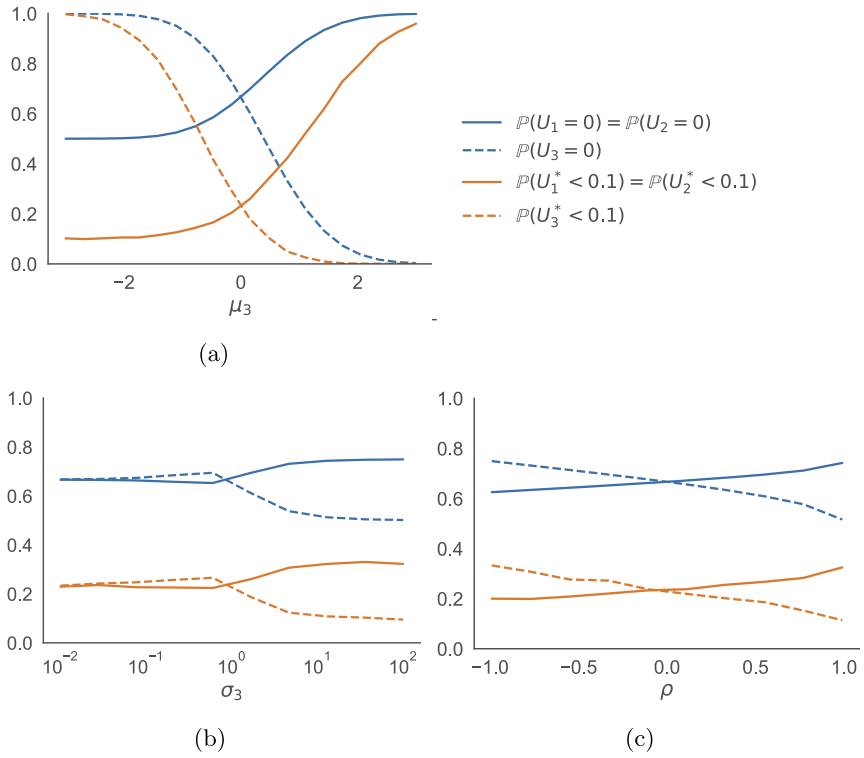


Figure 2: When there are more than two models, the bagged posterior (BayesBag) is far less likely than the usual posterior (Bayes) to strongly favor the wrong model (or an arbitrary equally good model) for a wide variety of mean and covariance structures of the asymptotic log-likelihoods. The asymptotic posterior probability of model $\mathfrak{m}_k$ is a random variable U_k (for Bayes) or U_k^* (for BayesBag), where $U_k \sim \mathsf{Bern}(\Phi_{-\mu_\infty,\Sigma_\infty}(0))$, $U_k^* = \Phi_{0,\Sigma_\infty}(c^{1/2}W^*)$, $W^* \sim \mathcal{N}(\mu_\infty, \Sigma_\infty)$, and $\mu_\infty \in \mathbb{R}^{K-1}$ and $\Sigma_\infty = \mathbb{R}^{(K-1)\times(K-1)}$ are, respectively, the asymptotic mean and covariance of the log-likelihood ratio of $\mathfrak{m}_k$ versus each other model (see Theorem 3.2). $U_k = 0$ represents the event that the Bayes posterior overwhelmingly rejects model $\mathfrak{m}_k$, and $U_k^* < 0.1$ is the event that the BayesBag posterior strongly rejects model $\mathfrak{m}_k$. Three scenarios are shown for the case of $K = 3$ models, for a range of values of $\mu'_\infty \in \mathbb{R}^3$ and $\Sigma'_\infty \in \mathbb{R}^{3\times3}$, the asymptotic mean and covariance of the log-likelihoods. **(a)** First, we vary μ_3, where $\mu'_\infty = (0, 0, \mu_3)^\top$ and the entries of Σ'_∞ are given by $\Sigma'_{\infty,i,j} = 0.5^{\mathbb{1}(i\neq j)}$. **(b)** Second, we vary σ_3, where $\mu'_\infty = (0, 0, 0)^\top$ and $\Sigma'_{\infty,i,j} = 0.5^{\mathbb{1}(i\neq j)}\sigma_3^{\mathbb{1}(i=3)}\sigma_3^{\mathbb{1}(j=3)}$. **(c)** Third, we vary ρ, where $\mu'_\infty = (0, 0, 0)^\top$ and $\Sigma'_{\infty,i,j} = \mathbb{1}(i = j) + \rho\mathbb{1}(i = 1, j = 2) + \rho\mathbb{1}(i = 2, j = 1)$.

where NA is short for "not available." The interpretation is as follows: $\mathcal{I}(f) \approx 0$ indicates no evidence of mismatch; $\mathcal{I}(f) > 0$ (respectively, $\mathcal{I}(f) < 0$) indicates the Bayes posterior is overconfident (respectively, under-confident); $\mathcal{I}(f) = \mathsf{NA}$ indicates that either there

is severe model-data mismatch or the required asymptotic assumptions do not hold (for example, due to multimodality in the posterior or small sample size). We refer the interested reader to Huggins and Miller (2019) for more justification and description of a non-asymptotic version of $\mathcal{I}$.

The reference model should be chosen such that if any model $\mathfrak{m} \in \mathfrak{M}$ is well-specified, then the reference model is well-specified. One common case is a finite set of models with partial order $\prec$ based on inclusion such that there exists a unique maximal model; in this case, the maximal model can be used as the reference model. More precisely, let $\mathcal{P}_{\mathfrak{m}} := \{p_{\theta_{\mathfrak{m}}}(\cdot \mid \mathfrak{m}) : \theta_{\mathfrak{m}} \in \Theta_{\mathfrak{m}}\}$. Then for models $\mathfrak{m}, \mathfrak{m}' \in \mathfrak{M}$, $\mathfrak{m} \prec \mathfrak{m}'$ if and only if $\mathcal{P}_{\mathfrak{m}} \subseteq \mathcal{P}_{\mathfrak{m}'}$, and $\mathfrak{m}$ is the unique maximal model if $\mathfrak{m}' \prec \mathfrak{m}$ for all $\mathfrak{m}' \in \mathfrak{M}$. Feature selection (Section 4) is an example of this type, where the maximal model includes all features. Another common situation is when all models have a set of shared, interpretable parameters, in which case we can define the reference model to be the disjoint union of all models $\mathfrak{m} \in \mathfrak{M}$. Phylogenetic tree reconstruction (Section 5) is an example of this type.

When there is more than one univariate quantity of inferential interest, we consider a collection of functions $f \in \mathcal{F}$ and suggest taking the most pessimistic mismatch value: $\mathcal{I}(\mathcal{F}) := \sup_{f \in \mathcal{F}} \mathcal{I}(f)$. In general, $\mathcal{F}$ can be chosen to reflect the quantities of interest in the application at hand. When $\theta \in \mathbb{R}^D$, two natural choices for the collection $\mathcal{F}$ are $\mathcal{F}_1 := \{\theta \mapsto w^\top \theta : \|w\|_2 = 1\}$ and $\mathcal{F}_{\text{proj}} = \{\theta \mapsto \theta_d : d = 1, \ldots, D\}$. In our experiments, we use the latter and therefore adopt the shorthand notation $\mathcal{I} := \mathcal{I}(\mathcal{F}_{\text{proj}})$.

3.3 Recommended workflow

Algorithm 1 outlines our recommended workflow for using BayesBag; here, $\dim(\Theta_{\mathfrak{m}})$ is the dimensionality of the parameter space of model $\mathfrak{m}$. In steps 1–3, we suggest computing the mismatch index with $M = N$ since the definition of the mismatch index is based on asymptotics, and thus, it is desirable to make M large in order to improve the accuracy of this asymptotic approximation. The mismatch index assesses the fit of the usual posterior, so there is no reason to use the same value of M that is used for robust inference with BayesBag. For very large datasets, it may be preferable to compute the mismatch index using $M < N$ in order to reduce the computation required.

Algorithm 1: Recommended workflow for BayesBag model selection.

Input: mismatch cutoff $\overline{\mathcal{I}}$ (default: 0.3),
model size cutoff factor ϱ (default: 1.0).

1 Compute Bayes posterior on Θ under the reference model.
2 Compute BayesBag posterior on Θ under the reference model, using $M = N$.
3 Compute mismatch index $\mathcal{I}$ using the results from steps 1 and 2.
4 **if** $\mathcal{I} < \overline{\mathcal{I}}$ *or* $\sum_{\mathfrak{m} \in \mathfrak{M}} \dim(\Theta_{\mathfrak{m}}) > \varrho N^{0.75}$ **then**
5 Compute BayesBag posterior on $\mathfrak{M}$ using $M = N^{0.95}$.
6 **else**
7 Compute BayesBag posterior on $\mathfrak{M}$ using $M = N^{0.75}$.

In steps 4–7, our recommendations for when to use $M = N^\alpha$ with $\alpha = 0.95$ versus $\alpha = 0.75$, and these particular values of α, should be taken as rough guidelines. The condition $\sum_{\mathfrak{m}\in\mathfrak{M}} \dim(\Theta_\mathfrak{m}) > \varrho N^{0.75}$ is meant to capture being in the "small-data" regime, where using very small bootstrap dataset sizes may result in unsatisfactory estimation accuracy. However, this precise condition may not always be appropriate; for example, it could be more appropriate to instead use $\max_{\mathfrak{m}\in\mathfrak{M}} \dim(\Theta_\mathfrak{m}) > \varrho N^{0.75}$ when the models are nested.

4 Simulation study

To validate our theory and assess the performance of BayesBag for model selection, we carry out a simulation study in the setting of feature selection for linear regression.

Model The data consist of regressors $Z_n \in \mathbb{R}^D$ and observations $Y_n \in \mathbb{R}$ for $n = 1, \dots, N$, and the goal is to predict Y_n given Z_n. For each $\gamma \in \{0,1\}^D$, define a model such that the dth regressor is included in the linear regression if and only if $\gamma_d = 1$. Letting $D_\gamma := \sum_{d=1}^D \gamma_d$ denote the number of regressors in model γ and $k^\star \in \{1, \dots, D\}$ denote the maximum number of regressors to include, we consider a collection of models $\mathfrak{M}_{k^\star} := \{\gamma \in \{0,1\}^D \,|\, D_\gamma \le k^\star\}$. Let $Z \in \mathbb{R}^{N\times D}$ denote the matrix with the nth row equal to Z_n and let Z_γ denote the submatrix of Z that includes the dth column if and only if $\gamma_d = 1$. Conditional on $\gamma \in \mathfrak{M}_{k^\star}$, the assumed model is

$$\begin{aligned}
&\sigma^2 \sim \Gamma^{-1}(a_0, b_0), \\
&\beta_d \,|\, \sigma^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2/\lambda), && d = 1, \dots, D_\gamma, \\
&Y_n \,|\, Z_\gamma, \beta, \sigma^2 \overset{\text{indep}}{\sim} \mathcal{N}(Z_{\gamma,n}^\top \beta,\, \sigma^2), && n = 1, \dots, N.
\end{aligned}$$

We parameterize the model as $\theta = (\theta_0, \dots, \theta_{D_\gamma}) = (\log\sigma^2, \beta_1, \dots, \beta_{D_\gamma}) \in \Theta_\gamma = \mathbb{R}^{D_\gamma+1}$. To perform posterior inference for γ, we analytically compute the marginal likelihood for each $\gamma \in \mathfrak{M}_{k^\star}$, integrating out σ^2 and β; specifically, for $Y := (Y_1, \dots, Y_N)^\top$, we use

$$p(Y \,|\, Z, \gamma) = \frac{b_0^{a_0}\Gamma(a_0 + N/2)}{(2\pi)^{N/2}\Gamma(a_0)} \frac{\lambda^{D_\gamma/2}}{b_\gamma^{a_0+N/2}|\Lambda_\gamma|^{1/2}},$$

where $\Lambda_\gamma := Z_\gamma^\top Z_\gamma + \lambda I$ and $b_\gamma := b_0 + Y^\top(I - Z_\gamma \Lambda_\gamma^{-1} Z_\gamma^\top)Y/2$. For the prior on $\gamma \in \mathfrak{M}_{k^\star}$, we let $Q_0(\gamma) \propto q_0^{D_\gamma}(1-q_0)^{D-D_\gamma}$, where $q_0 \in (0,1)$ is the prior inclusion probability of each component. Thus, the posterior probability of model γ is

$$Q(\gamma \,|\, Y, Z) = \frac{p(Y \,|\, Z, \gamma) Q_0(\gamma)}{\sum_{\gamma' \in \mathfrak{M}_{k^\star}} p(Y \,|\, Z, \gamma') Q_0(\gamma')}$$

and the *posterior inclusion probability* of the dth regressor is

$$Q(\gamma_d = 1 \,|\, Y, Z) := \frac{\sum_{\gamma \in \mathfrak{M}_{k^\star}} \gamma_d \, p(Y \,|\, Z, \gamma) Q_0(\gamma)}{\sum_{\gamma' \in \mathfrak{M}_{k^\star}} p(Y \,|\, Z, \gamma') Q_0(\gamma')}. \tag{5}$$

Data We simulate data by generating $Z_n \overset{\text{i.i.d.}}{\sim} G$, $\epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and

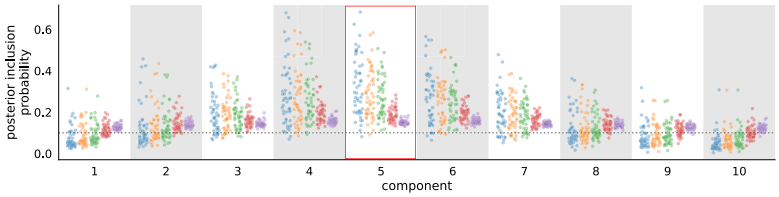
$$Y_n = f(Z_n)^\top \beta_\dagger + \epsilon_n \tag{6}$$

for $n = 1, \ldots, N$, with the regressor distribution G, the regression function f, and the coefficient vector $\beta_\dagger \in \mathbb{R}^D$ as described next. Using the linear regression function $f(z) = z$ results in well-specified data. To generate misspecified data, we use the nonlinear component-wise cubic function $f(z) = (z_1^3, \ldots, z_D^3)^\top$. We choose G and $\beta_\dagger$ in the spirit of genome-wide association study fine-mapping (Schaid, Chen and Larson, 2018) to simulate a scenario with many highly correlated regressors, of which only a few regressors are actually employed in the data-generating process. For $k \in \{1, 2\}$, we use a k-sparse vector (that is, a vector with k non-zero components) defined by setting $\beta_{\dagger d} = 1$ if $d \in \{\lfloor j(D + \frac{1}{2})/(k+1) \rfloor \,|\, j = 1, \ldots, k\}$ and $\beta_{\dagger d} = 0$ otherwise. For $h > 2$ and $\psi > 0$, $Z \sim G$ is defined by generating $\xi \sim \chi^2(h)$ and then $Z \,|\, \xi \sim \mathcal{N}(0, \Sigma)$, where the (d, d') entry of $\Sigma \in \mathbb{R}^{D \times D}$ is given by $\Sigma_{dd'} = \exp\{-(d-d')^2/\psi^2\}/(\xi_d \xi_{d'})$, and $\xi_d = \sqrt{\xi/(h-2)}$ if d is odd and $\xi_d = 1$ otherwise. The motivation for this data simulation procedure is to generate correlated regressors that have different tail behaviors while still having the same first two moments, since regressors are typically standardized to have mean 0 and variance 1. Note that, marginally, $Z_1, Z_3, \ldots$ are each rescaled t-distributed random variables with h degrees of freedom such that $\mathrm{Var}(Z_1) = 1$, and $Z_2, Z_4, \ldots$ are $\mathcal{N}(0,1)$.

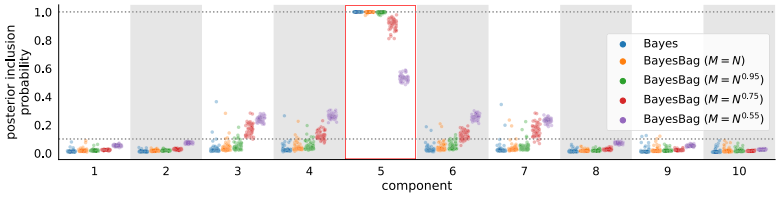
Experimental conditions We generate datasets under the k-sparse-linear and k-sparse-nonlinear settings according to (6) with $h = 10$, $\psi = 8$, and either $(D, k) = (10, 1)$ or $(D, k) = (20, 2)$. We set $q_0 = k/D$ and the model hyperparameters to $a_0 = 2$, $b_0 = 1$, and $\lambda = 16$, with the latter setting helping to penalize the addition of extraneous features. We consider $M = N^\alpha$ for $\alpha \in \{1, 0.95, 0.75, 0.55\}$. We consider $k^\star \in \{1, 2\}$ for 1-sparse data and $k^\star = 2$ for 2-sparse data. We then compute the posterior inclusion probabilities as defined in (5). Each experimental condition is replicated 50 times, resulting in 50 posterior inclusion probabilities for each regressor in each experimental setting.

Results We are interested in verifying the theory of Section 3 in the finite-sample regime, which suggests that when the model is misspecified, similar models may be assigned wildly varying probabilities under the usual posterior (Bayes), while the bagged posterior (BayesBag) probabilities will tend to be more balanced. Figures 3, 4, and S.1 to S.3 in the Supplementary Material show the Bayes and BayesBag posterior inclusion probabilities for each component, for all 50 replications. First, Figure 3 shows that when the model is correctly specified, the Bayes and BayesBag posteriors with $\alpha \geq 0.95$ behave similarly. However, BayesBag can be more stable even in this well-specified setting, exhibiting fewer outlier posterior inclusion probabilities. As α decreases, BayesBag yields substantially more conservative inferences, in the sense that the posterior inclusion probabilities tend to shrink toward the prior inclusion probability.

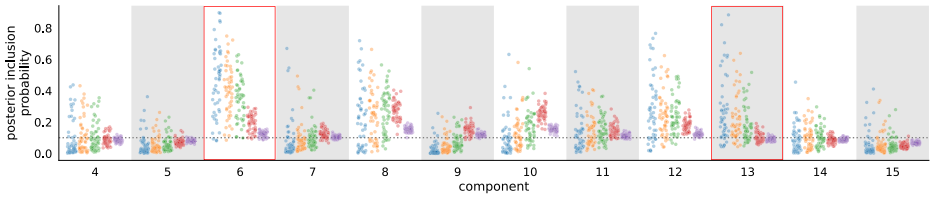
In the misspecified setting, the results are more interesting and subtle (Figures 4, S.1–S.3 in the Supplementary Material). Due to the misspecification and correlated regressors, it no longer holds in general that the components that were actually non-null in the data-generating process will be selected (see Section S.2 of the Supplementary



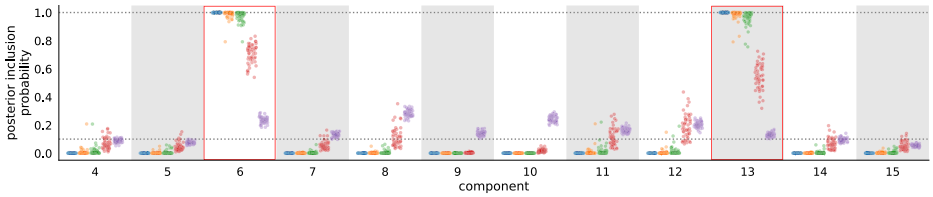
(a) 1-sparse-linear, $N = 5 \times 10^1$



(b) 1-sparse-linear, $N = 5 \times 10^3$

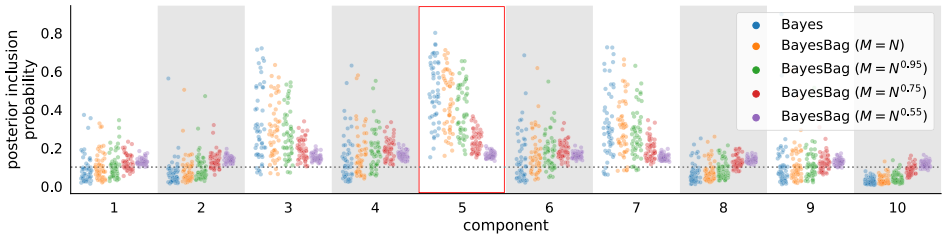


(c) 2-sparse-linear, $N = 10^2$

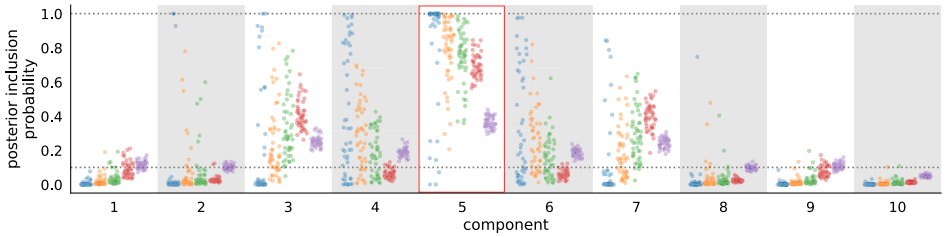


(d) 2-sparse-linear, $N = 10^3$

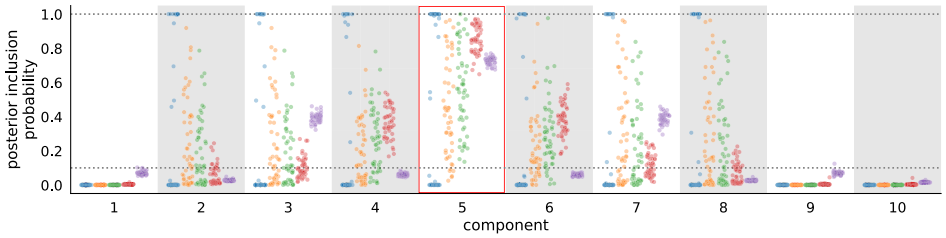
Figure 3: Simulation results for feature selection in linear regression with $k^\star = 2$ when the model contains the true distribution. The BayesBag posterior inclusion probabilities are similar to the Bayes posterior inclusion probabilities, but tend to shrink toward the prior inclusion probability (lower horizontal dotted line). The data was generated from the assumed model, $Y_n = Z_n^\top \beta_\dagger + \epsilon_n$ for $n = 1, \ldots, N$, where $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 1$, $Z_n \in \mathbb{R}^D$ is a vector of covariates, and $\beta_\dagger \in \mathbb{R}^D$ is a k-sparse vector, that is, $\beta_\dagger$ has k nonzero components. The prior on inclusion vectors $\gamma \in \{0,1\}^D$ is proportional to $q_0^{\sum \gamma_d}(1-q_0)^{D-\sum \gamma_d}$, where $q_0 = k/D$, with the constraint that $\sum_{d=1}^D \gamma_d \leq k^\star$. A conjugate prior is placed on the coefficients and σ^2 given γ. The posterior inclusion probabilities were computed by analytically integrating out the parameters and summing over all binary inclusion vectors γ. The figure shows results for simulations using (a) $D = 10$, $N = 50$, $k = 1$, (b) $D = 10$, $N = 5{,}000$, $k = 1$, (c) $D = 20$, $N = 100$, $k = 2$, and (d) $D = 20$, $N = 1{,}000$, $k = 2$. For each of these settings, 50 replicate datasets were generated, and the resulting posterior inclusion probabilities are shown. Components that were actually nonzero when generating the data are enclosed by red rectangles.



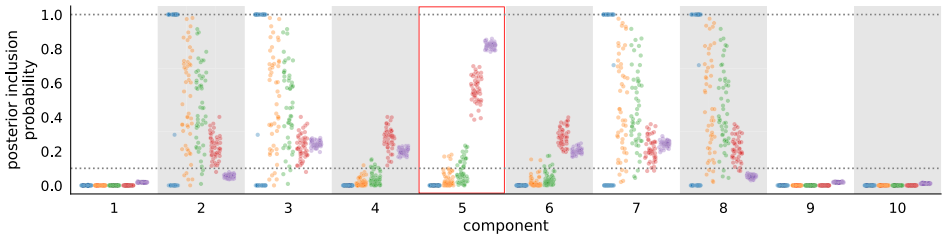
(a) $N = 5 \times 10^1$



(b) $N = 5 \times 10^2$



(c) $N = 5 \times 10^3$



(d) $N = 5 \times 10^4$

Figure 4: Simulation results for feature selection in linear regression with 1-sparse-nonlinear data (so the model is misspecified) and $k^\star = 2$. Everything is the same as in Figure 3(a, b), except that the data was generated using $Y_n = f(Z_n)^\top \beta_\dagger + \epsilon_n$, where $f(z) = (z_1^3, \ldots, z_D^3)^\top$. Results are shown for (a) $N = 50$, (b) $N = 500$, (c) $N = 5{,}000$, and (d) $N = 50{,}000$. See the caption of Figure 3 for further explanation. The Bayes posterior inclusion probabilities show considerable instability both (i) across datasets with N fixed and (ii) as N increases. Meanwhile, the BayesBag probabilities are much more stable, particularly for $M = N^\alpha$ with $\alpha \leq 0.75$. The component that was actually nonzero when generating the data is enclosed by a red rectangle; see the text for interpretation.

Material and Buja et al., 2019a,b). For the 1-sparse-nonlinear data, when $k^\star = 1$, the Bayes and BayesBag posteriors behave quite similarly and concentrate on component 5, which is asymptotically optimal; see Figure S.1 in the Supplementary Material. However, when $k^\star = 2$, two models are asymptotically optimal and equivalent, namely, the models with $\text{supp}(\gamma) = \{2, 3\}$ and $\text{supp}(\gamma) = \{7, 8\}$, where $\text{supp}(\gamma) := \{d : \gamma_d \neq 0\}$. Meanwhile, $\{4, 5\}$ and $\{5, 6\}$ are asymptotically equivalent but slightly less-than-optimal. As shown in Figure 4, in this case the Bayes posterior is unstable and, for large values of N, concentrates on $\{2, 3\}$ or $\{7, 8\}$ with equal probability. For $M \in \{N, N^{0.95}\}$, BayesBag places roughly uniformly distributed mass on the same four components for large values of N. Meanwhile, for $M \in \{N^{0.75}, N^{0.55}\}$, BayesBag is much more stable and puts more mass on component 4, 5, and 6. Thus, we see exactly the behaviors predicted by the asymptotic analyses in Section 3. We defer discussion of the results for 2-sparse-nonlinear data (Figures S.2 and S.3 in the Supplementary Material) to Section S.1 of the Supplementary Material.

Figures 5 and S.4 in the Supplementary Material show model-data mismatch index values for the reference model with $\gamma_d = 1$ for all $d = 1, \ldots, D$, on a representative subset of experimental configurations. For the k-sparse-linear data, the overall mismatch indices were either near zero or were NA, reflecting that the model is correctly specified but there are some issues with poor identifiability. For the k-sparse-nonlinear data, the mismatch indices were nearly all NA, reflecting that the model is misspecified and there may also be identifiability issues.

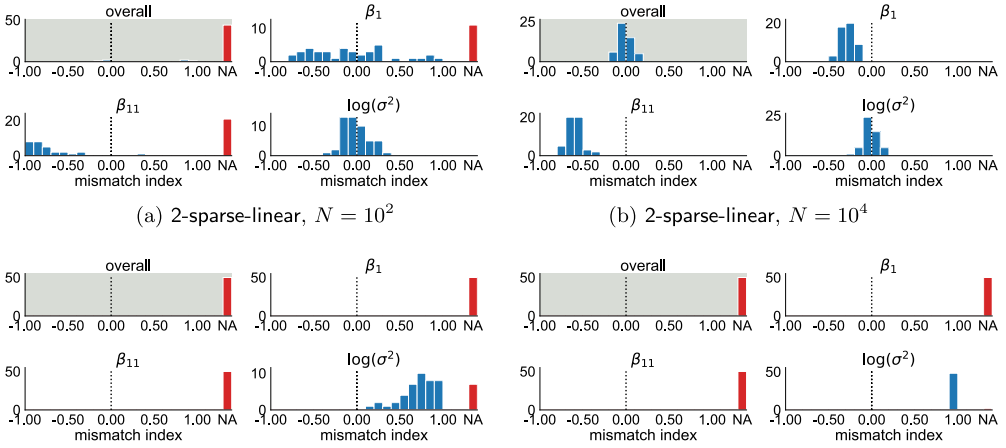
Summary Overall, the simulation results are in agreement with our asymptotic theory from Section 3: the behavior of Bayes can vary dramatically with the dataset size and the degree of misspecification, whereas BayesBag is much more stable. Additionally, the simulations provide insight into the behavior of the bagged posterior when M is sublinear in N. Of particular note is that $M = N^{0.95}$ yields noticeably improved stability with little loss of statistical efficiency. Meanwhile, for settings with substantial misspecification, taking $M = N^\alpha$ with $\alpha \in [0.55, 0.75]$ may be preferable – with the caveat that inferences will tend to be more conservative.

5 Applications

5.1 Feature selection for linear regression

We compare Bayesian model selection and BayesBag model selection for linear regression on four real-world datasets, summarized in Table 1. Based on our findings in Section 4, for BayesBag we consider $M = N^\alpha$ with $\alpha \in \{1.0, 0.95, 0.75\}$. We use a prior inclusion probability of $q_0 = 3/D$ and use $k^\star = D$ for the maximum number of nonzero components, except on the residential building dataset, where for computational tractability we use $k^\star = 3$. We set the model hyperparameters to $a_0 = 2$, $b_0 = 1$, and $\lambda = 1$.

We expect the parameters to be well-identified for all datasets except the residential building dataset, since the residential building dataset requires only 58 out of 104 principal components to explain 99% of the variance, whereas for the other three datasets,



(a) 2-sparse-linear, $N = 10^2$ (b) 2-sparse-linear, $N = 10^4$

(c) 2-sparse-nonlinear, $N = 10^2$ (d) 2-sparse-nonlinear, $N = 10^4$

Figure 5: Model-data mismatch indices $\mathcal{I}$ from the simulations on feature selection in linear regression. The overall $\mathcal{I}$ value and the $\mathcal{I}$ for selected parameters are shown in the case of 2-sparse data ($k = 2$) with $D = 20$ regressors. The figure shows histograms of $\mathcal{I}$ over 50 replicate datasets generated using the well-specified linear case of $f(z) = z$ (as in Figure 3) with (a) $N = 100$ and (b) $N = 10{,}000$, and the misspecified nonlinear case of $f(z) = (z_1^3, \ldots, z_D^3)^\top$ (as in Figure 4) with (c) $N = 100$ and (d) $N = 10{,}000$. We only display two components of β since the $\mathcal{I}$ values follow fairly similar distributions for all components. The results show that in the well-specified setting, when N is sufficiently large (panel (b)), $\mathcal{I}$ tends to be near zero, indicating correct specification as expected. An exception is that the $\mathcal{I}$ value for β_{11} is closer to -1, indicating that the Bayes posterior on β_{11} may be somewhat underconfident. When N is small (panel (a)), $\mathcal{I}$ is often NA in these simulations, reflecting the poor identifiability of the coefficients due to strong correlation in the regressors. Meanwhile, in the misspecified setting (panels (c) and (d)), $\mathcal{I}$ is typically NA for the coefficients, reflecting that the model is misspecified and there may also be identifiability issues.

D out of D principal components are needed to explain 99% of the variance. The model mismatch indices (for the reference model with $\gamma_d = 1$ for all $d = 1, \ldots, D$) are in agreement with expectations, since only the residential building dataset has a model mismatch index of NA. For the other datasets, the mismatch indices are 1.00 (California housing), 0.62 (Boston housing), and 0.03 (Diabetes), which suggests that the model is misspecified for the two housing datasets.

Figure 6 shows the posterior inclusion probabilities for all four datasets. To compare the reliability of the methods, we also run each method on subsets of the data obtained by randomly dividing each dataset into roughly equally sized splits (Figure 6). We use three splits for all datasets except for California housing, for which we use five splits since N is substantially larger. Generally, across splits, BayesBag produced lower-variance, more conservative posterior inclusion probabilities that are more consistent with the

Name	Model	N	D
California housing	LR	20,650	8
Boston housing	LR	506	13
Diabetes	LR	442	10
Residential building	LR	371	105
Whale mitochondrial coding DNA	PTR	10,605	14
Whale mitochondrial amino acids	PTR	3,535	14

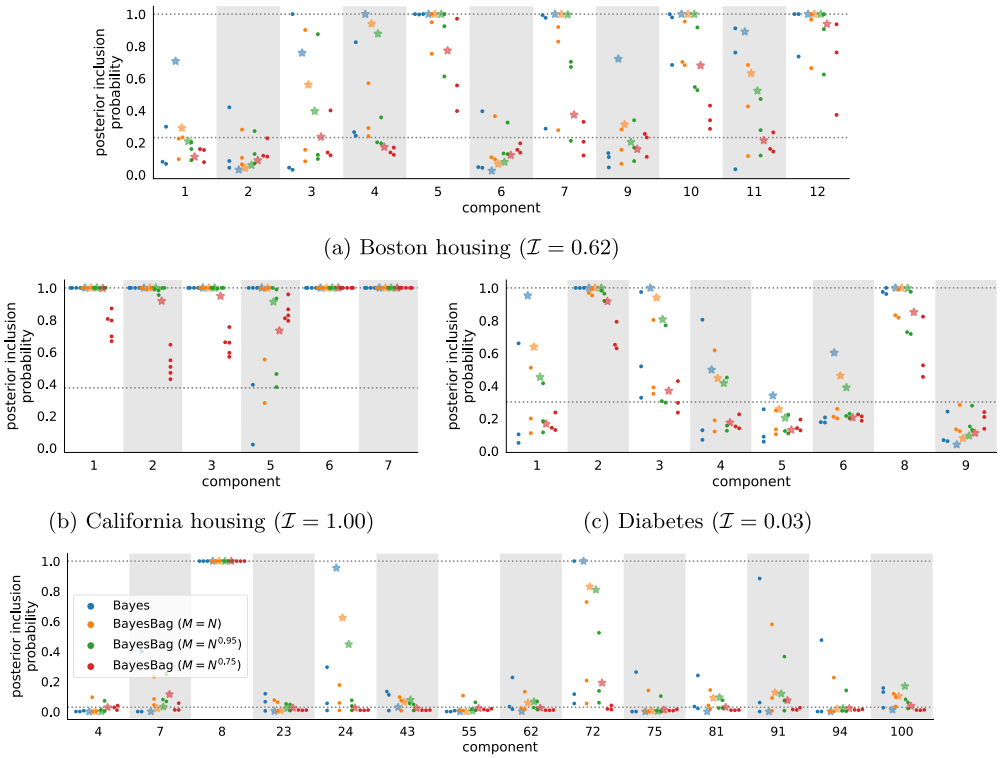
Table 1: Real-world datasets used in experiments. LR = linear regression, PTR = phylogenetic tree reconstruction. For LR, $N = \#$ samples and $D = \#$ covariates. For PTR, $N = \#$ features and $D = \#$ species.

posterior inclusion probabilities from the full datasets. BayesBag with $M = N^{0.75}$ is noticeably more conservative than Bayes and BayesBag with $M \in \{N^{0.95}, N\}$; for the two datasets with mismatch indices that suggest significant misspecification (California housing and residential building), such stability appears particularly desirable. These results are in agreement with the simulation results in Section 4.

5.2 Phylogenetic tree reconstruction

Finally, we investigate the use of BayesBag for reconstructing the phylogenetic tree of a collection of species based on their observed characteristics. This is an important model selection problem due to the widespread use of phylogeny reconstruction algorithms. Systematists have exhaustively documented that Bayesian model selection of phylogenetic trees can behave poorly. In particular, the posterior can provide contradictory results depending on what characteristics are used (for example, coding deoxyribonucleic acid [DNA] or amino acid sequences), what evolutionary model is used, or which outgroups are included (Alfaro, Zoller and Lutzoni, 2003; Buckley, 2002; Douady et al., 2003; Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004; Waddell, Kishino and Ota, 2002; Wilcox et al., 2002; Yang, 2007). We illustrate how BayesBag model selection provides reasonable inferences that are significantly more robust to the choice of data and model.

Models and data We use the whale dataset from Yang (2008), consisting of mitochondrial coding DNA from 13 whale species and the hippopotamus (Table 1). The hippopotamus is included as an "outgroup" species to identify the root of the tree, because the assumed evolutionary models are time-reversible and hence the trees are modeled as unrooted. We consider four DNA models (JC, HKY+C+Γ_5, GTR+Γ+I, and mixed+Γ_5) and one amino acid model (mtmam+Γ_5); see Yang (2008) for more details on these models and an explanation of the acronyms. For brevity, we refer to the models as, respectively, JC, HKY, GTR, mixed, and mtmam. To approximate the usual posterior (Bayes) and the bagged posterior (BayesBag), we use MrBayes 3.2 (Ronquist et al., 2012) with 2 independent runs, each with 4 coupled chains run for 1,000,000 total iterations (discarding the first quarter as burn-in). We confirm acceptable mixing



(a) Boston housing ($\mathcal{I} = 0.62$)

(b) California housing ($\mathcal{I} = 1.00$)

(c) Diabetes ($\mathcal{I} = 0.03$)

(d) Residential building ($\mathcal{I} = \mathsf{NA}$)

Figure 6: Application to feature selection on four real-world datasets; see Table 1 for dataset details. The assumed model is the same as in Section 4 (see also the caption of Figure 3), using a prior inclusion probability of $q_0 = 3/D$ (lower horizontal dotted line), where D is the number of regressors. To assess reproducibility, we randomly split each dataset into roughly equally sized parts, and computed the posterior inclusion probabilities for each split separately (indicated with a $\bullet$) as well as for the full dataset (indicated with a $\star$). As before, the posterior inclusion probabilities are computed by analytically integrating out the parameters and summing out all possible binary inclusion vectors. For computational tractability, we constrain the model for the residential building dataset to only allow up to $k^\star = 3$ nonzero components. For visual readability, we only display the components with at least one posterior inclusion probability greater than $\min(0.25, 3q_0)$. The BayesBag posterior inclusion probabilities exhibit greater reproducibility, in that (i) the between-split differences tend to be smaller and (ii) the differences between the split posterior inclusion probabilities and the full data posterior inclusion probabilities also tend to be smaller for BayesBag than Bayes.

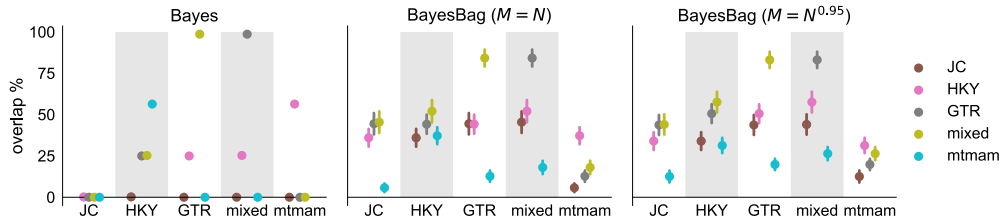
using the built-in convergence diagnostics for MrBayes. For BayesBag, we take $B = 100$ in all experiments and, since the number of models is very large, we only consider $M \in \{N, N^{0.95}\}$.

Evaluation Our goal is to investigate whether BayesBag avoids the self-contradictory inferences that Bayes produces. To this end, we compare the output of different configurations of the data, model, and inference method, as follows. We compute the set of trees in the 99% highest posterior probability (HPP) regions for each ⟨data, model, inference method⟩ configuration. For selected pairs of configurations, we then compute the overlap of the two 99% HPP regions in terms of (a) probability mass and (b) number of trees. Since the BayesBag posterior is approximated via Monte Carlo as in (2), we quantify the uncertainty in each overlap by reporting an 80% confidence interval for the overlapping mass. We compute these intervals using standard bootstrap methodology for a Monte Carlo estimate.

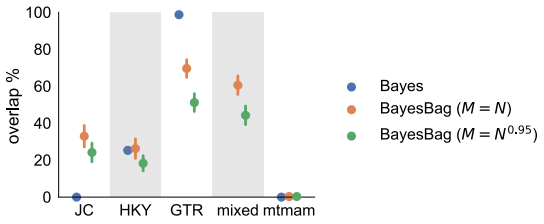
Results First, we look at the overlap between pairs of models. As shown in Figure 7(a) and Table S.1 in the Supplementary Material, there is substantially more overlap when using BayesBag. The difference is particularly noticeable when comparing JC (the simplest model) or mtmam (the amino acid model) to the other models. When using Bayes, JC has either 0% or (in one case) 0.2% overlap with the other models while mtmam only overlaps with HKY. Thus, these pairs of models produce contradictory results when using Bayesian model selection. On the other hand, when using BayesBag, all pairs of models have nonzero overlap, with typical amounts ranging from 30% to 50%. Hence, compared to Bayes, BayesBag provides results that are more consistent across models.

However, the good overlap between BayesBag posteriors does not necessarily mean that it is performing well, since it could simply be producing posteriors that are too diffuse, spreading the posterior mass over a very large number of trees. Notably (as expected), BayesBag with $M = N^{0.95}$ leads to a more diffuse posterior with 5–15 overlapping trees compared to 3–11 trees when $M = N$. To further investigate the possibility of the BayesBag posteriors being too diffuse, we consider the overlap of the BayesBag posterior for each model and the Bayes posterior for mixed, which is the most complex of the DNA models. As shown in Figure 7(b) and Table S.2 in the Supplementary Material, all of the BayesBag posteriors (with the exception of mtmam) put substantial posterior probability on the 99% HPP region of the Bayes mixed posterior. Moreover, all but BayesBag mtmam has two trees in the overlap, which is the maximum possible since the Bayes mixed 99% HPP region only contains two trees. Finally, using BayesBag with $M = N^{0.95}$ results in fairly small decreases (relative to BayesBag with $M = N$) in the mass on the two trees in the Bayes mixed 99% HPP region.

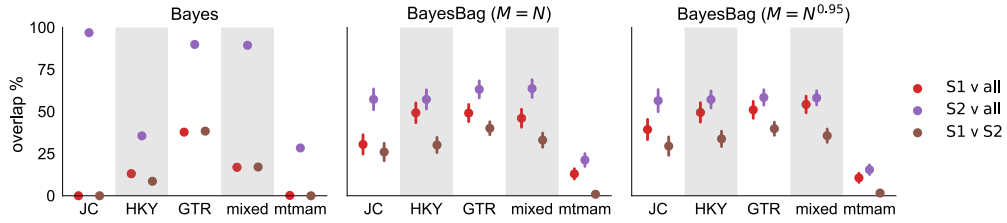
Next, we perform intra-model comparisons by considering three datasets: the complete whale dataset (denoted all) and two additional datasets formed by splitting the genomic data for each species in half (denoted S1 and S2). Ideally, for each model, we hope to see substantial overlap when comparing the results across these three datasets (all, S1, and S2). However, when using the Bayes posterior, there is little to no overlap in many cases, particularly for the simpler JC model and mtmam; see Figure 7(c) and Table S.3 in the Supplementary Material. Meanwhile, the BayesBag posteriors typically exhibit overlaps of between 21% and 56%, with less (though still nonzero) overlap with mtmam. These results suggest that BayesBag exhibits superior reproducibility in terms of uncertainty quantification.



(a) Inter-model comparison



(b) Comparison to mixed using usual Bayesian posterior



(c) Intra-model comparison

Figure 7: Application to phylogenetic tree inference on a whale genetics dataset. To assess reproducibility, we computed the posterior under five different evolutionary models (JC, HKY, GTR, mixed, mtmam). We quantify the similarity of posteriors by computing the overlapping probability mass of 99% highest posterior probability regions. To quantify uncertainty in the overlap due to Monte Carlo error, 80% confidence intervals are shown for the overlaps involving BayesBag. Panel (a) shows the posterior overlap for each pair of models. The usual posterior (Bayes) is quite sensitive to the choice of model, exhibiting $\approx 0\%$ overlap in many cases, for instance, between JC and the other models. Meanwhile, the bagged posterior (BayesBag) is more robust, exhibiting overlaps in a reasonable range. Panel (b) shows the overlap between the Bayes posterior for the mixed model, which is the most flexible of the DNA models, and the Bayes or BayesBag posterior for each other model. Panel (c) shows the overlap when using the same model on different subsets of the data — specifically, splitting the genomic data for each species into two halves (S1, S2) or using the complete data (all).

Finally, we compute the mismatch index for each model on the complete whale dataset, obtaining 0.23 (JC), NA (HKY), 0.47 (GTR), 0.84 (mixed), and 0.34 (mtmam). These mismatch indices suggest significant amounts of model misspecification, with the

simpler JC model likely underestimating the actual degree of misspecification. Thus, using the BayesBag posterior with $M = N^{0.95}$ appears to be advisable.

6 Discussion

In this paper, we have developed an approach to overcome the instability of Bayesian model selection when the models are all misspecified. This type of misspecification is common in scientific settings where idealized but interpretable models are commonly used (such as in systematics, population and cancer genetics, and economics). Our bagged posterior approach is theoretically justified, easy to use, and widely applicable. However, we see three potential limitations in practice. The first is that bagged posterior model selection tends to be more conservative, with posterior model probabilities farther from the extremes of zero and one. The recommended workflow discussed in Section 3.3 is designed to at least partially ameliorate this issue, however, this conservative behavior may be a necessary price for greater stability and reliability. The second limitation is the additional computational cost required for the naive estimation of the bagged model probabilities. The development of more computationally efficient alternatives is an important direction for future work. A final limitation is that our asymptotic theory only covers cases where the observations are independent. Extending the theory to cover important structured models like time-series and spatial models is another valuable direction for future work.

Supplementary Material

Supplementary Material: Reproducible Model Selection Using Bagged Posteriors (DOI: 10.1214/21-BA1301SUPP; .pdf).

References

Alfaro, M. E., Zoller, S. and Lutzoni, F. (2003). Bayes or Bootstrap? A Simulation Study Comparing the Performance of Bayesian Markov Chain Monte Carlo Sampling and Bootstrapping in Assessing Phylogenetic Confidence. *Molecular Biology and Evolution* **20** 255–266. 82, 98

Berk, R. H. (1966). Limiting Behavior of Posterior Distributions when the Model is Incorrect. *The Annals of Mathematical Statistics* **37** 51–58. MR0189176. doi: https://doi.org/10.1214/aoms/1177699477. 79, 81

Breiman, L. (1996). Bagging Predictors. *Machine Learning* **24** 123–140. 80

Buckley, T. R. (2002). Model Misspecification and Probabilistic Tests of Topology: Evidence from Empirical Data Sets. *Systematic Biology* **51** 509–523. 98

Bühlmann, P. (2014). Discussion of Big Bayes Stories and BayesBag. *Statistical Science* **29** 91–94. MR3201850. doi: https://doi.org/10.1214/13-STS460. 80, 83

Bühlmann, P. and Yu, B. (2002). Analyzing Bagging. *The Annals of Statistics* **30** 927–961. MR1926165. doi: https://doi.org/10.1214/aos/1031689014. 84

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K. and Zhao, L. H. (2019a). Models as Approximations I: Consequences Illustrated with Linear Regression. *Statistical Science* **34** 523–544. MR4048582. doi: https://doi.org/10.1214/18-STS693. 96

Buja, A., Brown, L., Kuchibhotla, A. K., Berk, R., George, E. and Zhao, L. H. (2019b). Models as Approximations II: A Model-Free Theory of Parametric Regression. *Statistical Science* **34** 545–565. MR4048583. doi: https://doi.org/10.1214/18-STS694. 96

Clarke, B. S. and Barron, A. R. (1990). Information-Theoretic Asymptotics of Bayes Methods. *Information Theory, IEEE Transactions on* **36** 453–471. MR1053841. doi: https://doi.org/10.1109/18.54897. 88

Douady, C. J., Delsuc, F., Boucher, Y., Doolittle, W. F. and Douzery, E. J. P. (2003). Comparison of Bayesian and Maximum Likelihood Bootstrap Measures of Phylogenetic Reliability. *Molecular Biology and Evolution* **20** 248–254. 80, 82, 84, 98

Huelsenbeck, J. P. and Rannala, B. (2004). Frequentist Properties of Bayesian Posterior Probabilities of Phylogenetic Trees Under Simple and Complex Substitution Models. *Systematic Biology* **53** 904–913. doi: https://doi.org/10.1080/10635150490522629. 98

Huggins, J. H. and Miller, J. W. (2019). Robust Inference and Model Criticism Using Bagged Posteriors. *arXiv.org* arXiv:1912.07104 [stat.ME]. 80, 83, 89, 91

Huggins, J. H. and Miller, J. W. (2022). Supplementary Material: Reproducible Model Selection Using Bagged Posteriors. *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1301SUPP. 88

Lemmon, A. R. and Moriarty, E. C. (2004). The Importance of Proper Model Assumption in Bayesian Phylogenetics. *Systematic Biology* **53** 265–277. 98

Meng, L. and Dunson, D. B. (2020). Comparing and Weighting Imperfect Models using D-Probabilities. *Journal of the American Statistical Association* **115** 1349–1360. MR4143470. doi: https://doi.org/10.1080/01621459.2019.1611140. 80, 82

Oelrich, O., Ding, S., Magnusson, M., Vehtari, A. and Villani, M. (2020). When are Bayesian Model Probabilities Overconfident? *arXiv.org* arXiv:2003.04026 [math.ST]. 80, 82, 84

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. and Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology* **61** 539–542. 98

Schaid, D. J., Chen, W. and Larson, N. B. (2018). From Genome-Wide Associations to Candidate Causal Variants by Statistical Fine-Mapping. *Nature Reviews Genetics* **19** 1–14. 93

Waddell, P. J., Kishino, H. and Ota, R. (2002). Very Fast Algorithms for Evaluating the Stability of ML and Bayesian Phylogenetic Trees from Sequence Data. *Genome Informatics* **13** 82–92. 80, 84, 98

Wilcox, T. P., Zwickl, D. J., Heath, T. A. and Hillis, D. M. (2002). Phylogenetic Relationships of the Dwarf Boas and a Comparison of Bayesian and Bootstrap Measures of Phylogenetic Support. *Molecular Phylogenetics and Evolution* **25** 361–371. doi: https://doi.org/10.1016/S1055-7903(02)00244-0. 82, 98

Yang, Z. (2007). Fair-Balance Paradox, Star-Tree Paradox, and Bayesian Phylogenetics. *Molecular Biology and Evolution* **24** 1639–1655. 98

Yang, Z. (2008). Empirical Evaluation of a Prior for Bayesian Phylogenetic Inference. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363** 4031–4039. 98

Yang, Z. and Zhu, T. (2018). Bayesian Selection of Misspecified Models is Overconfident and May Cause Spurious Posterior Probabilities for Phylogenetic Trees. *Proceedings of the National Academy of Sciences* **115** 1854–1859. MR3779786. doi: https://doi.org/10.1073/pnas.1712673115. 80, 82, 84

Acknowledgments

We thank Pierre Jacob for bringing P. Bühlmann's BayesBag paper to our attention, Ziheng Yang for sharing the whale dataset and his MrBayes scripts, Ryan Giordano and Pierre Jacob for helpful feedback on an earlier draft of this paper, Peter Grünwald, Natalia Bochkina, Mathieu Gerber, and Anthony Lee for helpful discussions, the AE and three reviewers for their constructive comments, and especially the third reviewer, who provided numerous insightful comments that substantially enhanced the scope and readability of the paper.