USING BAGGED POSTERIORS FOR ROBUST INFERENCE AND MODEL CRITICISM

BY JONATHAN H. HUGGINS[1] AND JEFFREY W. MILLER[2]

[1]*Department of Mathematics & Statistics, Boston University, huggins@bu.edu*

[2]*Department of Biostatistics, Harvard T. H. Chan School of Public Health, jwmiller@hsph.harvard.edu*

Standard Bayesian inference is known to be sensitive to model misspecification, leading to unreliable uncertainty quantification and poor predictive performance. However, finding generally applicable and computationally feasible methods for robust Bayesian inference under misspecification has proven to be a difficult challenge. An intriguing, easy-to-use, and widely applicable approach is to use bagging on the Bayesian posterior ("BayesBag"); that is, to use the average of posterior distributions conditioned on bootstrapped datasets. In this paper, we comprehensively develop the asymptotic theory of BayesBag, propose a model–data mismatch index for model criticism using BayesBag, and empirically validate our theory and methodology on synthetic and real-world data in linear regression (both feature selection and parameter inference), sparse logistic regression, a hierarchical mixed effects model, and phylogenetic tree reconstruction. We find that in the presence of significant misspecification, BayesBag yields more reproducible inferences, has better predictive accuracy, and selects correct models more often than the standard Bayesian posterior; meanwhile, when the model is correctly specified, BayesBag produces superior or equally good results for parameter inference and prediction, while being slightly more conservative for model selection. Overall, our results demonstrate that BayesBag combines the attractive modeling features of standard Bayesian inference with the distributional robustness properties of frequentist methods, providing benefits over both Bayes alone and the bootstrap alone.

1. Introduction. Bayesian inference is premised on the data being generated from the assumed model. In practice, however, it is widely recognized that models are (sometimes gross) approximations to reality (Box, 1979, 1980; Cox, 1990; Lehmann, 1990). Moreover, even when the model is nearly correct, the optimal parameter (in terms of log loss or Kullback–Leibler divergence) may be extremely unlikely under the prior distribution, which can bias the posterior distribution and lead to poor predictive performance. Thus, in order to effectively use Bayesian methods, it is crucial to be able to both diagnose and correct for mismatch between the model and the data during the model building process (Blei, 2014; Gelman et al., 2013). The task of diagnosing mismatch/misspecification is often termed "model criticism" or "model assessment" (Gelman and Shalizi, 2011; Vehtari and Ojanen, 2012). Yet, eventually model building must cease due either to resource limitations (e.g., the data analyst's time, knowledge about the phenomena under study, or computational capacity) or norms within a field limiting the set of models the analyst can consider.

Keywords and phrases: Bagging, Bernstein–Von Mises theorem, Bootstrap, Model criticism, Model misspecification, Uncertainty calibration

Therefore, whatever model is ultimately used, the analyst may still need to rely on robust inference methods to correct for any remaining model–data mismatch.

This article develops the theory and practice of *BayesBag*, a simple and widely applicable approach to robust Bayesian inference that also provides diagnostics for model criticism. Originally developed by Waddell, Kishino and Ota (2002) and Douady et al. (2003) in the context of phylogenetic inference and then independently proposed by Bühlmann (2014) (where the name was coined), the idea of BayesBag is to apply bagging (Breiman, 1996) to the Bayesian posterior. (As we show in Appendix A, it can also be interpreted as an approximation to Jeffrey conditionalization.) The *bagged posterior* $\pi^*(\theta \,|\, x)$ is defined 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:

$$\pi^*(\theta \,|\, x) := \frac{1}{N^M} \sum_{x^*} \pi(\theta \,|\, x^*), \tag{1}$$

where $\pi(\theta \,|\, x^*) \propto \pi_0(\theta) \prod_{m=1}^M p_\theta(x_m^*)$ is the standard posterior density given data x^* and the sum is over all possible N^M bootstrap datasets of M samples drawn with replacement from the original dataset. In practice, we can approximate $\pi^*(\theta \,|\, x)$ by generating B bootstrap datasets $x_{(1)}^*, \ldots, x_{(B)}^*$, where $x_{(b)}^*$ consists of M samples drawn with replacement from x, yielding the approximation

$$\pi^*(\theta \,|\, x) \approx \frac{1}{B} \sum_{b=1}^{B} \pi(\theta \,|\, x_{(b)}^*). \tag{2}$$

BayesBag is easy to use since the bagged posterior is simply an average over standard Bayesian posteriors, which means no additional algorithmic tools are needed beyond what a data analyst would use for posterior inference in the original model. While BayesBag does require more computational resources since one must approximate B posteriors (each conditioned on a bootstrap dataset) where typically $B \approx$ 50–100, each posterior can be approximated in parallel, which is ideal for modern cluster-based high-performance computing environments. Surprisingly, despite these attractive features, there has been little practical or theoretical investigation of BayesBag. In the only previous work of which we are aware, Bühlmann (2014) (which is a short discussion paper) presented only a few simulation results in a simple Gaussian location model, while Waddell, Kishino and Ota (2002) and Douady et al. (2003) undertook limited investigations in the setting of phylogenetic tree inference in papers focused primarily on speeding up model selection (in the former) and comparing Bayesian inference and the bootstrap (in the latter).

In this paper, we show that the bagged posterior has appealing statistical properties in the presence of model misspecification, while also being easy to use and computationally tractable on a range of practical problems. The bagged posterior integrates the attractive features of Bayesian inference—such as flexible hierarchical modeling and the use of prior information—with the distributional robustness of frequentist methods, nonparametrically accounting for sampling variability and model misspecification. Moreover, rather than just providing robustness to misspecification, our BayesBag methodology can simultaneously

diagnose the degree of misspecification, or more generally, the degree of model–data mismatch.

The organization and main contributions of the paper are as follows. In Section 2, we begin by briefly providing representative examples of the superior empirical performance of the bagged posterior compared to the standard posterior for both estimating the optimal parameter and selecting the optimal model, where optimality is in terms of log loss. Next, we introduce our BayesBag methodology in more detail and describe our *model–data mismatch index* for performing model criticism. This mismatch index is based on how much the standard and bagged posterior variances differ, compared to what would be expected if the model were correctly specified. In order to explain the empirical performance of BayesBag and justify our mismatch index, we sketch out our statistical theory for the bagged posterior.

We then proceed to fully develop our asymptotic theory of the bagged posterior in both the parameter inference and model selection settings. For parameter inference (Section 3), we prove that the bagged posterior is asymptotically normal (that is, it satisfies a Bernstein–Von Mises theorem) and that bagged posterior credible intervals for the optimal parameter are asymptotically conservative when the bootstrapped datasets are the same size as the original dataset. Moreover, we show that if the size of the bootstrapped datasets is appropriately selected, then the credible intervals have asymptotically correct frequentist coverage. For model selection (Section 4), we show that when multiple models have the same or very similar explanatory power for the true data-generating distribution, bagged model selection assigns more stable and appropriate posterior probabilities to each model. In short, we show that when used for parameter inference, the bagged posterior improves upon the standard posterior by accounting for sampling variance, as in traditional bootstrapping (Efron, 1979), while when used for model selection, the bagged posterior stabilizes model probabilities, in the spirit of bagging (Breiman, 1996; Bühlmann and Yu, 2002).

Next, we validate our theory and model–data mismatch index through simulation experiments in the setting of parameter inference and feature selection for linear regression (Section 5.1). Our results show that the mismatch index is useful for both (a) diagnosing misspecification in the likelihood as well as (b) detecting poorly chosen priors that either make the true parameter extremely unlikely or lead to poorly identified model parameters. Remarkably, we find that BayesBag often produces superior results compared to standard Bayesian inference even when the likelihood model is correct. We then explore the benefits of BayesBag over alternative robust approaches in a hierarchical mixed effects model (Section 5.2). Finally, we apply BayesBag and the mismatch index to real-world data using a variety of models: linear regression model selection, sparse logistic regression, and phylogenetic tree reconstruction (Section 6). On the real-world data, the mismatch index appears to accurately reflect the expected amount of misspecification. Overall, our empirical results demonstrate that in the presence of significant misspecification, the bagged posterior produces more stable inferences, has better predictive accuracy, and selects correct models more often than the standard posterior; meanwhile, when the model is correctly specified, the bagged posterior produces equally good or better results for parameter inference and prediction, while being slightly more conservative for model selection, when compared to the standard posterior. We conclude in Section 7 with a more detailed discussion of related work and possible extensions.

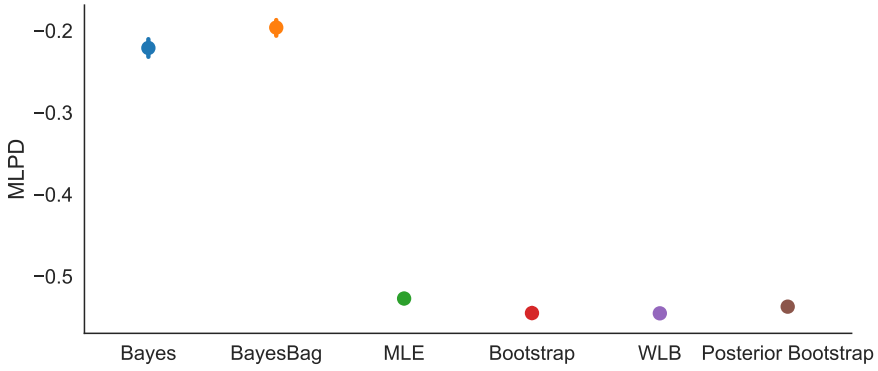


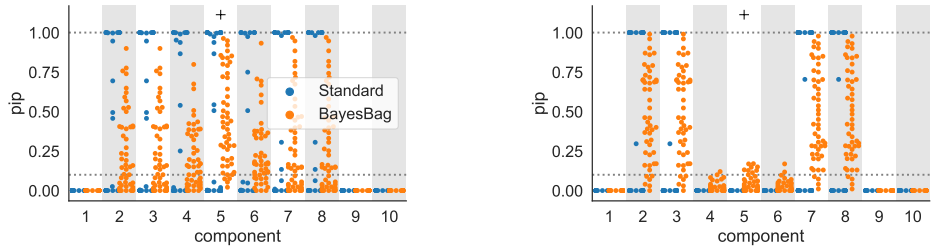
Fig 1: Predictive performance comparison of the standard posterior (Bayes), the bagged posterior (BayesBag), and four other methods based on maximum likelihood estimation (MLE). We show the mean log predictive density (MLPD) of each method on held-out data and 95% confidence intervals. The differences between Bayes, BayesBag, and the MLE-based methods are all statistically significant ($p < 0.0001$, paired t-test)

2. Methodology and motivation. In this section, we highlight the empirical benefits of the bagged posterior compared to the standard posterior, introduce our BayesBag methodology, and provide an overview of the statistical theory that justifies our methodology. Readers wishing to skip the technical details of our theoretical developments can safely proceed to Section 5 after reading this section.

2.1. *Two motivating examples.*

2.1.1. *Parameter inference and prediction.* The bagged posterior tends to more accurately reflect uncertainty given the observed data. As an example in the parameter inference setting, consider a three-level mixed effects logistic regression model inspired by Browne and Draper (2006). We simulated a misspecified data scenario where the data was generated to have correlations among the second-level effects, violating the assumption that they were independent in the model (see Section 5.2 for details). We compared the predictive performance of the standard posterior, the bagged posterior, and a variety of methods based on the maximum likelihood estimation (with the random effects integrated out): the standard MLE, the bootstrapped MLE, the weighted likelihood bootstrap (Newton and Raftery, 1994), and the posterior bootstrap (Lyddon, Walker and Holmes, 2018). The MLE-based methods performed substantially worse than the fully Bayesian methods, and the bagged posterior provided the best predictive performance (Fig. 1).

2.1.2. *Model selection.* As an example in the model selection setting, we applied linear regression to data generated from a nonlinear regression model with $D = 10$ correlated regressors, one of which (component 5) was "causal" (see Section 5.1.3 for details). For $N = 5 \times 10^3$ and 5×10^4, we generated 50 datasets of size N and computed the posterior inclusion probabilities (pips) of each regressor component, allowing at most two regression



(a) $N = 5 \times 10^3$ (b) $N = 5 \times 10^4$

Fig 2: Posterior inclusion probabilities (pips) for the standard posterior and the bagged posterior.

coefficients to be nonzero. Figure 2 summarizes the results. Due to the misspecification and correlated regressors, it does not hold in general (as it does in the well-specified case) that the causal components will be selected. We have set things up so that, by symmetry, components $5 - i$ and $5 + i$ $(i = 0, \ldots, 4)$ are equivalent. As $N \to \infty$, it is optimal to use component 3 (and/or component 7) and component 2 (and/or component 8). The standard pip for either component 3 or 7 is ≈ 1 (with the other ≈ 0) depending on the dataset, and likewise for components 2 and 8, demonstrating that the standard posterior is highly unstable. Meanwhile, the BayesBag pips for components 2, 3, 7, and 8 are roughly uniformly distributed between 0 and 1, thus avoiding false confidence that a particular component will be included or excluded. The standard posterior is also highly unstable as N increases: as more data is added, components with pip ≈ 1 will often flip to ≈ 0, and vice versa. This is illustrated by the pips ≈ 1 for components 4, 5, and 6 when $N = 5 \times 10^3$ that eventually go to zero when $N = 5 \times 10^4$. The BayesBag pips do not exhibit this instability. In Section 6.3, we observe similar behavior on the important real-world problem of phylogenetic tree inference.

2.2. *BayesBag methodology.* Recall from Eq. (2) that we can approximate the bagged posterior by averaging standard posteriors conditioned on each of B bootstrap datasets $x^*_{(1)}, \ldots, x^*_{(B)}$, where $x^*_{(b)} := (x^*_{(b)1}, \ldots, x^*_{(b)M})$ consists of M samples drawn with replacement from the original dataset $x = (x_1, \ldots, x_N)$. For each b, expectations with respect to $\pi(\theta \,|\, x^*_{(b)})$ can be computed by whatever method is most appropriate—for example, a closed-form solution, Markov chain Monte Carlo, or quadrature. In this section, we first discuss the choice of the bootstrap size M and the number of bootstrap datasets B and then we describe how to use BayesBag for model criticism.

2.2.1. *Choosing the bootstrap size for the bagged posterior.* A crucial question for the practical application of BayesBag is how to select the bootstrap dataset size M. We recommend $M = N$ as a good default choice because, as we describe in more detail shortly, it yields asymptotically conservative credible intervals in the parameter inference setting and stabilizes the posterior probabilities of each model in the model selection setting. Alternatively, our theory indicates an optimal choice of M that can be estimated as follows. For

a real-valued function of interest f, let v_N and v_N^* denote, respectively, the standard and bagged posterior variances of $f(\theta)$, where the bagged posterior is computed with $M = N$. Then an estimator for the asymptotically optimal bootstrap sample size for $f(\theta)$ is

$$\hat{M}_{\infty,\text{opt}}(f) := \frac{v_N^*}{v_N^* - v_N} N. \tag{3}$$

If $\hat{M}_{\infty,\text{opt}}(f)$ differs significantly from N, then we suggest recomputing the bagged posterior with $M = \hat{M}_{\infty,\text{opt}}(f)$. Note that since v_N and v_N^* are nonnegative, it follows that $\hat{M}_{\infty,\text{opt}}(f) \in (-\infty, 0) \cup [N, \infty]$. If $\hat{M}_{\infty,\text{opt}}(f) < 0$, then $v_N > v_N^*$, which indicates that using the asymptotically optimal estimate of bootstrap size is not appropriate in this case.

For a set of functions of interest $\mathcal{F}$, we suggest taking the most conservative sample size: $\hat{M}_{\infty,\text{opt}}(\mathcal{F}) := \inf_{f \in \mathcal{F}} \hat{M}_{\infty,\text{opt}}(f)$. In general, $\mathcal{F}$ can be chosen to reflect the quantity or quantities of interest to the ultimate statistical analysis. When $\theta \in \mathbb{R}^D$, two natural choices for the function class 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, hence, we will use the shorthand notation $\hat{M}_{\infty,\text{opt}} := \hat{M}_{\infty,\text{opt}}(\mathcal{F}_{\text{proj}})$.

2.2.2. *Choosing the number of bootstrap datasets for BayesBag.* In addition to choosing M, the other key question for the practical application of BayesBag is how to select the number of bootstrap datasets B. Assume that we can approximate $\pi(\theta \,|\, x_{(b)}^*)$ to high accuracy. Then evaluating the accuracy of the BayesBag approximation given by Eq. (2) reduces to the well-studied problem of estimating the accuracy of a simple Monte Carlo approximation (e.g., Koehler, Brown and Haneuse, 2009). In practice, we have found it sufficient to take $B = 50$ or 100 since the quantities we wish to estimate seem to be fairly low-variance. Thus, we suggest starting with $B = 50$, estimating the Monte Carlo error of any quantities of interest such as parameter means and variances, and then increasing B if the estimated error is unacceptably large. On the other hand, in some scenarios there may be a question of how best to balance the accuracy of a moderate-quality approximation to $\pi(\theta \,|\, x_{(b)}^*)$ (e.g., from a short Markov chain Monte Carlo run) with the number of bootstrap samples B. See Appendix D for a discussion of this computational trade-off.

2.2.3. *Model criticism with BayesBag.* We can also use $\hat{M}_{\infty,\text{opt}}(f)$ for model criticism, which is the task of diagnosing any mismatch between the assumed model and the observed data. Specifically, we define the *model–data mismatch index*

$$\mathcal{I}(f) := \begin{cases} 2N/\hat{M}_{\infty,\text{opt}}(f) - 1 & \text{if } \hat{M}_{\infty,\text{opt}}(f) \in [N, \infty) \\ \mathsf{NA} & \text{otherwise.} \end{cases} \tag{4}$$

This provides a simple, intuitive, and theoretically well-grounded method for measuring the fit of the model to the data. The interpretation of the mismatch index is as follows: $\mathcal{I}(f) \approx 0$ indicates no evidence of mismatch; $\mathcal{I}(f) > 0$ (respectively, $\mathcal{I}(f) < 0$) indicates the standard posterior is overconfident (respectively, under-confident); $\mathcal{I}(f) = \mathsf{NA}$ indicates either that the assumptions required to use $\hat{M}_{\infty,\text{opt}}(f)$ do not hold (e.g., due to multimodality in the posterior or small sample size) or there is a severe model–data mismatch. For a function class $\mathcal{F}$, we define $\mathcal{I}(\mathcal{F})$ by replacing $\hat{M}_{\infty,\text{opt}}(f)$ with $\hat{M}_{\infty,\text{opt}}(\mathcal{F})$ in Eq. (4) and we let $\mathcal{I} := \mathcal{I}(\mathcal{F}_{\text{proj}})$.

2.3. *Justification for BayesBag in the parameter inference setting.* In order to better understand the good empirical performance of BayesBag observed in Section 2.1, to explain why it makes sense to use $M = N$ as a default choice, and to justify the definitions of $\hat{M}_{\infty,\text{opt}}(f)$ and $\mathcal{I}(f)$, we next provide an informal sketch of the asymptotic behavior of the bagged posterior compared to the standard posterior. We cover parameter inference first, then model selection.

2.3.1. *Sketch of statistical theory for parameter inference.* To elucidate the behavior of the bagged posterior, we begin by deriving its mean and covariance in the case where $\theta \in \mathbb{R}^D$. Given data x, let X^* be a random bootstrap dataset and let $\vartheta^* \,|\, X^* \sim \pi(\theta \,|\, X^*)$ be distributed according to the standard posterior given data X^*. (We denote random variables with capital Latin letters, e.g., X rather than x, or "curly" Greek letters, e.g., ϑ rather than θ.) Marginalizing out X^*, we have $\vartheta^* \,|\, x \sim \pi^*(\theta \,|\, x)$. Let $\vartheta \,|\, x \sim \pi(\theta \,|\, x)$ and define $\mu(x) := \mathbb{E}(\vartheta \,|\, x) = \int \theta \, \pi(\theta \,|\, x) \mathrm{d}\theta$ to be the standard posterior mean given x. By the law of total expectation, the mean of the bagged posterior is

$$\mathbb{E}(\vartheta^* \,|\, x) = \mathbb{E}\big\{\mathbb{E}(\vartheta^* \,|\, X^*) \,|\, x\big\} = \mathbb{E}\{\mu(X^*) \,|\, x\} = \frac{1}{N^M} \sum_{x^*} \mu(x^*).$$

By the law of total covariance, the covariance matrix of the bagged posterior is

$$\begin{aligned} \mathrm{Cov}(\vartheta^* \,|\, x) &= \mathbb{E}\big\{ \mathrm{Cov}(\vartheta^* \,|\, X^*) \,|\, x\big\} + \mathrm{Cov}\left\{\mathbb{E}(\vartheta^* \,|\, X^*) \,|\, x\right\} \\ &= \mathbb{E}\{\Sigma(X^*) \,|\, x\} + \mathrm{Cov}\{\mu(X^*) \,|\, x\} \end{aligned} \tag{5}$$

where $\Sigma(x) := \mathrm{Cov}(\vartheta \,|\, x) = \int \{\theta - \mu(x)\}\{\theta - \mu(x)\}^\top \pi(\theta \,|\, x) \mathrm{d}\theta$ is the standard posterior covariance. In this decomposition of $\mathrm{Cov}(\vartheta^* \,|\, x)$, the first term approximates the mean of the posterior covariance matrix under the sampling distribution, and the second term approximates the covariance of the posterior mean under the sampling distribution. Thus, the first term reflects Bayesian model-based uncertainty averaged with respect to frequentist sampling variability, and the second term reflects frequentist sampling-based uncertainty of the Bayesian model-based estimate. To make these concepts more concrete, we consider a simple Gaussian location model.

EXAMPLE 2.1 (BayesBag for the Gaussian location model). Consider an i.i.d. Gaussian location model $x_n \sim \mathcal{N}(\theta, V)$ with known covariance matrix V, and assume a conjugate prior on the mean: $\theta \sim \mathcal{N}(0, V_0)$. Given data $x = (x_1, \ldots, x_N)$, the posterior distribution is $\theta \,|\, x \sim \mathcal{N}(R_N \bar{x}_N, V_N)$, where $\bar{x}_N := N^{-1} \sum_{n=1}^N x_n$, $R_N := (V_0^{-1} V / N + I)^{-1}$, and $V_N := (V_0^{-1} + N V^{-1})^{-1}$. For intuition, one can think of $R_N \approx I$ since $\|R_N - I\| = O(N^{-1})$. Meanwhile, the bagged posterior mean and covariance are

$$\begin{aligned} \mathbb{E}(\vartheta^* \,|\, x) &= \mathbb{E}(R_M \bar{X}^*_M \,|\, x) = R_M \bar{x}_N \\ \mathrm{Cov}(\vartheta^* \,|\, x) &= \mathbb{E}(V_M \,|\, x) + \mathrm{Cov}(R_M \bar{X}^*_M \,|\, x) = V_M + M^{-1} R_M \hat{\Sigma}_N R_M, \end{aligned} \tag{6}$$

where $\hat{\Sigma}_N := N^{-1} \sum_{n=1}^N (x_n - \bar{x}_N)(x_n - \bar{x}_N)^\top$ is the sample covariance. In particular, when $M = N$, these expressions simplify to $\mathbb{E}(\vartheta^* \,|\, x) = \mathbb{E}(\vartheta \,|\, x)$ and $\mathrm{Cov}(\vartheta^* \,|\, x) = \mathrm{Cov}(\vartheta \,|\, x) + N^{-1} R_N \hat{\Sigma}_N R_N$. Unlike the standard posterior, which simply assumes the

data have covariance V, the bagged posterior accounts for the true covariance of the data through the inclusion of the term involving $\hat{\Sigma}_N$. Thus, we see that the bagged posterior covariance is the sum of the Bayesian model-based uncertainty plus the frequentist sampling uncertainty. □

The decomposition of the bagged posterior covariance also gives insight into when we can expect the bagged posterior to be advantageous compared to the standard posterior. Again, the case of the Gaussian location model is instructive.

EXAMPLE 2.2 (Uncertainty calibration in the Gaussian location model). Suppose the data are $X_1, \ldots, X_N$ i.i.d. $\sim P_\circ$, and denote the mean and covariance of a single observation by $\mu_\circ := \mathbb{E}(X_1)$ and $\Sigma_\circ := \mathrm{Cov}(X_1)$, respectively. Assume $\mu_\circ$ and $\Sigma_\circ$ are finite. Then the optimal parameter (in terms of log loss) is $\theta_\circ = \mu_\circ$ and $\theta \,|\, X_{1:N}$ converges in distribution to a point mass at $\mu_\circ$ (see Section 3 for details). For quantifying uncertainty about $\theta_\circ$, the posterior is appropriately calibrated if the posterior covariance is equal to

$$\begin{aligned}\mathbb{E}[(R_N \bar{X}_N - \mu_\circ)(R_N \bar{X}_N - \mu_\circ)^\top] \\ &= N^{-1} R_N \Sigma_\circ R_N + (R_N - I)\mu_\circ \mu_\circ^\top (R_N - I) \\ &= N^{-1}\Sigma_\circ + O(N^{-2}).\end{aligned}$$

If the model is correctly specified, then $\Sigma_\circ = V$. Hence, in this case the posterior is correctly calibrated, as expected. If we choose $M = 2N$ then BayesBag is also correctly calibrated since the covariance of the bagged posterior is then approximately $N^{-1}V = N^{-1}\Sigma_\circ$. On the other hand, if we use the default choice of $M = N$ then the covariance of the bagged posterior is approximately $2N^{-1}V = 2N^{-1}\Sigma_\circ$, overestimating the true uncertainty by only a factor of 2.

If the model is misspecified, the posterior covariance underestimates the true uncertainty unless $\Sigma_\circ \preceq NV_N = VR_N$ (that is, unless $VR_N - \Sigma_\circ$ is positive semidefinite). More generally, when N is sufficiently large, the posterior covariance will underestimate (respectively, overestimate) the true uncertainty about $\theta^\top v$ for some $v \in \mathbb{R}^D$ if any eigenvalue of $V - \Sigma_\circ$ is negative (respectively, positive). Meanwhile, the bagged posterior covariance with $M = N$ is

$$V_N + N^{-1} R_N \hat{\Sigma}_N R_N \approx N^{-1}(V + \Sigma_\circ),$$

so it provides an (asymptotically) conservative uncertainty estimate. In the worst-case scenario of $V \ll \Sigma_\circ$, the standard posterior dramatically underestimates the true uncertainty about $\theta_\circ$ while the bagged posterior is correctly calibrated. □

Returning to the generic decomposition given by Eq. (5), we show in Section 3 that for any sufficiently smooth finite-dimensional parametric model, the covariance of the bagged posterior behaves (qualitatively) similarly to Examples 2.1 and 2.2. In particular, if $X_1, \ldots, X_N$ i.i.d. $\sim P_\circ$, then for $N \to \infty$,

$$\mathrm{Cov}(\vartheta^* \,|\, x) \approx M^{-1} J_\circ^{-1} + M^{-1} J_\circ^{-1} I_\circ J_\circ^{-1}, \tag{7}$$

where $J_\circ^{-1}$ is the "model" covariance (analogous to V) and $J_\circ^{-1} I_\circ J_\circ^{-1}$ is the "sandwich" covariance (analogous to $\Sigma_\circ$). Behavior analogous to that of the Gaussian location model case hold: (1) the (rescaled) sandwich covariance $N^{-1} J_\circ^{-1} I_\circ J_\circ^{-1}$ correctly quantifies the (asymptotic) uncertainty about the optimal parameter $\theta_\circ$; (2) when the model is correctly specified, $J_\circ^{-1} = J_\circ^{-1} I_\circ J_\circ^{-1}$; and (3) when the model is misspecified, typically $J_\circ^{-1} \neq J_\circ^{-1} I_\circ J_\circ^{-1}$.

2.3.2. *Justification of recommended bootstrap size.* Using the results from the previous section, we can fully justify our default recommendation of $M = N$ and our definitions for $\hat{M}_{\infty,\text{opt}}(f)$ and $\mathcal{I}(f)$. Let $\text{id} : \theta \mapsto \theta$ denote the identity function. Since we can redefine θ to be $f(\theta)$, without any loss of generality, we assume that $\theta \in \mathbb{R}$ and consider $\hat{M}_{\infty,\text{opt}} = \hat{M}_{\infty,\text{opt}}(\{\text{id}\}) = \hat{M}_{\infty,\text{opt}}(\text{id})$. Let $\sigma_\circ^2 := J_\circ^{-1}$ denote the model-based asymptotic variance and $s_\circ^2 := J_\circ^{-1} I_\circ J_\circ^{-1}$ denote the sampling-based (sandwich) variance.

Conceptually, the situation is as follows. Bootstrapping with $M = N$ typically increases the variance, and as M grows the variance decreases. We are trying to balance these two tendencies in order to match the frequentist sandwich (co)variance. In particular, the bagged posterior variance needs to be approximately $s_\circ^2/N$ in order to be well-calibrated. If follows from Eq. (7) that the bagged posterior variance when using a bootstrap sample size M is $v_M^* \approx (\sigma_\circ^2 + s_\circ^2)/M$. Setting $(\sigma_\circ^2 + s_\circ^2)/M = s_\circ^2/N$ and solving for M shows that we should choose

$$M = M_{\infty,\text{opt}} := (1 + \sigma_\circ^2/s_\circ^2)N. \tag{8}$$

Thus, if $s_\circ^2 = \sigma_\circ^2$ (i.e., the model variance is correctly specified), then we should choose $M = 2N$; this is in agreement with Example 2.2. If $s_\circ^2 > \sigma_\circ^2$, then the sampling-based term is larger, and we should choose $M \in [N, 2N)$. If $s_\circ^2 < \sigma_\circ^2$, then $M > 2N$ is preferred.

A conservative default would be to choose $M = N$ since this protects against having an over-confident posterior in the presence of misspecification and only over-inflates the posterior variance by a factor of 2 when the model is correct. Alternatively, we can estimate $M_{\infty,\text{opt}}$ using Eq. (8) by plugging in an estimate of $\sigma_\circ^2/s_\circ^2$. To obtain such an estimate, we use the fact that the posterior variance satisfies $v_N \approx \sigma_\circ^2/N$ and the bagged posterior variance satisfies $v_N^* \approx (\sigma_\circ^2 + s_\circ^2)/N$. Combining these two equations and solving, we find that $\sigma_\circ^2/s_\circ^2 \approx v_N/(v_N^* - v_N)$. Plugging this into Eq. (8) yields Eq. (3).

Using the finite-sample covariance expression in Eq. (6) for the bagged posterior under the Gaussian location model, we can also define a finite-sample version of $M_{\infty,\text{opt}}$, denoted $M_{\text{fs,opt}}$. To construct an estimator for $M_{\text{fs,opt}}$, let v_0 denote the prior variance and define the estimators $\hat{\sigma}_\circ^2 := N v_0 v_N/(v_0 - v_N)$ and

$$\hat{s}_\circ^2 := \frac{v_0^2}{(v_0 - v_N)^2}(v_N^* - v_N)N.$$

The estimator for $M_{\text{fs,opt}}$ is given by

$$\hat{M}_{\text{fs,opt}} := \frac{N}{2} + \frac{N\hat{\sigma}_\circ^2}{2\hat{s}_\circ^2} - \frac{\hat{\sigma}_\circ^2}{v_0} + \left\{ \left(\frac{N}{2} + \frac{N\hat{\sigma}_\circ^2}{2\hat{s}_\circ^2} \right)^2 - \frac{N\hat{\sigma}_\circ^2}{v_0} \right\}^{1/2} \tag{9}$$

when the right hand side of Eq. (9) is well-defined and positive; otherwise, we set $\hat{M}_{\text{fs,opt}} = N$ and $\mathcal{I} = \mathsf{NA}$. See Appendix F for the derivation of Eq. (9).

REMARK. Fundamentally, we construct $\hat{M}_{\infty,\text{opt}}$ and $\hat{M}_{\text{fs,opt}}$ using an estimator for the sandwich variance $s_\circ^2$, which may be difficult to accurately estimate. However, note that we can always default to the conservative choice $M = N$ when it is hard to estimate the optimal choice of M. Further, the optimal bootstrap sample size estimators will tend to be effective when $f(\theta)$ is roughly Gaussian-distributed, and since we only need to estimate the sandwich variance for the univariate quantity $f(\theta)$, it is plausible to find roughly Gaussian behavior even with relatively small samples sizes and even if θ is high-dimensional. Thus, the applicability of $\hat{M}_{\infty,\text{opt}}$ and $\hat{M}_{\text{fs,opt}}$ is greater it might first appear.

2.3.3. *Justification of the model–data mismatch index.* We can now explain our definition of the mismatch index $\mathcal{I}$. Let $\hat{M}_{\text{opt}}$ denote either $\hat{M}_{\infty,\text{opt}}$ or $\hat{M}_{\text{fs,opt}}$. When the model is correctly calibrated, we expect $\hat{M}_{\text{opt}} \approx 2N$, which leads to $\mathcal{I} \approx 0$. When the standard posterior is overconfident (respectively, under-confident), we expect $\hat{M}_{\text{opt}} < 2N$ (respectively, $\hat{M}_{\text{opt}} > 2N$), which leads to $\mathcal{I} > 0$ (respectively, $\mathcal{I} < 0$).

To understand why we set $\mathcal{I} = \mathsf{NA}$ when $\hat{M}_{\infty,\text{opt}} < N$, we consider the cases of $\hat{M}_{\infty,\text{opt}}$ and $\hat{M}_{\text{fs,opt}}$ separately. By construction, $\hat{M}_{\infty,\text{opt}} \in [N, \infty)$ unless $v_N^* < v_N$; in which case $\hat{M}_{\infty,\text{opt}} < 0$, which is nonsensical. On the other hand, $\hat{M}_{\text{fs,opt}} < N$ when $\hat{s}_\circ^2/N > 0.5 v_0\{1 + v_0/(2N\hat{\sigma}_\circ^2 + v_0)\}$; in other words, when the (estimated) optimal posterior variance is large relative to the prior variance, which indicates that the assumptions used to construct $\hat{M}_{\text{fs,opt}}$ do not hold (since posterior variance should generally be smaller than prior variance). In either case, $\hat{M}_{\text{opt}} < N$ indicates that either (1) there is severe model–data mismatch or (2) the posterior is multimodal or otherwise far from a Gaussian approximation. Hence, we choose to set $\mathcal{I} = \mathsf{NA}$ when $\hat{M}_{\text{opt}} < N$.

2.4. *Justification for BayesBag in the model selection setting.* In model selection, instead of a continuous parameter θ, we have a finite or countable set of models $\mathfrak{M}$. The posterior probability of a model $\mathfrak{m} \in \mathfrak{M}$ is $Q(\mathfrak{m} \,|\, x) \propto p(x \,|\, \mathfrak{m}) Q_0(\mathfrak{m})$, where $x = (x_1, \ldots, x_N)$, $p(x \,|\, \mathfrak{m})$ is the marginal likelihood, and $Q_0(\mathfrak{m})$ is the prior probability. In the notation of Eq. (1), the bagged posterior for model $\mathfrak{m}$ is

$$Q^*(\mathfrak{m} \,|\, x) := \frac{1}{N^M} \sum_{x^*} Q(\mathfrak{m} \,|\, x^*). \tag{10}$$

2.4.1. *Sketch of statistical theory for model selection.* As first noted in Berk (1966), when there are two or more models that explain the data equally well, the posterior typically does not converge on a single model. For instance, consider the case of two models, $\mathfrak{M} = \{1, 2\}$, and suppose $X = (X_1, \ldots, X_N)$ where $X_1, X_2, \ldots$ are i.i.d. For distinct misspecified models, if $\lim_{N\to\infty} \mathbb{E}\{\log p(X \,|\, 1) - \log p(X \,|\, 2)\} = 0$, then the posterior mass on model 1 converges in distribution to a $\mathsf{Bern}(1/2)$ random variable (Yang and Zhu, 2018):

$$Q(1 \,|\, X) \xrightarrow[N\to\infty]{\mathcal{D}} \mathsf{Bern}(1/2). \tag{11}$$

Since both models provide equally good approximations of the true data-generating distribution, the ideal outcome would be $Q(1 \,|\, X) = Q(2 \,|\, X) = 1/2$, but Eq. (11) describes the opposite behavior: a single model has posterior probability 1.

Meanwhile, BayesBag model selection does not exhibit this pathological behavior. Rather, as we show in Theorem 4.1, the bootstrap resampling stabilizes the model probabilities such that when $M = N$, the bagged posterior mass on model 1 converges in distribution to a $\mathsf{Unif}(0, 1)$ random variable:

$$Q^*(1 \mid X) \xrightarrow[N\to\infty]{\mathcal{D}} \mathsf{Unif}(0, 1).$$

Moreover, if we choose M such that $M = o(N)$, then the bagged posterior mass on model 1 has the ideal behavior of converging to 1/2:

$$Q^*(1 \mid X) \xrightarrow[N\to\infty]{\mathcal{D}} 1/2.$$

2.4.2. *Justification of recommended bootstrap size for model selection.* In practice, it seems implausible that two models would explain the true data-generating distribution *exactly* equally well. However, it turns out that even if model 1 dominates model 2 asymptotically, for a finite sample size it can happen that $N^{-1}\mathbb{E}\{\log p(X \mid 1) - \log p(X \mid 2)\} \approx 0$, such that model 2 dominates model 1 roughly half of the time. The analysis of Yang and Zhu (2018) was motivated by widespread observations of this type of phenomena in Bayesian phylogenetic tree reconstruction (Alfaro, Zoller and Lutzoni, 2003; Douady et al., 2003; Wilcox et al., 2002), though it certainly occurs more generally (Meng and Dunson, 2019) such as in economic modeling (Oelrich et al., 2020).

Thus, setting $M = N$ appears to be a good default choice that will still behave fairly well in the worst case. If large amounts of data are available and there is reason to believe that many models have similar expected log-likelihoods, a choice such as $M = \lceil N/\log_{10}(N) \rceil$ or $M = \lceil cN \rceil$ for a moderate value of c such at $1/4$ may be advisable.

3. Theory of parameter inference with BayesBag. In this section, we formally present our general results on BayesBag in the parameter inference setting. This section formalizes and generalizes the results sketched out in Section 2.3. Our main result (Theorem 3.2) is a Bernstein–Von Mises theorem for the bagged posterior under sufficiently regular finite-dimensional models. In particular, we show that while the standard Bayesian posterior may be arbitrarily under- or over-confident when the model is misspecified, the bagged posterior avoids over-confident uncertainty quantification by accounting for sampling variability. Since Theorem 3.2 is asymptotic in nature, it ignores the potentially significant finite-sample benefits of both the bootstrap and the posterior, neither of which requires the normality assumptions of our asymptotic analysis. Nevertheless, the theorem offers valuable statistical justification for BayesBag in general and for the use of $\hat{M}_{\infty,\text{opt}}$ and $\mathcal{I}$ in particular. Theorem 3.1 is a simpler version of the same result for the univariate Gaussian location model, in which case the statement and proof of the theorem are much more transparent.

3.1. *Background.* We now restate more precisely the setting that was informally introduced in Sections 1 and 2. Consider a model $\{P_\theta : \theta \in \Theta\}$ for independent and identically distributed (i.i.d.) data $x_1, \ldots, x_N$, where $x_n \in \mathbb{X}$ and $\Theta \subset \mathbb{R}^D$. Denote $x_{1:N} =$

$(x_1, \ldots, x_N)$ and suppose p_θ is the density of P_θ with respect to some reference measure. The standard Bayesian posterior given $x_{1:N}$ is

$$\Pi(\mathrm{d}\theta \,|\, x_{1:N}) := \frac{\prod_{n=1}^N p_\theta(x_n)}{p(x_{1:N})} \Pi_0(\mathrm{d}\theta),$$

where $\Pi_0(\mathrm{d}\theta)$ is the prior distribution and $p(x_{1:N}) := \int \{\prod_{n=1}^N p_\theta(x_n)\} \Pi_0(\mathrm{d}\theta)$ is the marginal likelihood. When convenient, we will use the shorthand notation $\Pi_N := \Pi(\cdot \,|\, x_{1:N})$.

Assume the observed data $X_1, \ldots, X_N$ is actually generated i.i.d. from some unknown distribution $P_\circ$. Suppose there is a unique parameter $\theta_\circ$ that minimizes the Kullback–Leibler divergence from $P_\circ$ to the model, or equivalently, $\theta_\circ = \arg\max_{\theta \in \Theta} \mathbb{E}\{\log p_\theta(X_1)\}$. Denote the log-likelihood by $\ell_\theta := \log p_\theta$, its gradient by $\dot{\ell}_\theta := \nabla_\theta \ell_\theta$, and its Hessian by $\ddot{\ell}_\theta := \nabla^2_\theta \ell_\theta$. Furthermore, define the information matrices $J_\theta := -\mathbb{E}\{\ddot{\ell}_\theta(X_1)\}$ and $I_\theta := \mathrm{Cov}\{\dot{\ell}_\theta(X_1)\}$.

Let $\hat{\theta}_N := \arg\max_\theta \prod_{n=1}^N p_\theta(X_n)$ denote the maximum likelihood estimator. Under regularity conditions, $\hat{\theta}_N$ is asymptotically normal in the sense that

$$N^{1/2}(\hat{\theta}_N - \theta_\circ) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_{\theta_\circ}^{-1} I_{\theta_\circ} J_{\theta_\circ}^{-1}), \tag{12}$$

where $J_{\theta_\circ}^{-1} I_{\theta_\circ} J_{\theta_\circ}^{-1}$ is known as the sandwich covariance (White, 1982). Under mild conditions, the Bernstein–Von Mises theorem (Kleijn and van der Vaart, 2012; van der Vaart, 1998, Ch. 10) guarantees that for $\vartheta \sim \Pi_N$,

$$N^{1/2}(\vartheta - \hat{\theta}_N) \,|\, X_{1:N} \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_{\theta_\circ}^{-1}). \tag{13}$$

Hence, the standard posterior is correctly calibrated, asymptotically, if the covariance matrices of the Gaussian distributions in Eqs. (12) and (13) coincide – that is, if $J_{\theta_\circ}^{-1} I_{\theta_\circ} J_{\theta_\circ}^{-1} = J_{\theta_\circ}^{-1}$, or equivalently, $I_{\theta_\circ} = J_{\theta_\circ}$. In particular, if $I_{\theta_\circ} = J_{\theta_\circ}$, then Bayesian credible sets are (asymptotically) valid confidence sets in the frequentist sense: sets of posterior probability $1 - \alpha$ contain the true parameter with $P_\circ^\infty$-probability $1 - \alpha$, under mild conditions.

If the model is well-specified, that is, if $P_\circ = P_{\theta_\dagger}$ for some parameter $\theta_\dagger \in \Theta$ (and thus $\theta_\circ = \theta_\dagger$ by the uniqueness assumption), then $I_{\theta_\circ} = J_{\theta_\circ}$ under very mild conditions. On the other hand, if the model is misspecified — that is, if $P_\circ \neq P_\theta$ for all $\theta \in \Theta$ — then although Eq. (13) still holds, typically $I_{\theta_\circ} \neq J_{\theta_\circ}$. If $I_{\theta_\circ} \neq J_{\theta_\circ}$, then the standard posterior is not correctly calibrated, and in fact, asymptotic Bayesian credible sets may be arbitrarily over- or under-confident.

Although Eqs. (12) and (13) are only asymptotic, we should take little comfort that the non-asymptotic situation will somehow be better. We have already seen in Example 2.2 that, no matter the sample size, the standard posterior for the Gaussian location model can be badly miscalibrated. The prior may help to down-weight *a priori* unlikely hypotheses, but it cannot account for misspecification in the likelihoods among "reasonable" parameter values.

3.2. *BayesBag for parameter inference.* Let $X^*_{1:M}$ denote a bootstrapped copy of $X_{1:N}$ with M observations; in other words, each observation X_n is replicated K_n times in $X^*_{1:M}$, where $K_{1:N} \sim \mathsf{Multi}(M, 1/N)$ is a multinomial-distributed count vector of length N. We define the *bagged posterior* $\Pi^*(\cdot \,|\, X_{1:N})$ by setting

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

for all measurable $A \subseteq \Theta$. (Note that this is equivalent to the informal definition in Eq. (1).) In order words, BayesBag uses bootstrapping to average posteriors over approximate realizations of data from the true data-generating distribution. Depending on M, the bagged posterior may be more diffuse or less diffuse than the standard posterior. To avoid notational clutter, we suppress the dependence of $\Pi^*(\cdot \,|\, X_{1:N})$ on M.

We will typically use the shorthand notation $\Pi^*_N := \Pi^*(\cdot \,|\, X_{1:N})$ and we let $\vartheta^* \,|\, X_{1:N} \sim \Pi^*_N$ denote a random variable distributed according to the bagged posterior. We assume Θ is an open subset of $\mathbb{R}^D$ and we write $\mathrm{d}\theta$ to denote Lebesgue measure on Θ. Further, we assume Π_N and Π^*_N have densities π_N and π^*_N, respectively, with respect to Lebesgue measure. Note that π^*_N exists if π_N exists.

3.3. *Asymptotic normality of BayesBag for the Gaussian location model.* Before developing a general Bernstein–Von Mises theorem for BayesBag, as a warmup we prove Theorem 3.1, a simpler version of the result in the case of the Gaussian location model from Examples 2.1 and 2.2. Although it is a special case of Theorem 3.2, the statement and proof of Theorem 3.1 are much easier to follow and it still captures the essence of the general result. For maximal clarity, we consider the case of univariate data.

THEOREM 3.1. *Let $X_1, X_2, \ldots \in \mathbb{R}$ i.i.d. such that $\mathbb{E}(|X_1|^3) < \infty$. Let $\vartheta^* \,|\, X_{1:N} \sim \Pi^*_N$ and suppose $c := \lim_{N\to\infty} M/N \in (0, \infty)$ for $M = M(N)$. Then for almost every $(X_1, X_2, \ldots)$,*

$$N^{1/2}\big\{\vartheta^* - \mathbb{E}(\vartheta^* \,|\, X_{1:N})\big\} \,|\, X_{1:N} \xrightarrow{\mathcal{D}} \mathcal{N}(0,\, V/c + \mathrm{Var}(X_1)/c). \tag{14}$$

In other words, with probability 1, the bagged posterior converges weakly to $\mathcal{N}(0,\, V/c + \mathrm{Var}(X_1)/c)$ after centering at its mean and scaling by $N^{1/2}$. The proof of Theorem 3.1 is in Appendix G.1.

3.4. *Bernstein–Von Mises theorem for BayesBag.* We now turn to the main result of this section: a general Bernstein–Von Mises theorem for BayesBag. Recall that if the standard posterior were asymptotically correctly calibrated, it would have asymptotic covariance $J^{-1}_{\theta_\circ} I_{\theta_\circ} J^{-1}_{\theta_\circ}$ (the sandwich covariance), whereas in fact it has asymptotic covariance $J^{-1}_{\theta_\circ}$. Letting $c := \lim_{N\to\infty} M/N$ as in Theorem 3.1, we show that the asymptotic covariance of the bagged posterior is $J^{-1}_{\theta_\circ}/c + J^{-1}_{\theta_\circ} I_{\theta_\circ} J^{-1}_{\theta_\circ}/c$, which mirrors the form of Eq. (14) but is much more general. Our technical assumptions are essentially the same as those used by Kleijn and van der Vaart (2012) to prove the Bernstein–Von Mises theorem under misspecification for the standard posterior.

For a measure ν and function f, we will make use of the shorthand $\nu(f) := \int f \mathrm{d}\nu$. Let $X_{1:\infty}$ denote the infinite sequence of data $(X_1, X_2, \ldots)$.

THEOREM 3.2. *Suppose $X_1, X_2, \ldots$ i.i.d. $\sim P_\circ$ and assume that:*

(i) *$\theta \mapsto \ell_\theta(X_1)$ is differentiable at $\theta_\circ$ in probability;*
(ii) *there is an open neighborhood U of $\theta_\circ$ and a function $m_{\theta_\circ} : \mathbb{X} \to \mathbb{R}$ such that $P_\circ(m_{\theta_\circ}^3) < \infty$ and for all $\theta, \theta' \in U$, $|\ell_\theta - \ell_{\theta'}| \leq m_{\theta_\circ} \|\theta - \theta'\|_2$ a.s.$[P_\circ]$;*
(iii) *$-P_\circ(\ell_\theta - \ell_{\theta_\circ}) = \frac{1}{2}(\theta - \theta_\circ)^\top J_{\theta_\circ}(\theta - \theta_\circ) + o(\|\theta - \theta_\circ\|_2^2)$ as $\theta \to \theta_\circ$;*
(iv) *$J_{\theta_\circ}$ is an invertible matrix;*
(v) *letting $\vartheta^* \sim \Pi_N^*$, it holds that conditionally on $X_{1:\infty}$, for almost every $X_{1:\infty}$, for every sequence of constants $C_N \to \infty$,*

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

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

Then, letting $\vartheta^* \sim \Pi_N^*$, we have that conditionally on $X_{1:\infty}$, for almost every $X_{1:\infty}$,

$$N^{1/2}(\vartheta^* - \theta_\circ) - \Delta_N \,|\, X_{1:N} \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_{\theta_\circ}^{-1}/c + J_{\theta_\circ}^{-1} I_{\theta_\circ} J_{\theta_\circ}^{-1}/c),$$

where $\Delta_N := N^{1/2} J_{\theta_\circ}^{-1}(\mathbb{P}_N - P_\circ)\dot{\ell}_{\theta_\circ}$ and $\mathbb{P}_N := N^{-1}\sum_{n=1}^N \delta_{X_n}$.

The proof of Theorem 3.2 is in Appendix G.2. To interpret this result, it is helpful to compare it to the behavior of the standard posterior. Under the conditions of Theorem 3.2, letting $\vartheta \sim \Pi_N$, then almost surely $N^{1/2}(\vartheta - \theta_\circ) - \Delta_N \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_{\theta_\circ}^{-1})$ by Kleijn and van der Vaart (2012, Theorem 2.1 and Lemma 2.1). Thus, the bagged posterior and the standard posterior for $N^{1/2}(\theta - \theta_\circ)$ have the same asymptotic mean, Δ_N, but the bagged posterior has asymptotic covariance $J_{\theta_\circ}^{-1}/c + J_{\theta_\circ}^{-1} I_{\theta_\circ} J_{\theta_\circ}^{-1}/c$ instead of $J_{\theta_\circ}^{-1}$.

3.5. *Extensions.* There are many possible extensions to Theorem 3.2.

Regression models. Our main results in this section as well as the next section on model selection (that is, Theorem 3.2, Theorem 4.1, and Corollary 4.2) apply equally well to the regression setting with random regressors where the data take the form $X_n = (Y_n, Z_n)$ and the models $p_\theta(y \,|\, z)$ are conditional.

Alternative bootstrap methods. Much of bootstrap theory extends beyond the multinomial distribution for $K_{1:N}$ to other distributions such as those where $K_1, \ldots, K_N$ are weakly correlated random variables with mean 1 and variance 1 (van der Vaart and Wellner, 1996). Thus, we conjecture that Theorem 3.2 also holds when we use the bagged posterior proportional to $\pi_0(\theta)\prod_{n=1}^N p_\theta(X_n)^{K_n}$, where $K_1, \ldots, K_N$ are independent nonnegative random variables satisfying $\mathbb{E}(K_n) = 1$ and $\mathrm{Var}(K_n) = 1$ $(n = 1, \ldots, N)$.

Dependent observations. We have also focused on the case of independent observations $X_1, \ldots, X_N$, but it is feasible to extend our theory and methodology to more complex models. One possibility is to draw on the rich existing bootstrap literature for time series and spatial models (Künsch, 1989; Peligrad, 1998). Alternatively, a model-based bootstrap approach could be used by employing a nonparametric or rich parametric model to approximate $P_\circ$. We leave investigation of these extensions for future work.

4. Theory of model selection with BayesBag. The issues with Bayesian model selection, which concerns a countable collection of models, are quite different from the challenges with Bayesian parameter inference, where we are dealing with continuous parameter space. The central issue we will be concerned with is that posterior model probabilities can be extremely unstable when there is model misspecification. In the spirit of bagging, BayesBag is able to stabilize the posterior model probabilities, thus improving reproducibility.

In Bayesian model selection, we have a countable set of models $\mathfrak{M}$. A model $\mathfrak{m} \in \mathfrak{M}$ has prior probability $Q_0(\mathfrak{m})$ 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 now an element of a model-specific parameter space with prior distribution $\Pi_0(\mathrm{d}\theta_\mathfrak{m} \,|\, \mathfrak{m})$. The posterior probability of $\mathfrak{m} \in \mathfrak{M}$ is $Q(\mathfrak{m} \,|\, X_{1:N}) \propto p(X_{1:N} \,|\, \mathfrak{m}) Q_0(\mathfrak{m})$.

Recall from Eq. (10) that the bagged posterior probability of model $\mathfrak{m} \in \mathfrak{M}$ is

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

We develop our asymptotic theory in the case of two models, $\mathfrak{M} = \{1, 2\}$. For the moment, we assume each model contains a single parameter value, that is, $|\Theta_\mathfrak{m}| = 1$, but we allow the observation model $p_N(X_n \,|\, \mathfrak{m})$ to depend on the number of observations, so that $p(X_{1:N} \,|\, \mathfrak{m}) = \prod_{n=1}^{N} p_N(X_n \,|\, \mathfrak{m})$. (We generalize to the case of nondegenerate parameter spaces $\Theta_\mathfrak{m}$ in Corollary 4.2.) Let $Z_N := \log p(X_{1:N} \,|\, 1) - \log p(X_{1:N} \,|\, 2)$ and $Z_{Nn} := \log p_N(X_n \,|\, 1) - \log p_N(X_n \,|\, 2)$ for $n = 1, \ldots, N$. Assume the data $X_1, X_2, \ldots$ are i.i.d. from some unknown distribution $P_\circ$.

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 $\lim_{N\to\infty} N^{1/2}\mathbb{E}(Z_{Nn}) = \mu_\infty \in \mathbb{R}$ while the variance remains fixed: $\mathrm{Var}(Z_{Nn}) = \sigma^2_\infty$. Thus, $\mathbb{E}(Z_N) \approx N^{1/2}\mu_\infty$ and $\mathrm{Std}(Z_N) = N^{1/2}\sigma_\infty$ when N is large, so whenever $\mu_\infty \neq 0$ the deviation of $\mathbb{E}(Z_N)$ from zero remains nontrivial relative to $\mathrm{Std}(Z_N)$—even in the asymptotic regime. The effect size $\delta_\infty := \mu_\infty/\sigma_\infty$ quantifies the amount of evidence in favor of model 1. If $\delta_\infty > 0$, then model 1 is favored, whereas model 2 is favored if $\delta_\infty < 0$.

Our next result, which is similar in spirit to the bagging result of Bühlmann and Yu (2002, Proposition 2.1), shows that (1) the posterior probability of model 1 converges to a Bernoulli random variable with parameter depending on δ_∞ and (2) when $M = \Theta(N)$, the bagged posterior probability of model 1 converges to a continuous random variable on $[0, 1]$ with a distribution that depends on δ_∞. Hence, in the context of model selection, BayesBag yields more stable and reproducible inferences than the standard posterior. Let $\Phi(t)$ denote the cumulative distribution function of the standard normal distribution.

THEOREM 4.1. *Let $X_1, X_2, \ldots$ i.i.d. $\sim P_\circ$ for some distribution $P_\circ$ and define $Z_{Nn} := \log p_N(X_n \,|\, 1) - \log p_N(X_n \,|\, 2)$. If*

(i) $\lim_{N\to\infty} N^{1/2}\mathbb{E}(Z_{Nn}) = \mu_\infty \in \mathbb{R}$,
(ii) $\mathrm{Var}(Z_{Nn}) = \sigma^2_\infty \in (0, \infty)$ *for all* N,
(iii) $\limsup_{N\to\infty} \mathbb{E}(|Z_{Nn}|^{2+\varepsilon}) < \infty$ *for some* $\varepsilon > 0$, *and*

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

then

1. *for the standard posterior,* $Q(1 \mid X_{1:N}) \xrightarrow{\mathcal{D}} U \sim \mathsf{Bern}(\Phi(\mu_\infty/\sigma_\infty))$;
2. *for the bagged posterior, if* $c > 0$, *then*

$$Q^*(1 \mid X_{1:N}) \xrightarrow{\mathcal{D}} U^*$$

where U^* *is a random variable on* $[0,1]$ *with probability density* $f(u) = \Phi'\big(c^{-1/2}\Phi^{-1}(u) - \mu_\infty/\sigma_\infty\big)c^{-1/2}/\Phi'(\Phi^{-1}(u))$ *for* $u \in (0,1)$; *and*

3. *for the bagged posterior, if* $c = 0$, *then*

$$Q^*(1 \mid X_{1:N}) \xrightarrow{P} 1/2.$$

In particular, if $\mu_\infty = 0$ *and* $c > 0$, *then we have* $Q(1 \mid X_{1:N}) \xrightarrow{\mathcal{D}} \mathsf{Bern}(1/2)$ *and* $Q^*(1 \mid X_{1:N}) \xrightarrow{\mathcal{D}} \mathsf{Unif}(0,1)$.

The proof is in Appendix G.3. Fig. 3 illustrates how Theorem 4.1 establishes the greater stability of BayesBag versus standard Bayes for model selection. Even for effect sizes $\delta_\infty > 1$, which should strongly favor model 1, the standard posterior overwhelmingly favors model 2 with non-negligible probability — that is, $\mathbb{P}(Q(1 \mid X_{1:N}) \approx 0)$ is non-negligible. Meanwhile, the probability that the bagged posterior strongly favors model 2 goes to zero rapidly as δ_∞ increases — that is, $\mathbb{P}(Q^*(1 \mid X_{1:N}) \approx 0) \to 0$ rapidly as δ_∞ grows. For example, when $\delta_\infty = 2$ and $c = 1$, $\mathbb{P}(U = 0) > 0.02$ whereas $\mathbb{P}(U^* < 0.1) < 7 \times 10^{-5}$. Thus, in this example, in 1 out of 50 experiments the standard posterior will overwhelmingly favor the "wrong" model, whereas BayesBag will somewhat strongly favor the wrong model in only around 7 out of 100,000 experiments.

In Corollary 4.2, we extend Theorem 4.1 to nondegenerate parameter spaces $\Theta_1 \subset \mathbb{R}^{D_1}$ and $\Theta_2 \subset \mathbb{R}^{D_2}$. To avoid tedious arguments, we only consider the case where $\mu_\infty = 0$. For $\mathfrak{m} \in \mathfrak{M}$, define $\ell_{\mathfrak{m},\theta_\mathfrak{m}}(X_n) := \log p_{\theta_\mathfrak{m}}(X_n \mid \mathfrak{m})$ and denote the optimal parameter by $\theta_{\mathfrak{m}\circ} := \arg\max_{\theta_\mathfrak{m} \in \Theta_\mathfrak{m}} \mathbb{E}\{\ell_{\mathfrak{m},\theta_\mathfrak{m}}(X_1)\}$.

For arbitrary data x, let $\Lambda_x := \log p(x \mid 1)Q_0(1) - \log p(x \mid 2)Q_0(2)$. We will assume that conditionally on $(X_1, X_2, \ldots)$, for almost every $(X_1, X_2, \ldots)$,

$$\Lambda_{X^*_{1:M}} = \frac{1}{2}(D_2 - D_1)\log N + \sum_{m=1}^{M} \log \frac{p_{\theta_{1\circ}}(X^*_m \mid 1)}{p_{\theta_{2\circ}}(X^*_m \mid 2)} + O_{P^+}(1), \tag{15}$$

where $X^*_{1:M}$ is bootstrapped from $X_{1:N}$ and $O_{P^+}(1)$ denotes a (random) quantity which is bounded in (outer) probability. Eq. (15) holds when $X^*_{1:M}$ is replaced by $X_{1:N}$, under standard regularity assumptions (Clarke and Barron, 1990). Thus, we expect Eq. (15) to hold under similar but slightly stronger conditions, since we must consider a triangular array rather than a sequence of random variables.

COROLLARY 4.2. *Let* $X_1, X_2, \ldots$ *i.i.d.* $\sim P_\circ$ *and assume the regularity conditions in Theorem 3.2 hold for both models* 1 *and* 2. *Further assume that Eq.* (15) *holds,* $\mathbb{E}\{\ell_{1,\theta_{1\circ}}(X_1) - \ell_{2,\theta_{2\circ}}(X_1)\} = 0$, *and* $\mathrm{Var}\{\ell_{1,\theta_{1\circ}}(X_1) - \ell_{2,\theta_{2\circ}}(X_1)\} \in (0,\infty)$. *Then the conclusions of Theorem 4.1 apply.*

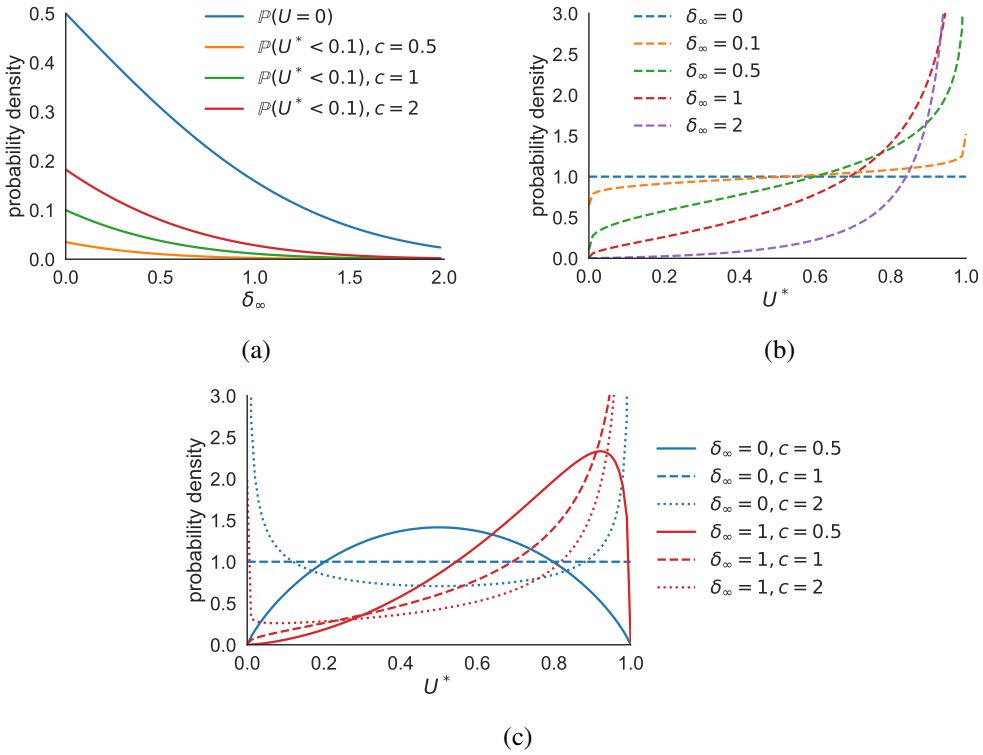


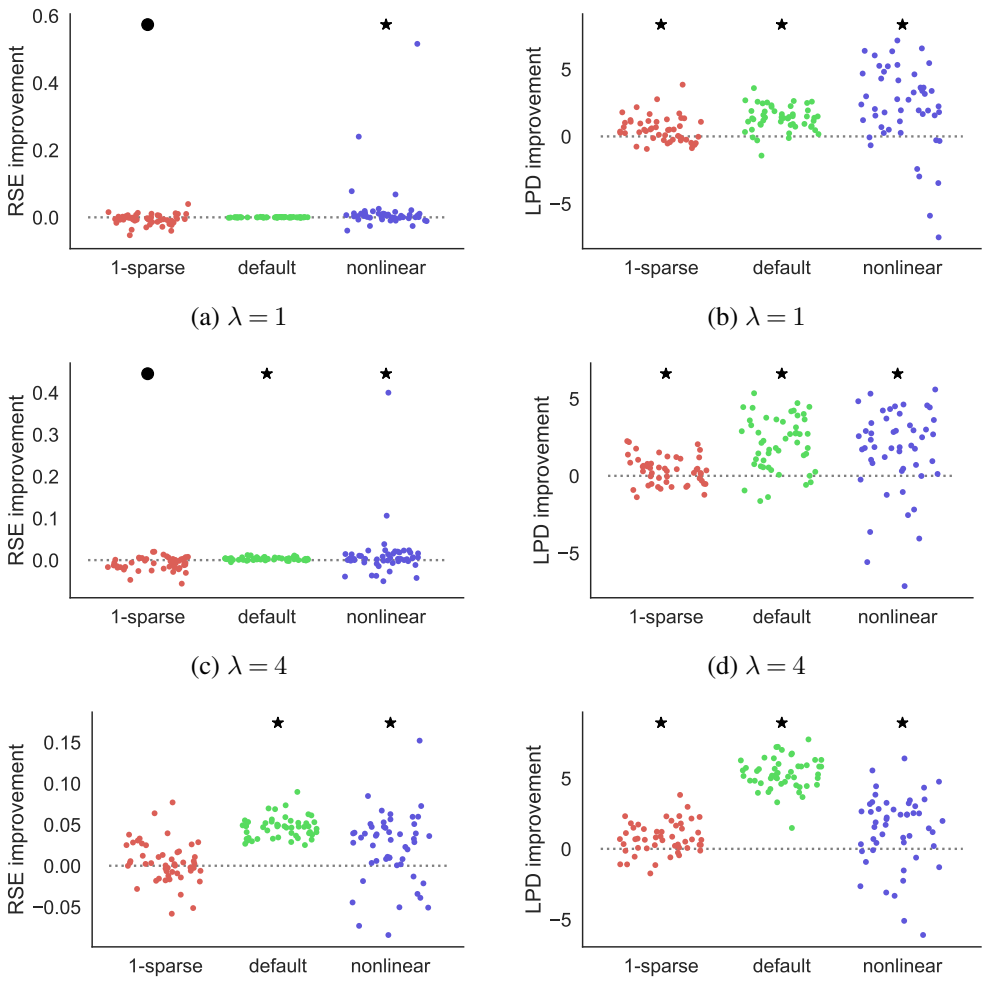
Fig 3: Asymptotic distribution of posterior probability of model 1 under the standard posterior (U) and bagged posterior (U^*). Larger values of the effect size $\delta_\infty = \mu_\infty/\sigma_\infty$ indicate stronger evidence for model 1. **(A)** Probabilities that $U = 0$ and $U^* < 0.1$ as a function of δ_∞. **(B)** Densities of U^* for a range of δ_∞ values, with $c = 1$. **(C)** Densities of U^* as δ_∞ and c vary.

The proof is in Appendix G.4.

5. Simulation studies.

5.1. *Linear regression.* We performed an extensive collection of simulations to assess the performance of BayesBag in the setting of linear regression. Linear regression is an ideal model for investigating the properties of BayesBag and the usefulness of the mismatch index $\mathcal{I}$ and optimal bootstrap sample size estimator $\hat{M}_{\text{opt}}$, since all computations of posterior quantities can be done in closed form, yet it is a rich enough model that we can explore many kinds of model–data mismatch. For linear regression, the data consist of regressors $Z_n \in \mathbb{R}^D$ and observations $Y_n \in \mathbb{R}$ $(n = 1, \ldots, N)$ while the parameter is $\theta = (\theta_0, \ldots, \theta_D) = (\log \sigma^2, \beta_1, \ldots, \beta_D) \in \mathbb{R}^{D+1}$. Assuming conjugate priors, the generative model is

$$\sigma^2 \sim \Gamma^{-1}(a_0, b_0)$$

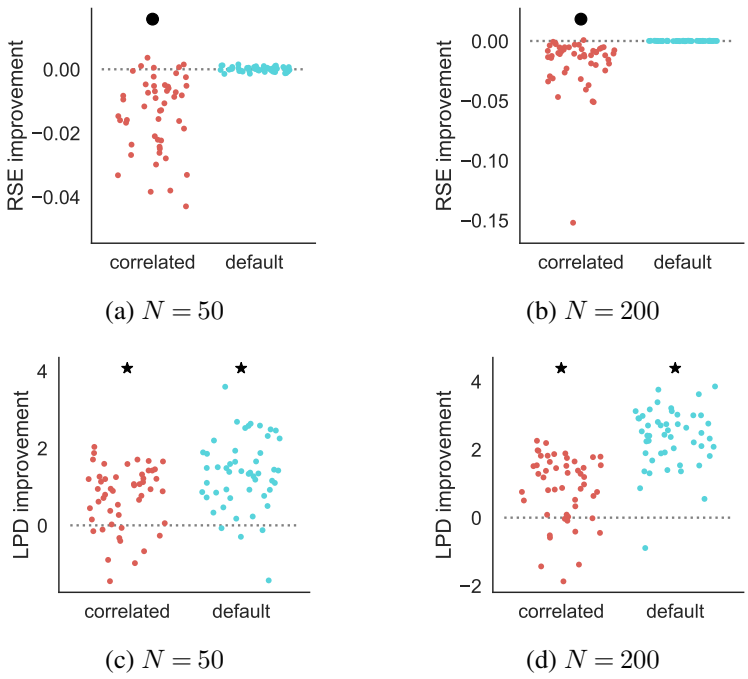


(a) $\lambda = 1$ (b) $\lambda = 1$

(c) $\lambda = 4$ (d) $\lambda = 4$

(e) $\lambda = 16$ (f) $\lambda = 16$

Fig 4: Parameter inference performance on default, 1-sparse, and nonlinear data for $\lambda \in \{1, 4, 16\}$. A positive improvement value indicates that the bagged posterior outperformed the standard posterior on that dataset. Scenarios marked with a $\star$ (respectively, $\bullet$) exhibited a statistically significant difference in the positive (respectively, negative) direction ($p < 0.05$, two-sided Wilcoxon signed-rank test). RSE = relative squared error of $\beta_\circ$. LPD = log posterior density at $\beta_\circ$.



(a) $N = 50$ (b) $N = 200$

(c) $N = 50$ (d) $N = 200$

Fig 5: Parameter inference performance on default and correlated data for $N \in \{50, 200\}$. See caption for Fig. 4 for further explanation.

$$\beta_d \,|\, \sigma^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2/\lambda) \qquad d = 1, \ldots, D,$$

$$Y_n \,|\, Z_n, \beta, \sigma^2 \overset{\text{indep}}{\sim} \mathcal{N}(Z_n^\top \beta, \sigma^2) \qquad n = 1, \ldots, N,$$

where a_0, b_0, and λ are hyperparameters that will be specified later. We simulated 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

$$Y_n = f(Z_n)^\top \beta_\dagger + \epsilon_n \tag{16}$$

for $n = 1, \ldots, N$, where we used two settings for each of the regressor distribution G, the regression function f, and the coefficient vector $\beta_\dagger \in \mathbb{R}^D$.

- **Regressor distribution** G**.** By default, we used $G = \mathcal{N}(0, I)$; we refer to this as the uncorrelated setting. Alternatively, we used a correlated setting, where, for $h = 10$, $Z \sim G$ was defined by generating $\xi \sim \chi^2(h)$ and then $Z \,|\, \xi \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{dd'} = \exp\{-(d - d')^2/64\}/(\xi_d \xi_{d'})$ and $\xi_d = \sqrt{\xi/(h-2)}^{\mathbb{1}(d \text{ is odd})}$. The motivation for the correlated sampling procedure was 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 standard normal.

- **Regression function f.** By default, we used a linear regression function $f(z) = z$. Alternatively, we used the nonlinear function $f(z) = (z_1^3, \ldots, z_D^3)^\top$.
- **Coefficient vector $\beta_\dagger$.** By default, we used a dense vector with $\beta_{\dagger d} = 2^{(5-d)/2}$ for $d = 1, \ldots, D$. Alternatively, we used a k-sparse vector with $\beta_{\dagger d} = 1$ if $d \in \{\lfloor j(D + \frac{1}{2})/(k + 1)\rfloor \,|\, j = 1, \ldots, k\}$ and $\beta_{\dagger d} = 0$ otherwise.

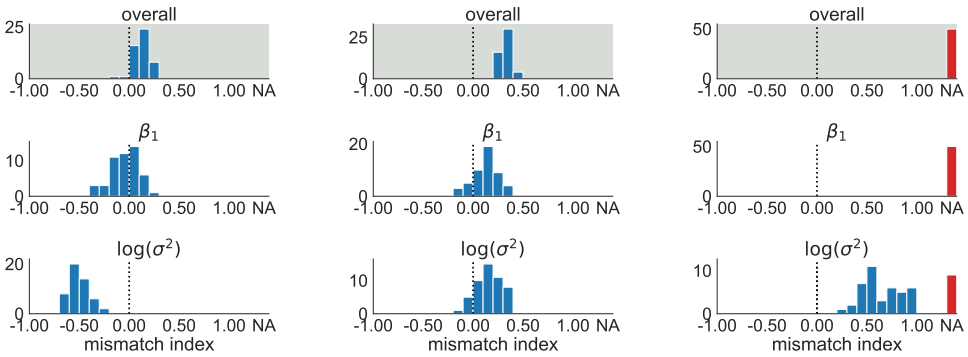
For brevity, we omit the default setting labels when indicating which settings of G, f, and $\beta_\dagger$ were used in each dataset. For example, we abbreviate uncorrelated-nonlinear-2-sparse as nonlinear-2-sparse, and correlated-linear-dense as correlated. We refer to the dataset with all three defaults as default.

Throughout this section, unless stated otherwise, the data were generated with $D = 10$ and $N = 50$, and the model hyperparameters were set to $a_0 = 2$, $b_0 = 1$, and $\lambda = 1$. Each experimental setting was replicated 50 times. BayesBag was run with $M = \hat{M}_{\text{fs,opt}}$ and $B = 100$ since pilot experiments revealed no noticeable differences when larger values of B were used.

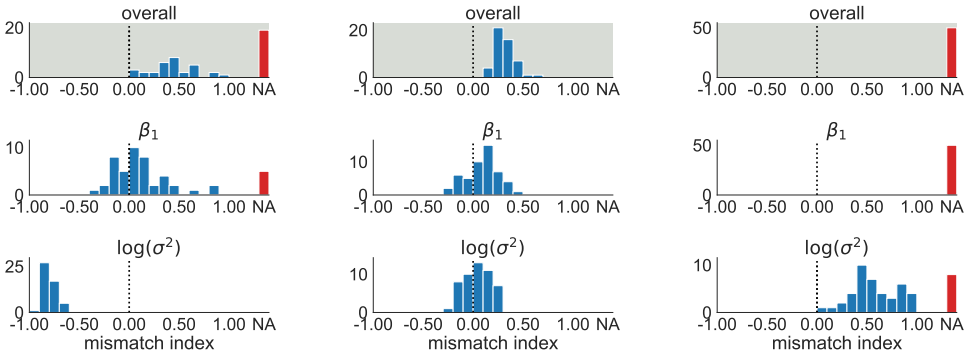
5.1.1. *Parameter inference.* We begin by assessing how well the standard posterior and the bagged posterior estimated the optimal coefficient vector $\beta_\circ := \arg\min_{\beta \in \mathbb{R}^d} \mathbb{E}\{(Y_1 - Z_1^\top \beta)^2\}$. For the linear simulation setting (regardless of G), $\beta_\circ = \beta_\dagger$; for the uncorrelated and nonlinear setting, $\beta_\circ = 3\beta_\dagger$. Let $\hat{\beta}$ and $\log \pi(\beta)$ denote, respectively, the (standard or bagged) posterior mean and log posterior density of β. We quantify estimation accuracy by computing the relative squared error (RSE) $\|\hat{\beta} - \beta_\circ\|_2^2 / \|\beta_\circ\|_2^2$ and the log posterior density (LPD) $\log \pi(\beta_\circ)$. For all experiments, we first ran BayesBag with $M = N$ in order to compute $\hat{M}_{\text{fs,opt}}$ and $\mathcal{I}$. If $\mathcal{I} \neq$ NA, we reran BayesBag using $M = \hat{M}_{\text{fs,opt}}$. If $\mathcal{I} =$ NA, we reran BayesBag using $M = 2N$ because we found that $\mathcal{I} =$ NA typically indicated a poorly chosen prior (either because the true parameter was unlikely or the model was poorly identified), which we could best mitigate by using more data.

Almost universally, the bagged posterior performed as well as or better than the standard posterior in terms of both relative squared error and log posterior density. Fig. 4 compares performance on default, 1-sparse, and nonlinear data for the varying prior choices $\lambda \in \{1, 4, 16\}$. The benefits of BayesBag were especially large for the excessively strong $\lambda = 16$ prior. As expected, BayesBag was also particularly effective on the misspecified nonlinear data. Fig. 5 compares performance on default and correlated data for $N \in \{50, 500\}$. Because the prior on β was very weak, there were significant identifiability issues when the data were heavily correlated—particularly in the small data regime of $N = 50$. The only case in which BayesBag performed noticeably worse than the standard posterior was in terms of relative squared error on the correlated data (Fig. 5A–B). However, BayesBag still performed better in terms of log posterior density (Fig. 5C–D), indicating superior calibration of the parameter estimate at the cost of a small additional bias.

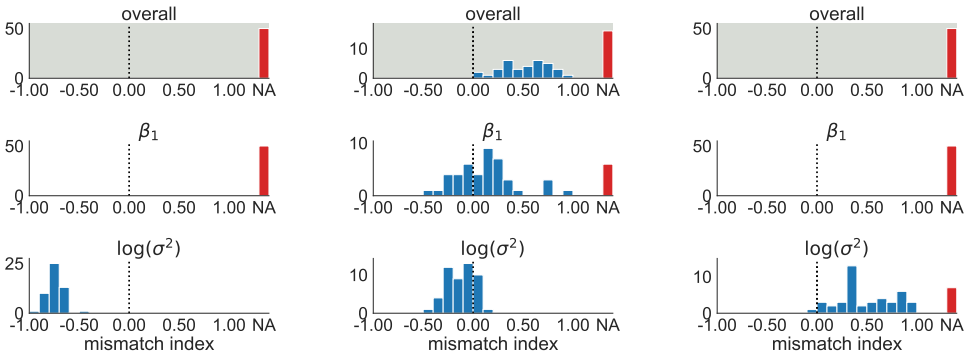
5.1.2. *Model criticism using the mismatch index.* Next, we reconsider the Section 5.1.1 scenarios from the perspective of model criticism. Our results demonstrate how $\mathcal{I}$ can be used to detect model–data mismatch when either (a) the likelihood is misspecified or (b)



(a) default, $\lambda = 1$ (b) 1-sparse, $\lambda = 1$ (c) nonlinear, $\lambda = 1$



(d) default, $\lambda = 4$ (e) 1-sparse, $\lambda = 4$ (f) nonlinear, $\lambda = 4$



(g) default, $\lambda = 16$ (h) 1-sparse, $\lambda = 16$ (i) nonlinear, $\lambda = 16$

Fig 6: Model–data mismatch indices $\mathcal{I}$ for selected parameters as well as the overall $\mathcal{I}$ value for default, 1-sparse, and nonlinear data for $\lambda \in \{1, 4, 16\}$. We only display one component of β since the $\mathcal{I}$ values followed the same distribution for all components.

the likelihood is well-specified but the prior is poorly chosen or some model parameters are poorly identified.

Dense versus sparse coefficient vector. We first consider the effects of varying prior choices. Fig. 6 compares the mismatch index with default and 1-sparse data for $\lambda \in \{1, 4, 16\}$. For the default data with a dense coefficient vector, the larger λ values result in a prior on β that is too concentrated near zero, leading to larger $\mathcal{I}$ values. The settings of $\lambda = 4$ and $\lambda = 16$ resulted in $\mathcal{I} = \mathsf{NA}$ for many of the datasets. For the 1-sparse data, on the other hand, a larger λ is more appropriate since all but one coefficient is zero. Thus, $\mathcal{I}$ was at most 0.5 for $\lambda \in \{1, 4\}$, although the stronger prior resulting from $\lambda = 16$ led to $\mathcal{I} = \mathsf{NA}$.

Linear versus nonlinear regression function. Finally, we consider misspecified data. Fig. 6 compares the mismatch index with default and nonlinear data for $\lambda \in \{1, 4, 16\}$. Due to the misspecification, $\mathcal{I} = \mathsf{NA}$ for all choices of λ.

Correlated versus uncorrelated regressors. See Appendix C.

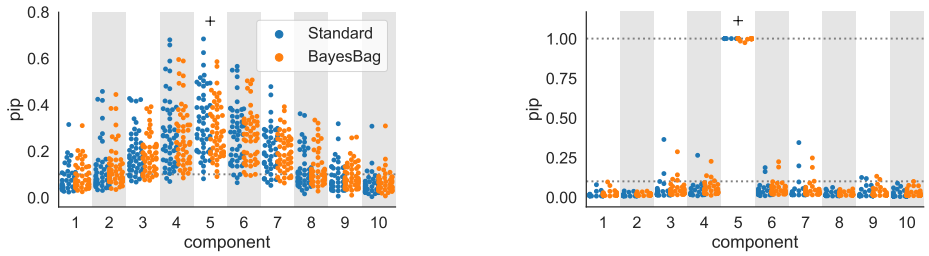
5.1.3. *Model selection.* We conclude our simulation study with feature selection in linear regression. For each $\gamma \in \{0, 1\}^D$, we define a model such that the dth component is included in the linear regression if and only if $\gamma_d = 1$. We consider a collection of models $\mathfrak{M}_{k^\star} := \{\gamma \in \{0, 1\}^D \mid \sum_{d=1}^D \gamma_d \leq k^\star\}$ where $k^\star \in \{1, \ldots, D\}$. 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 $\gamma_d = 1$. Letting $D_\gamma := \sum_{d=1}^D \gamma_d$, conditional on model γ, the parameter space is $\Theta_\gamma = \mathbb{R}^{D_\gamma + 1}$ and the generative model is

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

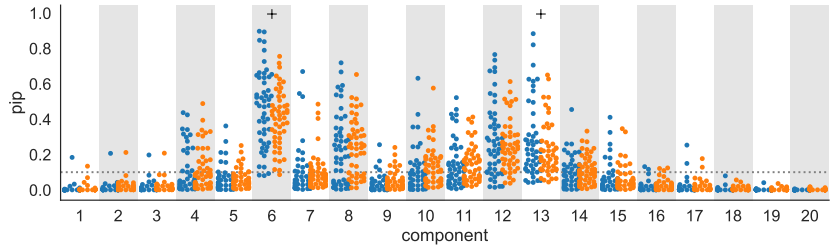
To perform posterior inference for γ, we analytically compute the marginal likelihood for each model γ, integrating out σ^2 and β. The prior for model $\gamma \in \mathfrak{M}_{k^\star}$ was $Q_0(\gamma) \propto q_0^{D_\gamma}(1 - q_0)^{D - D_\gamma}$, where $q_0 \in (0, 1)$ is the prior inclusion probability of a single component. For k-sparse data, we set $q_0 = k/D$.

In the spirit of GWAS fine-mapping (Schaid, Chen and Larson, 2018), we created synthetic data to simulate a scenario with many highly correlated regressors, of which only a few regressors are truly "causal." Specifically, we generated datasets under the correlated-k-sparse and correlated-k-sparse-nonlinear settings with either (a) $D = 10$, $N = 50$, and $k = 1$, or (b) $D = 20$, $N = 100$, and $k = 2$. We used $\lambda = 16$, as this helped to penalize the addition of extraneous features. We set $M = N$ per our default recommendation and $k^\star = 2$.

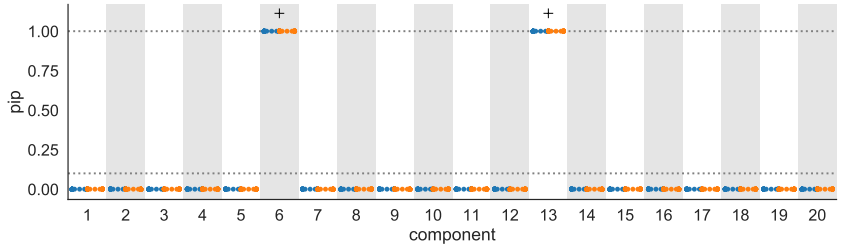
We are interested in verifying the theory of Section 4 in the finite-sample regime, which suggests that when the model is misspecified, similar models may be assigned wildly varying probabilities under the standard posteriors, while the bagged posterior probabilities will tend to be more balanced. In Figs. 7, 8 and C.2, we plot the standard and bagged posterior inclusion probabilities (pips) for each component for all 50 replications. First, Fig. 7 shows that when the model is correctly specified, Bayesian and BayesBag model selection behave



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



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



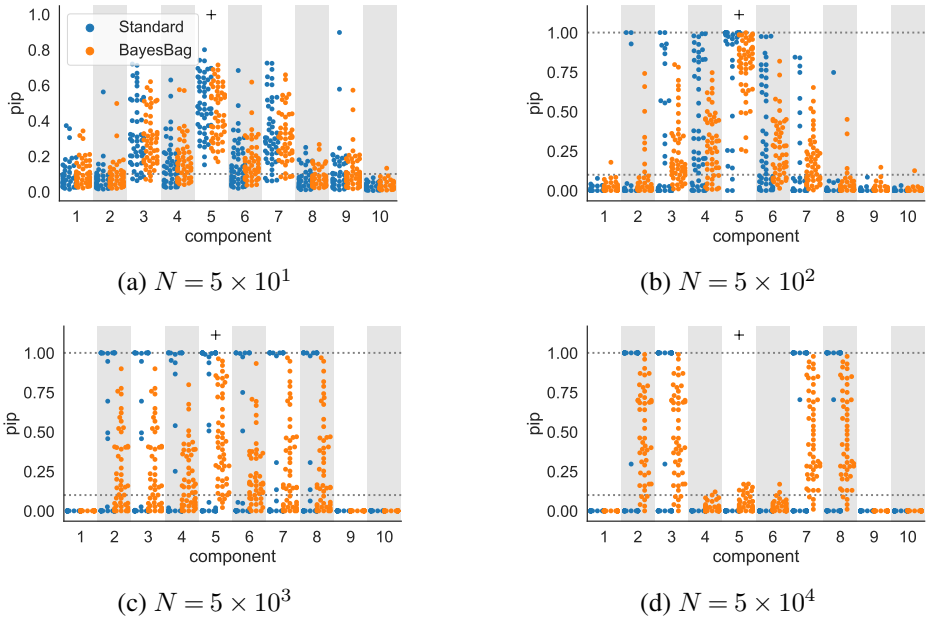
(d) correlated-2-sparse, $N = 10^4$

Fig 7: Posterior inclusion probabilities (pips) for well-specified data. Components used to generate the data are marked with a "+". The horizontal dotted lines indicate the prior inclusion probability and (when shown) the maximum inclusion probability (i.e., 1).

similarly. When N is small, BayesBag is slightly more conservative, assigning smaller posterior probabilities to both causal and non-causal components.

The results in the misspecified setting, shown in Figs. 8 and C.2, are more interesting and subtle. Due to the misspecification and correlated regressors, it no longer holds in general that the "causal" components will be selected. In fact, if $k^\star = D$, it is possible that all components will be selected — however, to maintain sparsity, we chose $k^\star = 2$. See Appendix E for derivations and further discussion; see also Buja et al. (2019a,b).

Figure 8 shows the results for correlated-1-sparse-nonlinear data. The regressor distribution G and coefficient vector $\beta_\dagger$ are such that, by symmetry, components $5 - i$ and $5 + i$ ($i = 0, \ldots, 4$) are equivalent. As $N \to \infty$, it is optimal to use component 3 (and/or component 7) and component 2 (and/or component 8). The Bayesian pip for either component 3 or



(a) $N = 5 \times 10^1$ (b) $N = 5 \times 10^2$

(c) $N = 5 \times 10^3$ (d) $N = 5 \times 10^4$

Fig 8: Posterior inclusion probabilities (pips) for misspecified correlated-1-sparse-nonlinear data. See caption for Fig. 7 for further explanation.

7 is ≈ 1 (with the other ≈ 0) and similarly for components 2 and 8, demonstrating that the standard posterior is highly unstable. On the other hand, the BayesBag pips for components 2, 3, 7, and 8 are close to uniformly distributed between 0 and 1, as expected. The Bayesian pips are also highly unstable as N increases, as illustrated by the pips ≈ 1 for components 4, 5, and 6 when $N = 5 \times 10^3$ that eventually go to zero as N increases; the BayesBag pips, on the other hand, do not exhibit this instability. Thus, we see exactly the unstable behavior predicted by Theorem 4.1 and Corollary 4.2. We defer discussion of Fig. C.2 and Fig. C.3 – which shows model–data mismatch index values for a representative subset of experimental configurations – to Appendix C.

5.2. *Hierarchical Mixed Effects Logistic Regression Model.* Next, we considered the canonical setting of mixed effects models, where Bayesian inference provides superior inferences compared to maximum likelihood and quasi-likelihood methods (Browne and Draper, 2006) and, even with significant amounts of data, can lead to dramatically different inferences (e.g., Giordano, Broderick and Jordan, 2018). Our objective was to compare the predictive performance of the bagged posterior (BayesBag) to both the standard Bayesian posterior and alternative likelihood-based methods. Specifically, we considered a 3-level logistic regression model with mixed effects and a balanced design:

$$v_k \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_v^2), \qquad u_{jk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_u^2),$$

$$Y_{ijk} \,|\, Z_{ijk}, u_{jk}, v_k \overset{\text{indep}}{\sim} \mathsf{Bern}(p_{ijk}), \qquad p_{ijk} = \text{logit}^{-1}(\beta_0 + Z_{ijk}^\top \beta + u_{jk} + v_k),$$

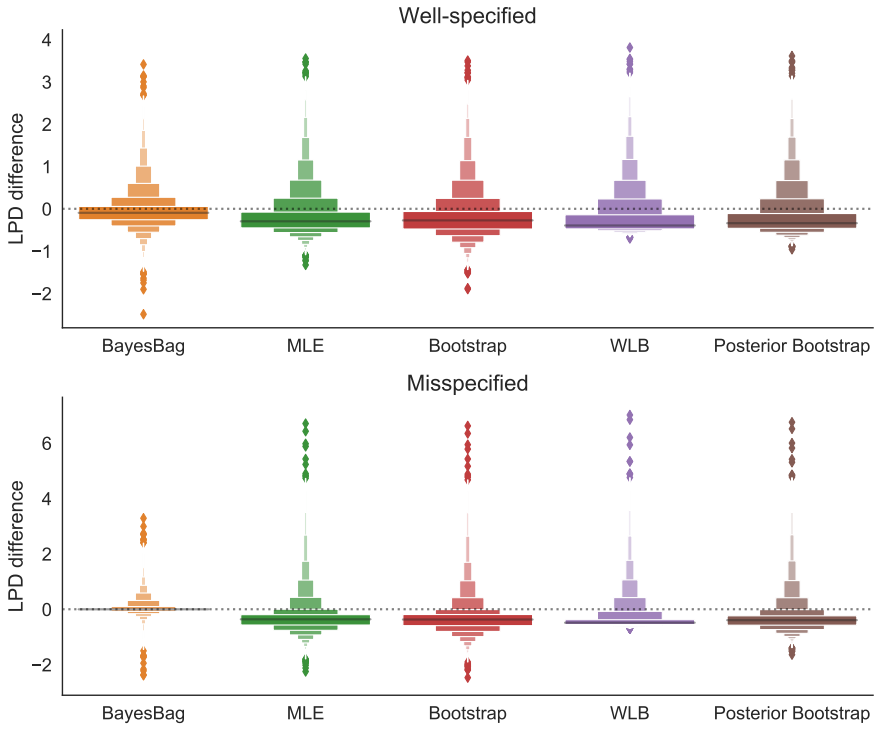


Fig 9: Differences between log predictive densities (LPDs) of the standard posterior and alternative methods on held-out data for the hierarchical mixed effects logistic regression model.

for $k = 1, \ldots, K$, $j = 1, \ldots, J$, $i = 1, \ldots, I$, $Z_{ijk} \in \mathbb{R}^D$, $\beta_0 \in \mathbb{R}$, and $\beta \in \mathbb{R}^D$. For example, Browne and Draper (2006) take Y_{ijk} to be a binary indicator of whether a woman received modern prenatal care during a pregnancy, with i indexing the birth, j indexing the mother, and k indexing the Guatemalan community to which the mother belonged. Note that in this particular example one would not expect a balanced design, but we used the balanced case for simplicity. For the Bayesian model, we used relatively diffuse priors:

$$\sigma_v \sim \mathsf{Unif}(0, 100), \qquad \sigma_u \sim \mathsf{Unif}(0, 100),$$

$$\beta_d \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 10^2), \quad d = 0, \ldots, D.$$

In our experiments, we took $I = 3$, $J = 8$, $K = 100$, and $D = 3$. We considered a well-specified scenario and a misspecified scenario. In the well-specified scenario, we generated covariates $Z_{ijkd} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and generated responses Y_{ijk} according to the assumed model with $\beta_0 = 0.65$, $\beta = (1, 1, 1)$, and $\sigma_v = \sigma_u = 3$. For the misspecified scenario, we generated data as in the well-specified case except that the random effects had unmodeled correlation structure: $(v_k, u_{1k}, \ldots, u_{Jk})$ was jointly Gaussian with correlation ρ between each pair of components, where $\rho = 0.99$.

We compared the predictive performance of the standard posterior, the bagged posterior, and four methods based on maximum likelihood estimation (with the random effects inte-

grated out): the standard MLE, the bootstrapped MLE, the weighted likelihood bootstrap (Newton and Raftery, 1994), and the posterior bootstrap (Lyddon, Walker and Holmes, 2018). See Section 7 for further discussion of these approaches. Fig. 9 shows the predictive performance of each method relative to the standard Bayesian posterior. Both the standard and bagged posteriors outperformed the MLE-based methods. In the well-specified scenario, standard Bayes was slightly better than BayesBag, as expected. Meanwhile, in the misspecified scenario, BayesBag had superior predictive performance compared to standard Bayes.

6. Experiments. In this section, we validate the effectiveness of BayesBag when applied to three diverse models using real-world data. Table 1 summarizes the real-world datasets we used for the three models. Some implementation details are deferred to Appendix B.

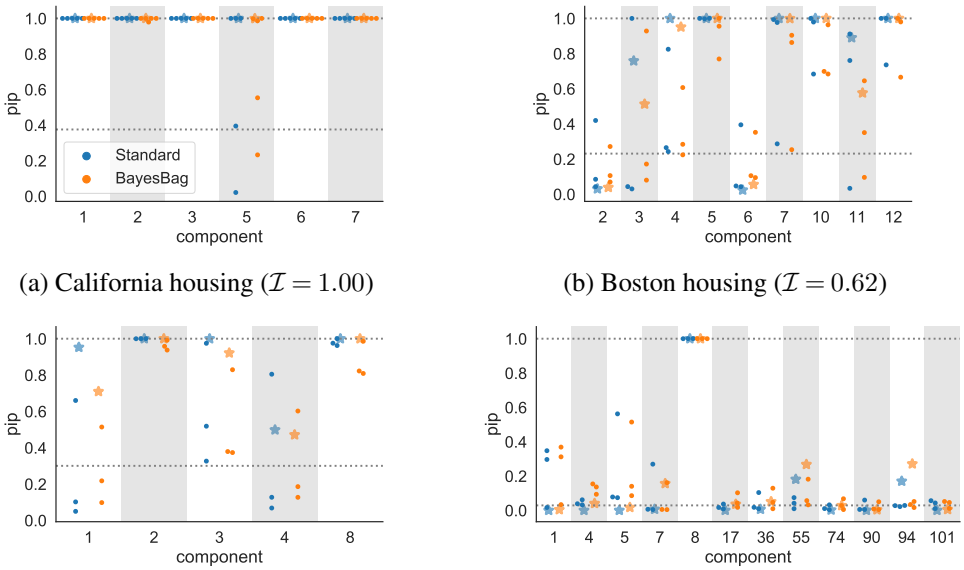
6.1. *Linear regression feature selection.* We compared standard Bayesian and BayesBag model selection for linear regression on four real-world datasets. We set $M = N$ per our default recommendation. In the notation of Section 5.1, we used a prior inclusion probability of $q_0 = 3/D$ and maximum nonzero components $k^\star = D$, except for the residential building dataset, where for computational tractability we used $k^\star = 3$. The residential building datasets required only 58 out of 104 principle components to explain 99% of the variance, whereas for the other three datasets, D out of D principle components were needed to explain 99% of the variance. Thus, we expected the parameters to be well-identified for all datasets except the residential building dataset. Therefore, we used $\lambda = 16$ for the residential building dataset and $\lambda = 1$ otherwise. The model mismatch indices (computed with $\gamma_d = 1$ for all $d = 1, \ldots, D$) were in agreement with expectations, as only the residential building dataset had a model mismatch index of NA. For the California housing, Boston housing, and diabetes datasets, we obtained mismatch indices of, respectively, 1.00, 0.62, and 0.03, indicating that the model was misspecified for the two housing datasets.

Fig. 10 shows the posterior inclusion probabilities (pips) for all four datasets. To compare the reliability of the methods, we also ran each method on subsets of the data obtained by randomly dividing each dataset into k roughly equally sized splits. We used $k = 3$ splits for all datasets except for California housing, for which we used $k = 5$ since N was substantially larger. Fig. 10 shows the pips for these splits as well. Generally, across splits, BayesBag produced lower-variance, more conservative pips that were more consistent with the pips from the full datasets. In general, BayesBag has a regularizing effect in which the pips tend to shrink toward the prior inclusion probability. These results were consistent with the simulation experiments in Section 5.1.3.

6.2. *Sparse logistic regression.* We next considered a sparse logistic regression model. We used the model and four cancer microarray datasets from Piironen and Vehtari (2017). Since we do not have access to ground truth parameters, we followed the procedure of Piironen and Vehtari (2017) and computed the mean log predictive density (MLPD) on 50 random train–test splits of each dataset, holding out 20% as test data on each split. Because the posteriors for all datasets are highly multimodal, $\hat{M}_{\text{opt}}$ and $\mathcal{I}$ were not applicable. Instead, a small pilot run using $M \in \{N, 1.5N, 2N\}$ suggested $M = 2N$ provided the best

TABLE 1
Real-world datasets used in experiments. LR = linear regression, BC = binary classification, PTR = phylogenetic tree reconstruction.

Name	Model	N	D
California housing	LR	20,650	8
Boston housing	LR	506	13
Diabetes	LR	442	10
Residential building	LR	371	105
Colon	BC	62	2,000
Leukemia	BC	72	7,129
Ovarian	BC	54	1,536
Prostate	BC	102	5,966
Whale mitochondrial coding DNA	PTR	14	10,605
Whale mitochondrial amino acids	PTR	14	3,535



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

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

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

(d) Residential building ($\lambda = 16$, $\mathcal{I} = \mathsf{NA}$)

Fig 10: Posterior inclusion probabilities (pips) for four real-world datasets when the data is split (•) and for the full dataset ($\star$). Only components with pips above the prior inclusion probability are shown. The horizontal dotted lines indicate the prior inclusion probability and the maximum inclusion probability (i.e., 1).

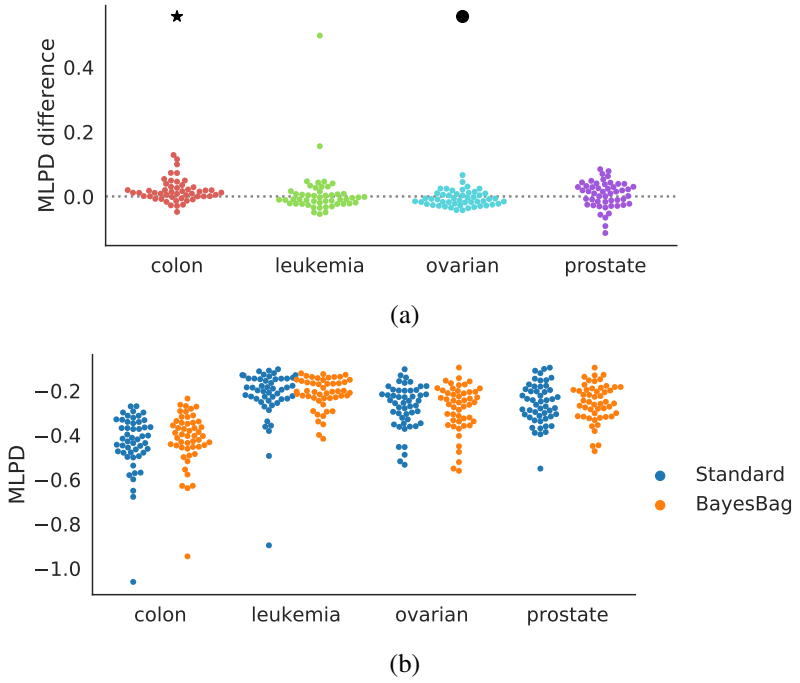
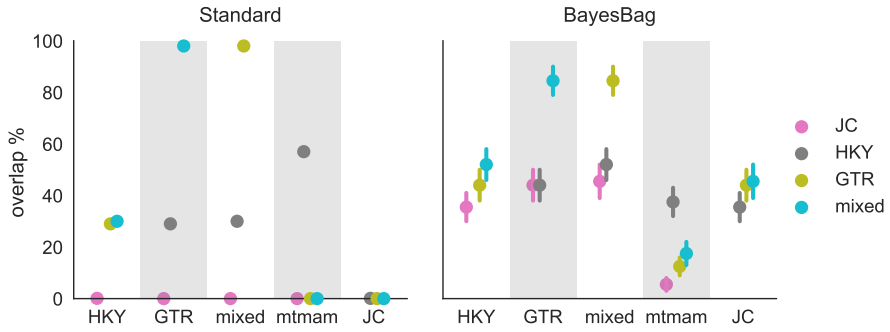


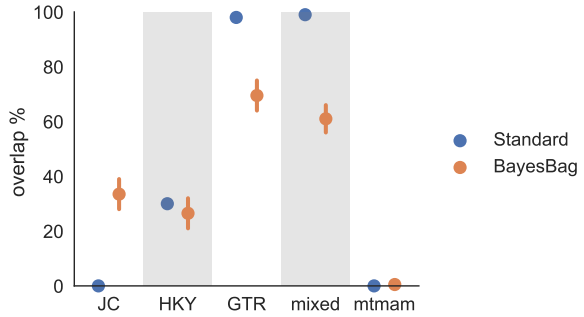
Fig 11: Predictive performance on 20% held-out test data across 50 random splits of the data for sparse logistic regression. (A) Differences in mean log predictive density (MLPD) between BayesBag and standard Bayes; a positive difference means BayesBag outperformed standard Bayes. Datasets marked with a $\star$ (respectively, $\bullet$) exhibited a statistically significant difference in the positive (respectively, negative) direction ($p < 0.05$, two-sided Wilcoxon signed-rank test). (B) MLPD of BayesBag and standard Bayes. The ratios of variances of the MLPDs for colon, leukemia, ovarian, and prostrate were, respectively, 1.2, 3.1, 0.89, and 1.1.

performance for BayesBag. We used $B = 50$ bootstrap samples to approximate the bagged posterior. (Nearly identical results were obtained with $B = 25$, which indicated that it was not necessary to make B larger.) The results, shown in Fig. 11, suggest that the predictive performance of the bagged posterior is (1) equal or slightly better on average and (2) more stable—that is, lower-variance—across splits. The only exception was the ovarian dataset, where BayesBag had slightly lower MLPD on average and slightly higher variance MLPDs. However, the average MLPD difference was only -0.007 nats on the ovarian dataset, while it was 0.018 nats on the colon dataset. The average MLPD difference for the leukemia and prostate dataset were -0.008 and 0.006 nats, respectively, although these differences were not statistically significant.

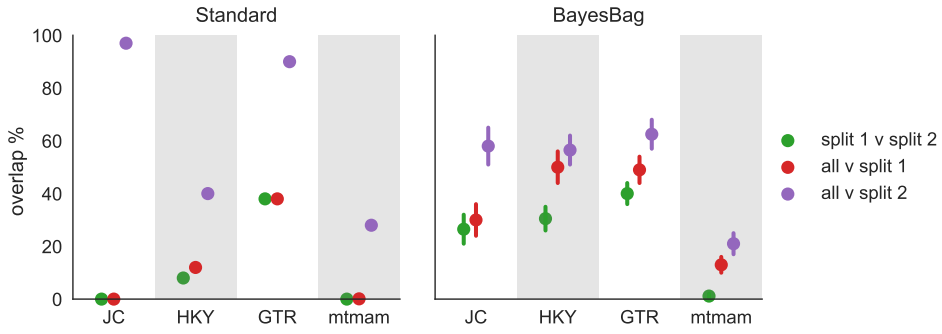
6.3. *Phylogenetic tree reconstruction.* Finally, we investigate the use of BayesBag model selection for reconstructing the phylogenetic tree of a collection of species based on their observed characteristics. This is an important model selection problem, due to the



(a) Inter-model comparison



(b) Comparison to mixed using standard Bayesian model selection



(c) Intra-model comparison

Fig 12: Comparison of standard Bayesian and BayesBag model selection consistency on the whale dataset in terms of overlap of 99% high-probability regions (HPRs). To quantify uncertainty in the overlap due to Monte Carlo error, 80% bootstrapped confidence intervals are provided for the BayesBag overlaps.

widespread use of phylogeny reconstruction algorithms. Systematists have exhaustively documented that Bayesian model selection of phylogenetic trees can behave poorly. In particular, the standard Bayesian approach can provide contradictory results depending on what characteristics are used (for example, coding 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. We used the whale dataset from Yang (2008), which consists of mitochondrial coding DNA from 13 whale species and the hippopotamus. (The hippopotamus was included as an "outgroup" species to identify the root of the tree, because the evolutionary models are time-reversible and hence the trees are modeled as unrooted.) We considered 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. For brevity, we refer to the models as JC, HKY, GTR, mixed, and mtmam, respectively. For the BayesBag settings, we used $M = N$ and $B = 100$ in all experiments.

Our goal was to investigate whether BayesBag can avoid the self-contradictory inferences produced by the standard posterior. To this end, we compared the output of different configurations of the data, model, and inference method, as follows. We computed the set of trees in the 99% high-probability regions (HPRs) for each $\langle$data, model, inference method$\rangle$ configuration. We then computed the overlap of the 99% HPRs in terms of both probability mass and number of trees. Since we approximated the bagged posterior via Monte Carlo with $B = 100$ (as in Eq. (2)), we quantify the uncertainty in these overlaps by reporting 80% confidence intervals. (We computed these confidence intervals using standard bootstrap methodology for a Monte Carlo estimate.)

For the first experiment, we looked at the overlap between pairs of models when using the standard posterior and bagged posterior. As shown in Fig. 12(A) and Table C.1, there was substantially more overlap when using the bagged posterior. The difference is particularly noticeable when comparing JC (the simplest model) or mtmam (the amino acid model) to the other models. When using the standard posterior, JC had either 0% or (in one case) 0.1% overlap with the other models while mtmam only overlapped with HKY. Meanwhile, when using BayesBag, all models had nonzero overlap, with typical amounts ranging from 30% to 50%. Hence, when applied to different models, BayesBag did not produce contradictory results.

However, the good overlap exhibited by BayesBag 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. To investigate this possibility, we considered the overlap of the bagged posterior for each model and the standard posterior for mixed, which is the most complex of the DNA models. As shown in Fig. 12(B) and Table C.2, all the bagged posteriors (with the exception of mtmam) put substantial posterior probability on the 99% HPR of the standard mixed posterior. Moreover, all but BayesBag mtmam had two trees in the overlap, which was the maximum possible since the standard mixed 99% HPR only contained two trees.

Next, we performed 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). Since the results for GTR and

mixed were very similar, we only report results for JC, HKY, GTR, and mtmam. Ideally, for a fixed model, we would expect to see substantial overlap when comparing the results for these three datasets. However, when using the standard posterior, there was little to no overlap in many cases, particularly for the simpler DNA models and the amino acid model; see Fig. 12(C) and Table C.3. On the other hand, the bagged posteriors typically exhibited 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.

Finally, we computed the mismatch index for each model on the complete whale dataset, obtaining values of 0.21 (JC), 0.16 (HKY), 0.47 (GTR), 0.84 (mixed), and 0.34 (mtmam). These mismatch indices indicate significant but not overwhelming amounts of model misspecification, with the simpler models perhaps underestimating the actual amount of misspecification. In our experiments, we used $M = N$ for BayesBag. However, Douady et al. (2003) found that BayesBag yielded similar results to standard maximum likelihood bootstrap for phylogenetic tree reconstruction, which suggests that using $M = N$ may be too conservative. Combined with the finding of moderate values for the model–data mismatch index, it would be worth investigating the use of BayesBag with $M > N$. Although our theoretical results for model selection suggest the use of $M \leq N$, phylogenetic tree inference may—at least in certain ways—behave more like parameter inference due to the very large number of trees as well as the importance of inferring a significant number of tree-agnostic parameter values.

7. Discussion. We conclude by situating BayesBag in the wider literature on robust Bayesian inference and model criticism. With this additional context in place, we highlight the strengths of our approach, while also suggesting fruitful directions for future development.

7.1. *Robust Bayesian inference.* Two common themes emerge when surveying existing methods for robust Bayesian inference. First, many methods require choosing a free parameter. But the proposals for choosing free parameters are either (a) heuristic, (b) strongly dependent on being in the asymptotic regime, or (c) computationally prohibitive for most real-world problems. Second, those methods without a free parameter all lose key parts of what make the Bayesian approach attractive. For example, they strongly rely on asymptotic assumptions, make a Gaussian assumption, or do not incorporate a prior distribution.

The power posterior is perhaps the most widely studied method for making the posterior robust to model misspecification (Grünwald, 2012; Grünwald and van Ommen, 2017; Holmes and Walker, 2017; Lyddon, Holmes and Walker, 2019; Miller and Dunson, 2018; Syring and Martin, 2019). For a likelihood function $L(\theta)$, prior distribution Π_0, and any $\zeta \geq 0$, the *ζ-power posterior* is defined as $\Pi^{(\zeta)}(\mathrm{d}\theta) \propto L(\theta)^{\zeta}\Pi_0(\mathrm{d}\theta)$. Hence, $\Pi^{(1)}$ is equal to the standard posterior and $\Pi^{(0)}$ is equal to the prior. Typically, ζ is set to a value between these two extremes, as there is significant theoretical support for the use of power posteriors with $\zeta \in (0, 1)$ (Bhattacharya, Pati and Yang, 2019; Grünwald, 2012; Miller and Dunson, 2018; Royall and Tsou, 2003; Walker and Hjort, 2001). However, there are two significant methodological challenges. First, computing the power posterior often requires new

computational methods or additional approximations, particularly in latent variable models (Antoniano-Villalobos and Walker, 2013; Miller and Dunson, 2018). Second, choosing an appropriate value of ζ can be difficult. Grünwald (2012) proposes SafeBayes, a theoretically sound method which is evaluated empirically in de Heide et al. (2019); Grünwald and van Ommen (2017). However, SafeBayes is computationally prohibitive except with simple models and very small datasets. In addition, the underlying theory relies on strong assumptions on the model class. Many other methods for choosing ζ have been suggested, but they are either heuristic or rely on strong asymptotic assumptions such as the accuracy of the plug-in estimator for the sandwich covariance (Holmes and Walker, 2017; Lyddon, Holmes and Walker, 2019; Miller and Dunson, 2018; Royall and Tsou, 2003; Syring and Martin, 2019).

More in the spirit of BayesBag are a number of bootstrapped point estimation approaches (Chamberlain and Imbens, 2003; Lyddon, Holmes and Walker, 2019; Lyddon, Walker and Holmes, 2018; Newton and Raftery, 1994; Rubin, 1981). However, unlike BayesBag, these methods compute a collection of *maximum a posteriori* (MAP) or *maximum likelihood* (ML) estimates. The weighted likelihood bootstrap of Newton and Raftery (1994) and a generalization proposed by Lyddon, Holmes and Walker (2019) do not incorporate a prior, and therefore lose many of the benefits of Bayesian inference. The related approach of Lyddon, Walker and Holmes (2018), which includes the weighted likelihood bootstrap and standard Bayesian inference as limiting cases, draws the bootstrap samples partially from the posterior and partially from the empirical distribution. Unfortunately, there is no accompanying theory to guide how much the empirical distribution and posterior distribution should be weighted relative to each other—nor rigorous robustness guarantees. Moreover, bootstrapped point estimation methods can behave poorly when the MAP and ML estimates are not well-behaved—for example, due to the likelihood being peaked (or even tending to infinity) in a region of low posterior probability.

Müller (2013) suggests replacing the standard posterior with a Gaussian distribution with covariance proportional to a plug-in estimate of the sandwich covariance. However, the applicability of such an approach seems rather limited, since it requires that the dataset be large enough that (1) the sandwich covariance can be well-estimated, and (2) the posterior uncertainty can be represented as approximately Gaussian. If both these conditions hold, then Bayesian inference may be adding minimal value anyway.

7.2. *Misspecification and decision theory.* When the model is well-specified, Bayesian inference is the optimal procedure for updating beliefs in light of new data, no matter the loss function (Bernardo and Smith, 2000; Robert, 1994). However, when the model is misspecified, our analysis of BayesBag only shows (near) optimality under log loss. When some other loss function is ultimately of interest, there is no reason to assume BayesBag will provide high-quality inferences – although the log loss does serve as a reasonable and universally-applicable default choice. However, when the model is misspecified and a loss function is available, generalized belief updating (that is, using a Gibbs posterior) may be more appropriate (Bissiri, Holmes and Walker, 2016; Syring, Hong and Martin, 2019; Syring and Martin, 2017). It is conceptually straightforward to combine our BayesBag methodology with generalized belief updates to obtain better-calibrated inferences that are (near) optimal for the loss function of interest.

7.3. *Model criticism.* Methods for model criticism typically involve "predictive checks." For example, prior and posterior predictive checks compare the observed data to data generated from the prior and posterior predictive distributions, respectively (Box, 1980; Gelman, Meng and Stern, 1996; Guttman, 1967; Rubin, 1984; Vehtari and Ojanen, 2012). The version most closely related to BayesBag is the *population* predictive check (Ranganath and Blei, 2019), which is based on a fusion of Bayesian and frequentist thinking. The population predictive check aims to avoid the data-reuse of posterior predictive checks by comparing data generated from the posterior predictive to data generated from the true distribution. As a computable approximation to the ideal population predictive check, Ranganath and Blei (2019) suggest bootstrapping datasets and computing the predictive check on the data not included in the bootstrap sample. Like all predictive checks, the population predictive check requires choosing a measure of data similarity, which BayesBag and our model–data mismatch index do not require. Having this additional degree of freedom could be useful if the data analyst knows what aspects of the data they wish to focus on, but it may also be an unwelcome burden. An additional challenge with all predictive checks is that, when the goal is to diagnose misspecification, an additional calibration step much be performed (Hjort, Dahl and Steinbakk, 2006; Robins, van der Vaart and Ventura, 2000). The mismatch index, on the other hand, is already on an interpretable scale.

7.4. *The benefits of BayesBag.* In view of previous work, the BayesBag approach has a number of attractive features that make it flexible, easy-to-use, and widely applicable. From a methodological perspective, BayesBag is general-purpose. It relies only on carrying out standard posterior inference, it is applicable to a wide range of models, and it can make full use of modern probabilistic programming tools: the only other requirement is the design of a bootstrapping scheme. Although this paper focuses on using BayesBag with independent observations, future applications can draw on the large literature devoted to adapting the bootstrap to more complex models such as those involving time-series and spatial data; moreover, in hierarchical models such as multilevel regression models (Gelman and Hill, 2006), it would be straightforward to bootstrap data within each group. BayesBag is also general-purpose in the sense that it is useful no matter whether the ultimate goal of Bayesian inference is parameter estimation, prediction, or model selection.

Another appeal of BayesBag as a methodology is that its only two hyperparameters are straightforward to choose. First, the bootstrap dataset size M has a natural, theoretically well-justified choice of $M = N$ that results in conservative inferences. However, when the posterior quantities of interest are sufficiently close to being Gaussian-distributed, the bootstrap dataset size can also be selected in a data-driven way using $\hat{M}_{\infty,\text{opt}}$ or $\hat{M}_{\text{fs,opt}}$, making inferences less conservative when the data supports it. See Section 2.3.2 for further discussion of these points. Second, as described in Section 2.2.2, validating that the number of bootstrap datasets B is sufficiently large only requires computing simple Monte Carlo error bounds. Moreover, defaulting to $B = 50$ or 100 appears to be an empirically sound choice across a range of problems.

In terms of computation, while there is an additional cost due to the need to compute the posterior for each bootstrapped dataset, it is trivial to compute the bootstrapped posteriors in parallel. Nonetheless, speeding up BayesBag with more specialized computational methods

could be worthwhile in some applications. For example, we suggest one simple approach to speeding up Markov chain Monte Carlo (MCMC) in Appendix D. Pierre Jacob has proposed using more advanced unbiased MCMC techniques for potentially even greater computational efficiency.[1]

A final benefit of BayesBag is that it incorporates robustness features of frequentist methods into Bayesian inference without sacrificing the core benefits of the Bayesian approach such as flexible modeling and the use of prior information. This synthesis of Bayesian and frequentist approaches compares favorably to existing methods, which, as we described above, either sacrifice some useful part of standard Bayesian inference or introduce tuning parameters that are difficult to choose. Indeed, BayesBag can actually diagnose how much robustness is necessary via the the model–data mismatch index. An exciting direction for future work is to better understand the finite-sample properties of BayesBag and the mismatch index.

Acknowledgments. Thanks to Pierre Jacob for bringing P. Bühlmann's BayesBag paper to our attention and to Ziheng Yang for sharing the whale dataset and his MrBayes scripts. Thanks also to Ryan Giordano and Pierre Jacob for helpful feedback on an earlier draft of this paper, and to Peter Grünwald, Natalia Bochkina, Mathieu Gerber, and Anthony Lee for helpful discussions.

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.

ANTONIANO-VILLALOBOS, I. and WALKER, S. G. (2013). Bayesian Nonparametric Inference for the Power Likelihood. *Journal of Computational and Graphical Statistics* **22** 801–813.

BERK, R. H. (1966). Limiting Behavior of Posterior Distributions when the Model is Incorrect. *The Annals of Mathematical Statistics* **37** 51–58.

BERNARDO, J. M. and SMITH, A. F. M. (2000). *Bayesian Theory*. Wiley, New York.

BHATTACHARYA, A., PATI, D. and YANG, Y. (2019). Bayesian fractional posteriors. *The Annals of Statistics* **47** 39–66.

BISSIRI, P. G., HOLMES, C. C. and WALKER, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **78** 1103–1130.

BLEI, D. M. (2014). Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models. *Annual Review of Statistics and Its Application* **1** 203–232.

BOX, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. In *Robustness in Statistics* 201–236. Elsevier.

BOX, G. E. P. (1980). Sampling and Bayes' Inference in Scientific Modelling and Robustness. *Journal of the Royal Statistical Society. Series A (General)* **143** 383–430.

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

BROWNE, W. J. and DRAPER, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis* **1** 473–514.

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

BÜHLMANN, P. (2014). Discussion of Big Bayes Stories and BayesBag. *Statistical Science* **29** 91–94.

[1] https://statisfaction.wordpress.com/2019/10/02/bayesbag-and-how-to-approximate-it/

BÜHLMANN, P. and YU, B. (2002). Analyzing bagging. *The Annals of Statistics* **30** 927–961.

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.

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.

CHAMBERLAIN, G. and IMBENS, G. (2003). Nonparametric applications of Bayesian inference. *Journal of Business Economic Statistics* **21** 12–18.

CHEN, L. H. Y., GOLDSTEIN, L. and SHAO, Q.-M. (2010). *Normal Approximation by Stein's Method*. *Probability and Its Applications*. Springer Science & Business Media, Berlin, Heidelberg.

CLARKE, B. S. and BARRON, A. R. (1990). Information-theoretic asymptotics of Bayes methods. *Information Theory, IEEE Transactions on* **36** 453–471.

COX, D. R. (1990). Role of Models in Statistical Analysis. *Statistical Science* **5** 169–174.

DAWID, A. P. (2011). Posterior model probabilities. In *Philosophy of Statistics* 607–630. Elsevier, New York.

DE HEIDE, R., KIRICHENKO, A., MEHTA, N. and GRÜNWALD, P. D. (2019). Safe-Bayesian Generalized Linear Regression. *arXiv.org* **arXiv:1910.09227 [math.ST]**.

DIACONIS, P. and ZABELL, S. L. (1982). Updating subjective probability. *Journal of the American Statistical Association* **77** 822–830.

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.

EFRON, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7** 1–26.

GELMAN, A. and HILL, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

GELMAN, A., MENG, X.-L. and STERN, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6** 733–807.

GELMAN, A. and SHALIZI, C. R. (2011). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*.

GELMAN, A., CARLIN, J., STERN, H., DUNSON, D. B., VEHTARI, A. and RUBIN, D. B. (2013). *Bayesian Data Analysis*, Third ed. Chapman and Hall/CRC.

GIORDANO, R., BRODERICK, T. and JORDAN, M. I. (2018). Covariances, Robustness, and Variational Bayes. *Journal of Machine Learning Research* **19** 1–49.

GRÜNWALD, P. D. (2012). The Safe Bayesian: Learning the Learning Rate via the Mixability Gap. In *Algorithmic Learning Theory* 169–183.

GRÜNWALD, P. D. and VAN OMMEN, T. (2017). Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis* **12** 1069–1103.

GUTTMAN, I. (1967). The Use of the Concept of a Future Observation in Goodness-Of-Fit Problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **29** 83–100.

HJORT, N. L., DAHL, F. A. and STEINBAKK, G. H. (2006). Post-Processing Posterior Predictive p Values. *Journal of the American Statistical Association* **101** 1157–1174.

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

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.

JEFFREY, R. C. (1968). Probable Knowledge. In *The Problem of Inductive Logic* (I. Lakatos, ed.) 166–180. North-Holland, Amsterdam.

JEFFREY, R. C. (1990). *The Logic of Decision*, 2nd ed. University of Chicago Press.

KALLENBERG, O. (2002). *Foundations of Modern Probability*, 2nd ed. Springer, New York, NY.

KLEIJN, B. J. K. and VAN DER VAART, A. W. (2012). The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics* **6** 354–381.

KOEHLER, E., BROWN, E. and HANEUSE, S. J. P. A. (2009). On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses. *The American Statistician* **63** 155–162.

KÜNSCH, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics* **17** 1217–1241.

LEHMANN, E. L. (1990). Model specification: the views of Fisher and Neyman, and later developments. *Statistical Science* **5** 160–168.

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

LYDDON, S. P., HOLMES, C. C. and WALKER, S. G. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika* **106** 465–478.

LYDDON, S. P., WALKER, S. G. and HOLMES, C. C. (2018). Nonparametric learning from Bayesian models with randomized objective functions. In *Advances in Neural Information Processing Systems*.

MAMMEN, E. (1992). Bootstrap, wild bootstrap, and asymptotic normality. *Probability Theory and Related Fields* **93** 439–455.

MENG, L. and DUNSON, D. B. (2019). Comparing and weighting imperfect models using D-probabilities. *Journal of the American Statistical Association* **0** 1–33.

MILLER, J. W. and DUNSON, D. B. (2018). Robust Bayesian Inference via Coarsening. *Journal of the American Statistical Association* **114** 1113–1125.

MÜLLER, U. K. (2013). Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix. *Econometrica: Journal of the Econometric Society* **81** 1805–1849.

NEWTON, M. A. and RAFTERY, A. E. (1994). Approximate Bayesian Inference with the Weighted Likelihood Bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)* **56** 3–46.

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]**.

PELIGRAD, M. (1998). On the blockwise bootstrap for empirical processes for stationary sequences. *The Annals of Probability* **26** 877–901.

PIIRONEN, J. and VEHTARI, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics* **11** 5018–5051.

RANGANATH, R. and BLEI, D. M. (2019). Population Predictive Checks. *arXiv.org* **arXiv:1908.00882 [stat.ME]**.

ROBERT, C. P. (1994). *The Bayesian Choice*. Springer, New York, NY.

ROBINS, J. M., VAN DER VAART, A. W. and VENTURA, V. (2000). Asymptotic Distribution of P Values in Composite Null Models. *Journal of the American Statistical Association* **95** 1143–1156.

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.

ROYALL, R. and TSOU, T.-S. (2003). Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65** 391–404.

RUBIN, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics* **9** 130–134.

RUBIN, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* **12** 1151–1172.

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.

SYRING, N., HONG, L. and MARTIN, R. (2019). Gibbs posterior inference on value-at-risk. *Scandavian Actuarial Journal* **2019** 548–557.

SYRING, N. and MARTIN, R. (2017). Gibbs posterior inference on the minimum clinically important difference. *Journal of Statistical Planning and Inference* **187** 67–77.

SYRING, N. and MARTIN, R. (2019). Calibrating general posterior credible regions. *Biometrika* **106** 479–486.

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. University of Cambridge.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer, New York.

VEHTARI, A. and OJANEN, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys* **6** 142–228.

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. International Conference on Genome Informatics* **13** 82–92.

WALKER, S. G. and HJORT, N. L. (2001). On Bayesian consistency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63** 811–821.

WHITE, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica: Journal of the Econometric Society* **50** 1–25.

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.

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

YANG, Z. (2008). Empirical evaluation of a prior for Bayesian phylogenetic inference. *Philosophical Transactions of the Royal Society B: Biological Sciences* **363** 4031–4039.

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.

APPENDIX A: INTERPRETATION OF THE BAGGED POSTERIOR IN TERMS OF JEFFREY CONDITIONALIZATION

The bagged posterior has an insightful interpretation in terms of Jeffrey conditionalization. This interpretation elegantly unifies the Bayesian and frequentist elements of the bagged posterior which might otherwise seem challenging to interpret coherently (e.g., the covariance decomposition in Eq. (5)).

Suppose we have a model $p(x, y)$ of two variables x and y. In the absence of any other data or knowledge, we would quantify our uncertainty in x and y via the marginal distributions $p(x) = \int p(x|y)p(y)\mathrm{d}y$ and $p(y) = \int p(y|x)p(x)\mathrm{d}x$, respectively. Now, suppose we are informed that the true distribution of x is $p_\circ(x)$, but we are not given any samples of x or y. We would then quantify our uncertainty in x via $p_\circ(x)$, and a natural way to quantify our uncertainty in y is via $q(y) := \int p(y|x)p_\circ(x)\mathrm{d}x$. The idea is that $q(x, y) := p(y|x)p_\circ(x)$ updates the model to have the correct distribution of x, while remaining as close as possible to the original model $p(x, y)$. This is referred to as Jeffrey conditionalization (Diaconis and Zabell, 1982; Jeffrey, 1968, 1990).

Suppose $x = (x_1, \ldots, x_N)$ is the data and $y = \theta$ is a parameter, so that $p(x, y) = p(x_{1:N}, \theta)$ is the joint distribution of the data and the parameter. If we are informed that the true distribution of the data is $p_{\circ N}(x_{1:N})$, then the Jeffrey conditionalization approach above is to quantify our uncertainty in θ by

$$q(\theta) = \int p(\theta \,|\, x_{1:N}) p_{\circ N}(x_{1:N}) \mathrm{d}x_{1:N}. \tag{17}$$

Now, suppose we are not informed of the true distribution exactly, but we are given data $X_1, \ldots, X_N$ i.i.d. $\sim p_\circ$. Since the empirical distribution $\mathbb{P}_N := N^{-1} \sum_{n=1}^N \delta_{X_n}$ is a consistent estimator of $p_\circ$ and $p_{\circ N}(x_{1:N}) = \prod_{n=1}^N p_\circ(x_n)$, it is natural to plug in $\prod_{n=1}^N \mathbb{P}_N$ for $p_{\circ N}$ in Eq. (17). Doing so, we arrive at the bagged posterior:

$$q(\theta) \approx \int p(\theta \,|\, x_{1:N}) \prod_{n=1}^N \mathbb{P}_N(\mathrm{d}x_n) = \mathbb{E}\big\{p(\theta \,|\, X^*_{1:N}) \,|\, X_{1:N}\big\}$$

where $X^*_1, \ldots, X^*_N$ i.i.d. $\sim \mathbb{P}_N$ given $X_{1:N}$.

APPENDIX B: ADDITIONAL EXPERIMENTAL DETAILS

Hierarchical Mixed Effects Logistic Regression Model. We computed maximum likelihood estimates with the `R` package `lme4`, which uses a Laplace approximation to integrate out the random effects; for prediction we used Monte Carlo to integrate out the random effects. To approximate the standard and bagged posteriors, we used Stan's implementation of dynamic Hamiltonian Monte Carlo with 4 chains (for the standard posterior) or 2 chains (for the bagged posterior) each run for 2,000 total iterations (discarding the first half as burn-in). For BayesBag we used $B = 100$ bootstrap datasets. For the MLE-based bootstrap methods we used 500 bootstrap datasets.

Sparse logistic regression. We used M. Betancourt's Stan implementation of the model from Piironen and Vehtari (2017).[2] To approximate the standard and bagged posteriors, we used Stan's implementation of dynamic Hamiltonian Monte Carlo with 4 chains each run for 2,000 total iterations (discarding the first half as burn-in). We used Stan's built-in convergence diagnostics in our preliminary experiments to confirm acceptable mixing; however, we then turned the diagnostics off because they significantly increased runtime.

Phylogenetic tree reconstruction. To approximate the standard and bagged posteriors, we used 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 confirmed acceptable mixing using the MrBayes built-in convergence diagnostics.

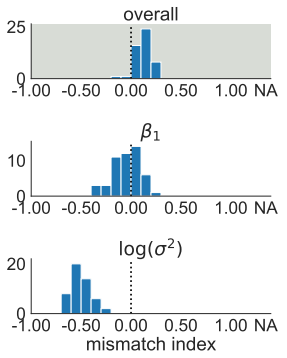
APPENDIX C: ADDITIONAL FIGURES AND TABLES

The mismatch index with correlated versus uncorrelated regressors. Continuing the discussion in Section 5.1.2, we consider cases involving poor identifiability in the model. Fig. C.1 compares the mismatch index with default and correlated data for $N \in \{50, 200\}$. The poor identifiability of the correlated data was correctly detected by $\mathcal{I}$. The identifiability issue becomes less severe with more data, which is reflected in the $\mathcal{I}$ values clustered around zero when $N = 200$. On the other hand, no identifiability issues were present for the uncorrelated default data, resulting in $\mathcal{I}$ values that were appropriately clustered near 0.

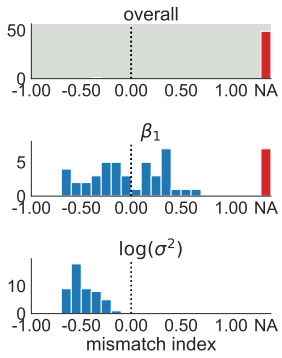
More on linear regression feature selection with simulated data. Figure C.2 shows similar results for correlated-2-sparse-nonlinear data. Note, however, that in this case it is asymptotically optimal to select one of the causal components (13) but not optimal to select the other causal component (6); rather, using either component 5 or 7 provides a better fit than component 6. Even though component 13 is asymptotically optimal, the standard pips for components near 13 sometimes remain at or close to 1 even when N is in the thousands. The BayesBag pips do not display the same instability.

Mismatch index results for linear regression feature selection with simulated data. Fig. C.3 shows model–data mismatch index values for a representative subset of experimental configurations (computed with $\gamma_d = 1$ for all $d = 1, \ldots, D$). For the correlated-k-sparse data, the mismatch indices were either near zero or NA, reflecting that the model is correctly specified but there are some issues with poor identifiability. For the correlated-k-sparse-nonlinear data, the mismatch indices were almost all NA, reflecting that the model is misspecified and there may also be identifiability issues.

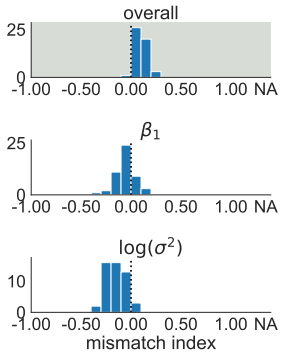
[2] https://betanalpha.github.io/assets/case_studies/bayes_sparse_regression.html



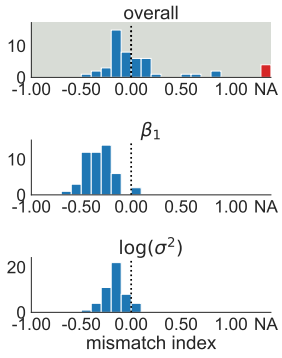
(a) default, $N = 50$



(b) correlated, $N = 50$

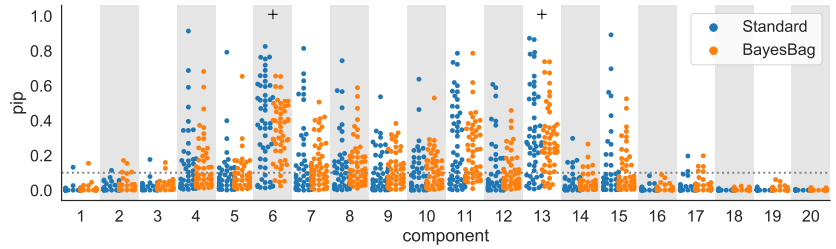


(c) default, $N = 200$

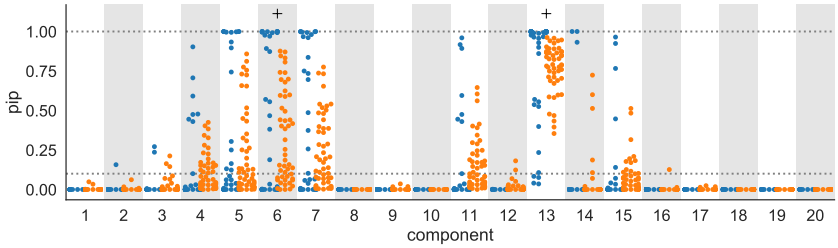


(d) correlated, $N = 200$

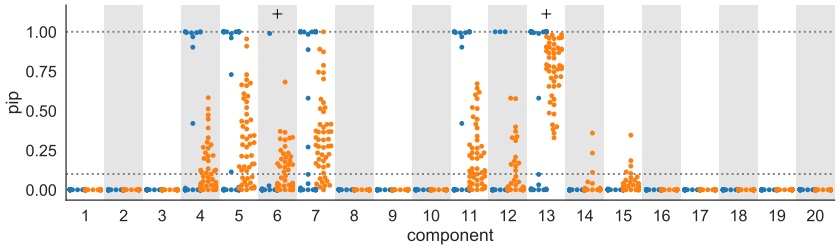
Fig C.1: Model–data mismatch indices $\mathcal{I}$ for selected parameters as well as the overall $\mathcal{I}$ value for default and correlated data for $N \in \{50, 200\}$. We only display one component of β since the $\mathcal{I}$ values followed very similar distributions for all components.



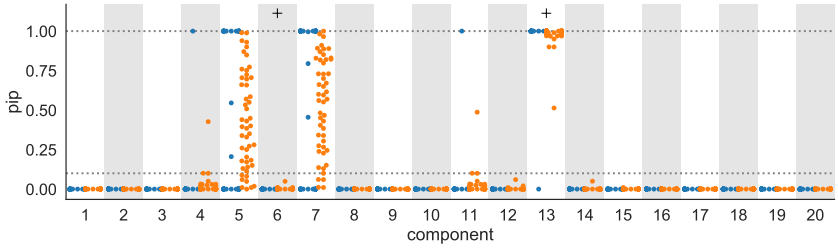
(a) $N = 10^2$



(b) $N = 10^3$

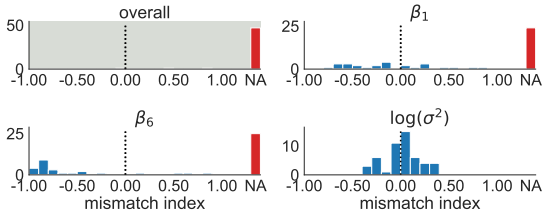


(c) $N = 10^4$

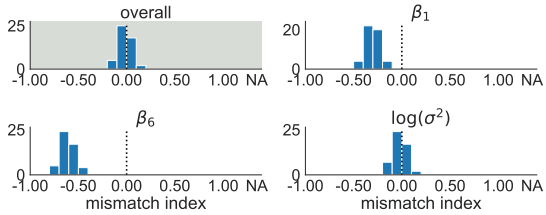


(d) $N = 10^5$

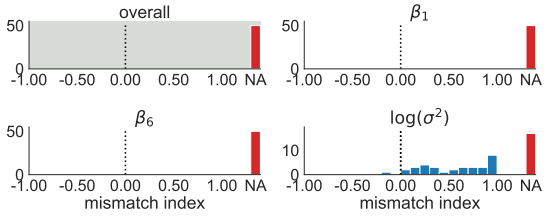
Fig C.2: Posterior inclusion probabilities (pips) for misspecified correlated-2-sparse-nonlinear data. See caption for Fig. 7 for further explanation.



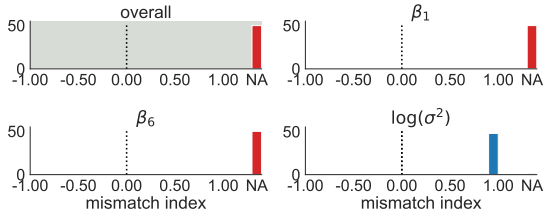
(a) correlated-1-sparse, $N = 5 \times 10^1$



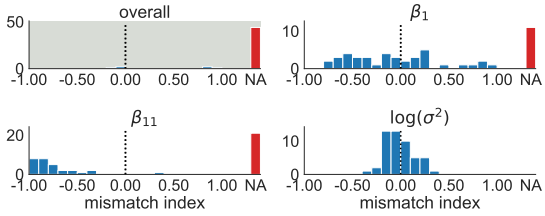
(b) correlated-1-sparse, $N = 5 \times 10^3$



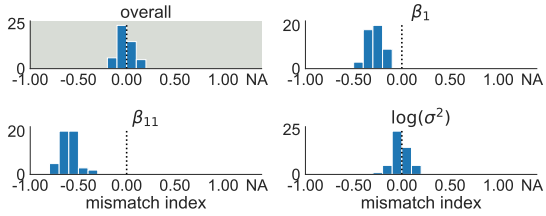
(c) correlated-1-sparse-nonlinear, $N = 5 \times 10^1$



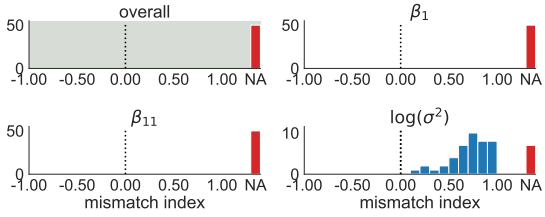
(d) correlated-1-sparse-nonlinear, $N = 5 \times 10^3$



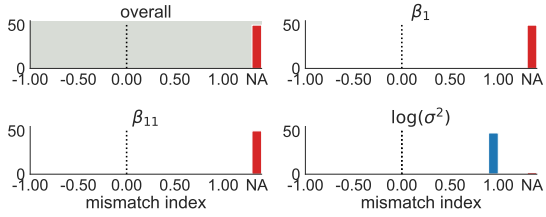
(e) correlated-2-sparse, $N = 10^2$



(f) correlated-2-sparse, $N = 10^4$



(g) correlated-2-sparse-nonlinear, $N = 10^2$



(h) correlated-2-sparse-nonlinear, $N = 10^4$

Fig C.3: Model–data mismatch indices $\mathcal{I}$ for selected parameters as well as the overall $\mathcal{I}$ value. We only display two components of β since the $\mathcal{I}$ values follow fairly similar distributions for all components.

TABLE C.1

Overlap between the posteriors for each pair of models, given the whale dataset, when using standard Bayes or BayesBag. The "mass" column shows the overlap of the 99% high-probability regions and "# trees" shows the number of trees in common. For BayesBag, "mass" shows 80% confidence intervals for the overlaps and "# trees" shows the median number of common trees.

Comparison			Standard		BayesBag	
			mass	# trees	mass (80% CI)	# trees
JC	vs	HKY	0.1%	1	(30%, 41%)	4
JC	vs	GTR	0%	0	(38%, 50%)	3
JC	vs	mixed	0%	0	(39%, 52%)	4
JC	vs	mtmam	0%	0	(3%, 8%)	3
HKY	vs	GTR	29%	1	(38%, 50%)	10
HKY	vs	mixed	30%	2	(46%, 58%)	10
HKY	vs	mtmam	57%	2	(32%, 43%)	8
GTR	vs	mixed	98%	1	(79%, 90%)	11
GTR	vs	mtmam	0%	0	(9%, 16%)	8
mixed	vs	mtmam	0%	0	(13%, 22%)	9

TABLE C.2

Overlap between the standard posterior for the mixed model and the bagged posterior for each model, given the whale dataset. The form of each data entry and the BayesBag parameters are the same as Table C.1.

Model	Standard		BayesBag	
	mass	# trees	mass (80% CI)	# trees
JC	0%	0	(28%, 39%)	2
HKY	30%	1	(21%, 32%)	2
GTR	98%	1	(64%, 74%)	2
mixed	99%	2	(56%, 66%)	2
mtmam	0%	0	(0.5%, 0.5%)	1

TABLE C.3

Comparison of self-consistency on whale dataset. The form of each data entry and the BayesBag parameters are the same as Table C.1.

Model	Comparison	Standard		BayesBag	
		mass	# trees	mass (80% CI)	# trees
JC	S1 vs S2	0%	0	(21%, 32%)	4
	all vs S1	0%	0	(24%, 36%)	5
	all vs S2	97%	1	(51%, 65%)	4
HKY	S1 vs S2	8%	3	(26%, 35%)	9
	all vs S1	12%	4	(44%, 56%)	9
	all vs S2	40%	4	(51%, 62%)	11
GTR	S1 vs S2	38%	2	(36%, 44%)	14
	all vs S1	38%	1	(44%, 54%)	12
	all vs S2	90%	1	(57%, 68%)	11
mtmam	S1 vs S2	0%	0	(0.3%, 2%)	9
	all vs S1	0.1%	1	(10%, 16%)	24
	all vs S2	28%	4	(17%, 25%)	24

APPENDIX D: COMPUTATION

When the posterior can be computed in closed form, using BayesBag is straightforward. If, however, approximate sampling methods such as Markov chain Monte Carlo are necessary, the computational cost could become substantial. In such cases we propose the basic scheme described in Algorithm 1, although more advanced approaches could also be developed. In short, the idea is to run a single long chain (or set of chains) on the standard posterior, then use the sampler hyperparameters and posterior samples to initialize shorter chains that sample from many different bootstrap datasets. Our proposed algorithm facilitates the use of $\hat{M}_{\text{opt}}$ and $\mathcal{I}$ since it outputs samples from the standard posterior as well as the bagged posterior.

If the approximation of $\pi(\theta \,|\, x^*_{(b)})$ is not very accurate (e.g., because it requires a time-consuming Markov chain Monte Carlo run), then we face a tradeoff between the error due to approximating each $\pi(\theta \,|\, x^*_{(b)})$ and the Monte Carlo error due to the BayesBag approximation given in Eq. (2). When using Markov chain Monte Carlo, we recommend assessing on how accurate different length Markov chains are likely to be by running long chains for the standard posterior, then using this information to decide on the best trade off between the length of the Markov chains and number of bootstrap datasets. Such an approach should not result in much wasted computation since we suggest obtaining a high-quality approximation to the standard posterior no matter what, as this permits computing quantities such as the optimal bootstrap sample size and the model–data mismatch index.

Algorithm 1 Basic BayesBag Sampler

Require: A Markov chain Monte Carlo procedure MCMC(x, T, θ_{init}, β_{init}) that returns adapted sampler hyperparameters and T approximate samples from $\Pi(\cdot \,|\, x)$, with the sampler initialized at θ_{init} with hyperparameters β_{init}

Require: Data x, "large" sample number T, "small" sample number T^*, number of bootstrap datasets B, initial hyperparameters β_{init}

$\beta, \theta_{1:T} \leftarrow$ MCMC(x, T, β_{init})
for $b = 1, \ldots, B$ **do**
    Generate a new bootstrap dataset $x^*_{(b)}$ of size M from x
    Sample $\theta^*_{(b)\text{init}}$ uniformly from $\theta_{1:T}$
    $\beta_{(b)}, \theta^*_{(b)1:T^*} \leftarrow$ MCMC($x^*_{(b)}$, T^*, $\theta^*_{(b)\text{init}}$, β)
end for
$\theta^*_{1:BT^*} \leftarrow$ concatenate$(\theta^*_{(1)1:T^*}, \ldots, \theta^*_{(B)1:T^*})$
return posterior samples $\theta_{1:T}$ and BayesBag samples $\theta^*_{1:BT^*}$

APPENDIX E: FEATURE SELECTION IN LINEAR REGRESSION

We derive the KL-optimal linear regression parameters for data generated as in our simulation studies (Section 5.1). In particular, we show that when the model is misspecified, even if the "causal" regression coefficients are sparse, the Kullback–Leibler (KL)-optimal regression coefficients may not be. Assuming a linear regression model and that the data follow Eq. (16), we have

$$\begin{aligned}
&\mathbb{E}\{\log p(Y_n \,|\, Z_n, \beta, \sigma^2)\} \\
&= -\frac{1}{2\sigma^2}\mathbb{E}\{(Y_n - Z_n^\top \beta)^2\} - \frac{1}{2}\log(2\pi\sigma^2) \\
&= -\frac{1}{2\sigma^2}\mathbb{E}\{(f(Z_n)^\top \beta_\dagger + \epsilon_n - Z_n^\top \beta)^2\} - \frac{1}{2}\log(2\pi\sigma^2) \\
&= -\frac{1}{2\sigma^2}\mathbb{E}\{\beta_\dagger^\top f(Z_n) f(Z_n)^\top \beta_\dagger + \beta^\top Z_n Z_n^\top \beta - 2\beta_\dagger^\top f(Z_n) Z_n^\top \beta\} \\
&\quad - \frac{1}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2).
\end{aligned}$$

Thus,

$$\sigma^2 \nabla_\beta \mathbb{E}\{\log p(Y_n \,|\, Z_n, \beta, \sigma^2)\} = -\mathbb{E}(Z_n Z_n^\top)\beta + \mathbb{E}\{Z_n f(Z_n)^\top\}\beta_\dagger,$$

so the optimal coefficient vector is

$$\beta_\circ = \mathbb{E}(Z_n Z_n^\top)^{-1}\mathbb{E}\{Z_n f(Z_n)^\top\}\beta_\dagger.$$

Thus, when f is not the identity and the regressors are no independent, in general $\beta_\circ$ will be dense even if $\beta_\dagger$ is sparse.

Let $\Sigma_{ZZ} := \mathbb{E}(Z_n Z_n^\top)$, $\Sigma_{Zf} := \mathbb{E}\{Z_n f(Z_n)^\top\}$, and $\Sigma_{ff} := \mathbb{E}\{f(Z_n) f(Z_n)^\top\}$. For the optimal coefficient vector, we have

$$\begin{aligned}
&\mathbb{E}\{\log p(Y_n \,|\, Z_n, \beta_\circ, \sigma^2)\} \\
&= -\frac{1}{2\sigma^2}\left[\beta_\dagger^\top \Sigma_{ff}\beta_\dagger + \beta_\dagger^\top \Sigma_{Zf}^\top \Sigma_{ZZ}^{-1}\Sigma_{Zf}\beta_\dagger - 2\beta_\dagger^\top \Sigma_{Zf}\Sigma_{ZZ}^{-1}\Sigma_{Zf}\beta_\dagger\right] \\
&\quad - \frac{1}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2). \\
&= -\frac{1}{2\sigma^2}\beta_\dagger^\top \Sigma_{ff}\beta_\dagger + \frac{1}{2\sigma^2}\beta_\dagger^\top \Sigma_{Zf}^\top \Sigma_{ZZ}^{-1}\Sigma_{Zf}\beta_\dagger - \frac{1}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2).
\end{aligned}$$

Thus, the optimal variance is

$$\sigma_\circ^2 = \left(1 + \beta_\dagger^\top \Sigma_{ff}\beta_\dagger - \beta_\dagger^\top \Sigma_{Zf}^\top \Sigma_{ZZ}^{-1}\Sigma_{Zf}\beta_\dagger\right)_+.$$

Now plugging in the optimal variance, we have

$$\begin{aligned}
&\mathbb{E}\{\log p(Y_n \,|\, Z_n, \beta_\circ, \sigma_\circ^2)\} \\
&= \begin{cases} -\log(2e\pi) - \log\left(1 + \beta_\dagger^\top \Sigma_{ff}\beta_\dagger - \beta_\dagger^\top \Sigma_{Zf}^\top \Sigma_{ZZ}^{-1}\Sigma_{Zf}\beta_\dagger\right) & \sigma_\circ^2 > 0 \\ \infty & \sigma_\circ^2 = 0. \end{cases}
\end{aligned}$$

APPENDIX F: DERIVATION OF FINITE-SAMPLE OPTIMAL BOOTSTRAP SIZE

We conclude with a derivation of the finite-sample optimal bootstrap sample size estimator $\hat{M}_{\text{fs,opt}}$. Recall that a potential shortcoming of using $\hat{M}_{\infty,\text{opt}}$ to choose the optimal bootstrap sample size is that it does not account for the influence of the prior. If the prior remains influential, then $v_N^* - v_N$ may be deceptively small, leading $\hat{M}_{\infty,\text{opt}}$ to be too large and the resulting bagged posterior to be overconfident. To account for the effect of the prior, we can instead use Eq. (6). Define $\sigma_\circ^2$ and $s_\circ^2$ as in Section 2.3.2. Eq. (6) yields the following more refined approximation $v_M^* \approx (R_M \sigma_\circ^2 + R_M^2 s_\circ^2)/M$ to the bagged posterior variance. Note that since $\sigma_\circ^2$ now plays the role that V played in the Gaussian location model, $R_M = (1 + \sigma_\circ^2 v_0^{-1}/M)^{-1}$. Since v_M^* needs to be approximately $s_\circ^2/N$ in order to be well-calibrated, we set $(R_M \sigma_\circ^2 + R_M^2 s_\circ^2)/M = s_\circ^2/N$ and solve for M, which yields

$$M_{\text{fs,opt}} = \frac{N}{2} + \frac{N\sigma_\circ^2}{2s_\circ^2} - \frac{\sigma_\circ^2}{v_0} + \left\{ \left(\frac{N}{2} + \frac{N\sigma_\circ^2}{2s_\circ^2} \right)^2 - \frac{N\sigma_\circ^2}{v_0} \right\}^{1/2}.$$

It remains to derive the estimators for $\sigma_\circ^2$ and $s_\circ^2$. Solving $v_N = \sigma_\circ^2 v_0/(N v_0 + \sigma_\circ^2)$ for $\sigma_\circ^2$ yields the finite-sample estimator $\hat{\sigma}_\circ^2 = N v_0 v_N/(v_0 - v_N)$ and plugging $\hat{\sigma}_\circ^2$ into the definition of R_N yields the estimator $\hat{R}_N = 1 - v_N/v_0$. Combining these yields the finite-sample estimator

$$\hat{s}_\circ^2 := \frac{v_0^2}{(v_0 - v_N)^2}(v_N^* - v_N)N.$$

Observe that we recover the asymptotic versions of the variance estimators and $M_{\infty,\text{opt}}$ by taking $v_0 \to \infty$.

APPENDIX G: PROOFS

NOTATION. The characteristic function of a distribution η on $\mathbb{R}^K$ is denoted $\psi_\eta(t) := \int \exp(it^\top x)\eta(\mathrm{d}x)$ for $t \in \mathbb{R}^K$. We use $\xrightarrow{P}$ to denote convergence in probability and $\xrightarrow{P_+}$ to denote convergence in outer probability.

G.1. Proof of Theorem 3.1. We use the classical characteristic function approach to proving central limit theorems. For $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, the characteristic function of $\mathcal{N}(\mu, \sigma^2)$ is

$$\psi_{\mathcal{N}(\mu,\sigma^2)}(t) = \exp(i\mu t - \sigma^2 t^2/2), \qquad t \in \mathbb{R}. \tag{18}$$

For $L \in \mathbb{N}$ and $p_1, \ldots, p_K \geq 0$ with $\sum_{k=1}^K p_k = 1$, the characteristic function of the multinomial distribution $\mathsf{Multi}(L, p)$ is

$$\psi_{\mathsf{Multi}(L,p)}(t) = \left(\sum_{k=1}^K p_k e^{it_k} \right)^L, \qquad t \in \mathbb{R}^K. \tag{19}$$

Let $\widetilde{\Pi}(\cdot \,|\, X_{1:M}^*) := \mathcal{N}(N^{1/2} R_M(\bar{X}_M^* - \bar{X}_N), NV_M)$, noting that this is the distribution of $N^{1/2}\{\vartheta^* - \mathbb{E}(\vartheta^* \,|\, X_{1:N})\} \,|\, X_{1:M}^*$. Similarly, let $\widetilde{\Pi}^*(\cdot \,|\, X_{1:N})$ denote the distribution

of $N^{1/2}\{\vartheta^* - \mathbb{E}(\vartheta^* \,|\, X_{1:N})\} \,|\, X_{1:N}$. Let $Y_{Nn} := N^{1/2} R_M (X_n - \bar{X}_N)$ and let $K_{1:N} \sim \text{Multi}(M, 1/N)$. Using Eqs. (18) and (19), we have

$$\begin{aligned}
\psi_{\widetilde{\Pi}^*(\cdot \,|\, X_{1:N})}(t) &= \mathbb{E}\{\psi_{\widetilde{\Pi}(\cdot \,|\, X^*_{1:M})}(t) \,|\, X_{1:N}\} \\
&= \mathbb{E}\big[\exp\big\{itM^{-1}\textstyle\sum_{n=1}^N K_n Y_{Nn} - NV_M t^2/2\big\} \,|\, X_{1:N}\big] \\
&= \left\{\frac{1}{N}\sum_{n=1}^N \exp(itM^{-1}Y_{Nn})\right\}^M \exp(-NV_M t^2/2). \qquad (20)
\end{aligned}$$

Let $\hat{V}_N := N^{-1}\sum_{n=1}^N (X_n - \bar{X}_N)^2$. By Taylor's theorem, $e^{is} = 1 + is - s^2/2 + \mathcal{R}(s)$ where $\mathcal{R}(s) \leq |s|^3/3$. Since $N^{-1}\sum_{n=1}^N Y_{Nn} = 0$, the first factor of Eq. (20) can be expanded as

$$\begin{aligned}
&\left\{\frac{1}{N}\sum_{n=1}^N \left(1 + itM^{-1}Y_{Nn} - \tfrac{1}{2}t^2 M^{-2} Y_{Nn}^2 + \mathcal{R}(tM^{-1}Y_{Nn})\right)\right\}^M \\
&= \left\{1 - \tfrac{1}{2}t^2 \frac{NR_M^2}{M^2}\hat{V}_N + \frac{1}{N}\sum_{n=1}^N \mathcal{R}(tM^{-1}Y_{Nn})\right\}^M . \qquad (21)
\end{aligned}$$

Since $\mathcal{R}(s) \leq |s|^3/3$,

$$\sum_{n=1}^N \mathcal{R}(tM^{-1}Y_{Nn}) \leq \frac{t^3 N^{3/2} R_M^3}{3M^3} \sum_{n=1}^N |X_n - \bar{X}_N|^3.$$

Using $|X_n - \bar{X}_N|^3 \leq |X_n|^3 + 3|X_n|^2|\bar{X}_N| + 3|X_n||\bar{X}_N|^2 + |\bar{X}_N|^3$, and applying the the strong law of large numbers to each factor, we have

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^N |X_n - \bar{X}_N|^3 \overset{a.s.}{<} \infty.$$

Hence, almost surely, for all $t \in \mathbb{R}$, $\sum_{n=1}^N \mathcal{R}(tM^{-1}Y_{Nn}) \to 0$ as $N \to \infty$. Further, note that $M/N \to c$, $R_M \to 1$, and $\hat{V}_N \overset{a.s.}{\to} \text{Var}(X_1)$ as $N \to \infty$. Now, we use the fact that if $a_N \to a$ and $c_N \to c$, then $(1 + a_N/N)^{Nc_N} \to \exp(a)^c$. Thus, almost surely, for all t, Eq. (21) converges to $\exp(-\frac{1}{2}t^2 \text{Var}(X_1)/c)$. Combining this with Eq. (20), and noting that $NV_M \to V/c$, we have that almost surely, for all $t \in \mathbb{R}$,

$$\psi_{\widetilde{\Pi}^*(\cdot \,|\, X_{1:N})}(t) \to \exp(-\tfrac{1}{2}t^2(\text{Var}(X_1)/c + V/c)).$$

The result follows by Lévy's continuity theorem (Kallenberg, 2002, Theorem 5.3).

G.2. Proof of Theorem 3.2. To de-clutter the notation, we abbreviate $J_\circ := J_{\theta_\circ}$, $I_\circ := I_{\theta_\circ}$, and $\dot{\ell}_\circ := \dot{\ell}_{\theta_\circ}$. Define

$$\begin{aligned}
\mathbb{P}^*_N &:= M^{-1}\sum_{n=1}^N K_n \delta_{X_n}, \\
\Delta^*_N &:= N^{1/2} J_\circ^{-1}(\mathbb{P}^*_N - \mathbb{P}_N)\dot{\ell}_{\theta_\circ},
\end{aligned}$$

the empirical process $\mathbb{G}_N = N^{1/2}(\mathbb{P}_N - P_\circ)$, and the bootstrap empirical process $\mathbb{G}^*_N = M^{1/2}(\mathbb{P}^*_N - \mathbb{P}_N)$. The conditions of van der Vaart (1998, Lemma 19.31) hold by assumption, so for any sequence $h_1, h_2, \ldots \in \mathbb{R}^D$ bounded in probability,

$$\mathbb{G}_N\{N^{1/2}\lambda_N - h_N^\top \dot{\ell}_\circ\} \xrightarrow{P} 0,$$

where $\lambda_N = \ell_{\theta_\circ + h_N/N^{1/2}} - \ell_{\theta_\circ}$. By van der Vaart and Wellner (1996, Theorem 3.6.3), for almost every $X_{1:\infty}$, conditional on $X_{1:\infty}$, $\mathbb{G}^*_N$ and $\mathbb{G}_N$ both converge weakly to the same limiting process. For the remainder of the proof we condition on $X_{1:\infty}$, so all statements will hold for almost every $X_{1:\infty}$. It follows that

$$\mathbb{G}^*_N\{N^{1/2}\lambda_N - h_N^\top \dot{\ell}_\circ\} \xrightarrow{P_+} 0. \tag{22}$$

By the proof of Kleijn and van der Vaart (2012, Lemma 2.1),

$$|N\mathbb{P}_N\lambda_N - \mathbb{G}_N h_N^\top \dot{\ell}_\circ - \tfrac{1}{2}h_N^\top J_\circ h_N| \xrightarrow{P_+} 0$$

and, following the same reasoning, we can expand the lefthand side of Eq. (22) and multiply though by $c^{1/2}$ to get

$$c^{1/2}(NM)^{1/2}\mathbb{P}^*_N\lambda_N - c^{1/2}\mathbb{G}^*_N h_N^\top \dot{\ell}_\circ - c^{1/2}(NM)^{1/2}\mathbb{P}_N\lambda_N \xrightarrow{P_+} 0$$

and hence

$$M\mathbb{P}^*_N\lambda_N - (c^{1/2}\mathbb{G}^*_N + c\mathbb{G}_N)h_N^\top \dot{\ell}_\circ - \tfrac{1}{2}h_N^\top(cJ_\circ)h_N \xrightarrow{P_+} 0.$$

Since $c\,\mathbb{G}_N h_N^\top \dot{\ell}_\circ = h_N^\top(cJ_\circ)\Delta_N$ and $c^{1/2}\mathbb{G}^*_N h_N^\top \dot{\ell}_\circ (cN/M)^{1/2} = h_N^\top(cJ_\circ)\Delta^*_N$ by the definitions of Δ_N and Δ^*_N, it follows that for every compact $K \subset \Theta$,

$$\sup_{h\in K}\left|M\mathbb{P}^*_N(\ell_{\theta_\circ + h/N^{1/2}} - \ell_{\theta_\circ}) - h^\top(cJ_\circ)(\Delta_N + \Delta^*_N) - \tfrac{1}{2}h^\top(cJ_\circ)h\right| \xrightarrow{P_+} 0.$$

We apply Kleijn and van der Vaart (2012, Theorem 2.1) to conclude that, letting $\vartheta^{*\prime}|X^*_{1:M} \sim \Pi(\cdot\,|\,X^*_{1:M})$, the total variation distance between the distribution of $N^{1/2}(\vartheta^{*\prime} - \theta_\circ)\,|\,X^*_{1:M}$ and $\mathcal{N}(\Delta_N + \Delta^*_N, J_\circ^{-1}/c)$ converges to zero in outer probability. Compared to the notation of Kleijn and van der Vaart (2012, Theorem 2.1), we have $X^*_{1:M}$ in place of $X^{(n)}$, $\mathbb{P}^M_N$ in place of $P_0^{(n)}$, $cJ_\circ$ in place of V_{θ^*}, and $\Delta_N + \Delta^*_N$ in place of Δ_{n,θ^*}. Hence, uniformly in $t \in \mathbb{R}^D$, the absolute difference in their characteristic functions must also converge to zero in outer probability. Let $\epsilon_N(t)$ (and similarly $\bar{\epsilon}_N(t)$) denote a function that satisfies $\limsup_{N\to\infty}\sup_{t\in\mathbb{R}}\epsilon_N(t) = 0$. We can therefore write the characteristic function of $N^{1/2}(\vartheta^* - \theta_\circ) - \Delta_N \,|\, X_{1:N}$ evaluated at $t \in \mathbb{R}^D$ as

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{i\Delta^{*\top}_N t - t^\top J_\circ^{-1}t/(2c)\right\}\,|\,X_{1:N}\right] + \epsilon_N(t) \\
&= \mathbb{E}\left[\exp\left\{iN^{1/2}\mathbb{P}^*_N\dot{\ell}_\circ^\top J_\circ^{-1}t\right\}\,\middle|\,X_{1:N}\right]\exp\left\{-iN^{1/2}\mathbb{P}_N\dot{\ell}_\circ^\top J_\circ^{-1}t\right\} \\
&\quad\times\exp\left\{-t^\top J_\circ^{-1}t/(2c)\right\} + \epsilon_N(t).
\end{aligned} \tag{23}$$

Letting $\delta\dot{\ell}_\circ(X_n) := \dot{\ell}_\circ(X_n) - \mathbb{P}_N\dot{\ell}_\circ$, we can further expand the first line of Eq. (23) to get

$$\begin{aligned}
&\mathbb{E}\left[\exp\left\{iN^{1/2}M^{-1}\sum_{n=1}^{N}K_n\dot{\ell}_\circ(X_n)^\top J_\circ^{-1}t\right\}\,\middle|\, X_{1:N}\right]\exp\left\{-iN^{1/2}\mathbb{P}_N\dot{\ell}_\circ^\top J_\circ^{-1}t\right\}\\
&= \left[\frac{1}{N}\sum_{n=1}^{N}\exp\left\{\frac{iN^{1/2}\dot{\ell}_\circ(X_n)^\top J_\circ^{-1}t}{M}\right\}\right]^M\exp\left\{-iN^{1/2}\mathbb{P}_N\dot{\ell}_\circ^\top J_\circ^{-1}t\right\}\\
&= \left[\frac{1}{N}\sum_{n=1}^{N}\exp\left\{\frac{iN^{1/2}\delta\dot{\ell}_\circ(X_n)^\top J_\circ^{-1}t}{M}\right\}\right]^M\\
&= \left[\frac{1}{N}\sum_{n=1}^{N}\left\{1+\frac{iN^{1/2}\delta\dot{\ell}_\circ(X_n)^\top J_\circ^{-1}t}{M}-\frac{N(\delta\dot{\ell}_\circ(X_n)^\top J_\circ^{-1}t)^2}{2M^2}+\mathcal{R}_n\right\}\right]^M\\
&= \left\{1-\frac{Nt^\top J_\circ^{-1}\mathbb{P}_N(\delta\dot{\ell}_\circ\delta\dot{\ell}_\circ^\top)J_\circ^{-1}t}{2M^2}+\mathcal{R}_n\right\}^M, \qquad (24)
\end{aligned}$$

where (recalling the notation from the proof of Theorem 3.1)

$$\mathcal{R}_n := \mathcal{R}\left(\frac{iN^{1/2}\delta\dot{\ell}_\circ(X_n)^\top J_\circ^{-1}t}{M}\right).$$

Arguing as in the proof of Theorem 3.1 and using assumption (ii), we conclude that $\lim_{N\to\infty}\sum_{n=1}^{N}\mathcal{R}_n = 0$.

Note that $M/N \to c$, and $\mathbb{P}_N(\delta\dot{\ell}_\circ\delta\dot{\ell}_\circ^\top) \overset{a.s.}{\to} I_\circ$ as $N\to\infty$. Now, we use the fact that if $a_N \to a$ and $c_N \to c$, then $(1+a_N/N)^{Nc_N} \to \exp(a)^c$. Combining all these observations with Eqs. (23) and (24), we have that, for all $t\in\mathbb{R}^D$, the characteristic function of $N^{1/2}(\vartheta^* - \theta_\circ) \mid X_{1:N}$ evaluated at t is

$$\exp\left\{i\Delta_N^\top t - t^\top J_\circ^{-1}t/(2c) - t^\top J_\circ^{-1}I_\circ J_\circ^{-1}t/(2c)\right\} + \epsilon_N(t) + \bar{\epsilon}_N(t).$$

The result follows from Lévy's continuity theorem (Kallenberg, 2002, Theorem 5.3).

G.3. Proof of Theorem 4.1. We first prove a simple uniform central limit theorem that is needed for our proof of Theorem 4.1. For a random variable ξ, let $\mathcal{L}(\xi)$ denote its law. For real-valued random variables ξ, ξ', let $d_K(\mathcal{L}(\xi),\mathcal{L}(\xi')) := \sup_{t\in\mathbb{R}}|\mathbb{P}(\xi\le t) - \mathbb{P}(\xi'\le t)|$ denote the Kolmogorov distance.

PROPOSITION G.1. *For a triangular array $\xi_{Nn}\sim P_N$ $(N=1,2,\ldots;n=1,\ldots,N)$ of independent random variables, if (i) $N^{1/2}\mathbb{E}(\xi_{N1})\to\mu\in\mathbb{R}$ as $N\to\infty$, (ii) $\mathrm{Var}(\xi_{N1}) = \sigma^2\in(0,\infty)$ for all N, and (iii) $\limsup_{N\to\infty}\mathbb{E}\{|\xi_{N1}-\mathbb{E}(\xi_{N1})|^{2+\varepsilon}\}<\infty$ for some $\varepsilon>0$, then $W_N := N^{-1/2}\sum_{n=1}^{N}\xi_{Nn}$ satisfies*

$$\lim_{N\to\infty} d_K(\mathcal{L}(W_N),\mathcal{N}(\mu,\sigma^2)) = 0.$$

PROOF. Let $\tilde{\xi}_{Nn} := \xi_{Nn} - \mathbb{E}(\xi_{Nn})$ and $\tilde{W}_N := N^{-1/2}\sum_{n=1}^N \tilde{\xi}_{Nn}$. By Chen, Goldstein and Shao (2010, Thm. 3.2, Thm. 3.3, Eq. 3.14), for any $\alpha \in (0,1)$,

$$d_K(\mathcal{L}(\tilde{W}_N), \mathcal{N}(0,\sigma^2)) \leq 4\left(\sigma^{-2}\mathbb{E}\{|\tilde{\xi}_{Nn}|^2 \mathbb{1}(|\tilde{\xi}_{Nn}| > \alpha\sigma N^{1/2})\} + \alpha\right)^{1/2}.$$

Further,

$$\begin{aligned}
\mathbb{E}\left\{|\tilde{\xi}_{N1}|^2 \mathbb{1}(|\tilde{\xi}_{N1}| > \alpha\sigma N^{1/2})\right\} &\leq \mathbb{E}\{|\tilde{\xi}_{N1}|^{2+\varepsilon}\}^{2/(2+\varepsilon)}\, \mathbb{E}\{\mathbb{1}(|\tilde{\xi}_{N1}| > \alpha\sigma N^{1/2})\}^{\varepsilon/(2+\varepsilon)} \\
&= \mathbb{E}\{|\tilde{\xi}_{N1}|^{2+\varepsilon}\}^{2/(2+\varepsilon)}\, \mathbb{P}(|\tilde{\xi}_{N1}| > \alpha\sigma N^{1/2})^{\varepsilon/(2+\varepsilon)} \\
&\leq \mathbb{E}\{|\tilde{\xi}_{N1}|^{2+\varepsilon}\}^{2/(2+\varepsilon)}\, \left(\alpha^2 N\right)^{-\varepsilon/(2+\varepsilon)} \xrightarrow[N\to\infty]{} 0,
\end{aligned}$$

where we have used Hölder's inequality, Chebyshev's inequality, assumption (ii), and assumption (iii). Since we can make α arbitrarily small, we have

$$\lim_{N\to\infty} d_K(\mathcal{L}(\tilde{W}_N), \mathcal{N}(0,\sigma^2)) = 0. \tag{25}$$

Since the cumulative distribution function of $Z \sim \mathcal{N}(\mu,\sigma^2)$ is Lipschitz for some constant $C > 0$, $|\mathbb{P}(Z < t) - \mathbb{P}(Z < s)| \leq C|t-s|$. Let $\tilde{Z} := Z - \mu \sim \mathcal{N}(0,\sigma^2)$ and note that $W_N = \tilde{W}_N + \mu_N$ where $\mu_N := N^{1/2}\mathbb{E}(\xi_{N1})$. Thus, for all $t \in \mathbb{R}$, letting $\tilde{t} := t - \mu_N$, we have

$$\begin{aligned}
&|\mathbb{P}(W_N < t) - \mathbb{P}(Z < t)| \\
&= |\mathbb{P}(\tilde{W}_N + \mu_N < \tilde{t} + \mu_N) - \mathbb{P}(\tilde{Z} + \mu < \tilde{t} + \mu_N)| \\
&= |\mathbb{P}(\tilde{W}_N < \tilde{t}) - \mathbb{P}(\tilde{Z} < \tilde{t} + \mu_N - \mu)| \\
&\leq |\mathbb{P}(\tilde{W}_N < \tilde{t}) - \mathbb{P}(\tilde{Z} < \tilde{t})| + |\mathbb{P}(\tilde{Z} < \tilde{t}) - \mathbb{P}(\tilde{Z} < \tilde{t} + \mu_N - \mu)| \\
&\leq |\mathbb{P}(\tilde{W}_N < \tilde{t}) - \mathbb{P}(\tilde{Z} < \tilde{t})| + C|\mu_N - \mu|.
\end{aligned}$$

By Eq. (25), the previous display, and assumption (i), it follows that $\sup_t |\mathbb{P}(W_N < t) - \mathbb{P}(Z < t)| \to 0$ as $N \to \infty$.

□

Proof of Theorem 4.1, part (1). Let $Z_{N0} := \log Q_0(1) - \log Q_0(2)$ denote the log prior ratio, let $W_N := N^{-1/2}\sum_{n=0}^N Z_{Nn}$, and let $W_\infty \sim \mathcal{N}(\mu_\infty, \sigma_\infty^2)$. It follows from Proposition G.1 with $\xi_{Nn} := Z_{Nn} + Z_{N0}/N$ that

$$\lim_{N\to\infty} d_K(\mathcal{L}(W_N), \mathcal{L}(W_\infty)) = 0, \tag{26}$$

where the Minkowski inequality and assumption (iii) of Theorem 4.1 verify assumption (iii) of Proposition G.1. In particular, Eq. (26) implies that $W_N \xrightarrow{\mathcal{D}} W_\infty$.

Letting $\phi_N(t) = \{1 + \exp(-N^{1/2}t)\}^{-1}$, we can write the posterior probability of model 1 as $Q(1 \mid X_{1:N}) = \phi_N(W_N)$. Since $\phi_N(t) \to \mathbb{1}(t > 0)$ pointwise for $t \neq 0$, it follows from the continuous mapping theorem (Kallenberg, 2002, Theorem 4.27) that $\phi_N(W_N) \xrightarrow{\mathcal{D}} \mathbb{1}(W_\infty > 0)$. Since $\mathbb{1}(W_\infty > 0) \sim \mathsf{Bern}(\Phi(\mu_\infty/\sigma_\infty))$, we have $Q(1 \mid X_{1:N}) \xrightarrow{\mathcal{D}} \mathsf{Bern}(\Phi(\mu_\infty/\sigma_\infty))$.

Proof of Theorem 4.1, parts (2) and (3). Let

$$W_N^* := M^{-1/2}\left(Z_{N0} + \sum_{n=1}^{N} K_{Nn} Z_{Nn}\right),$$

where $K_{N,1:N} \sim \text{Multi}(M, 1/N)$ is independent of $(X_1, X_2, \ldots)$. Furthermore, let $\Delta_N^* := W_N^* - (M/N)^{1/2} W_N$ and, independently of $(X_1, X_2, \ldots)$, let $\Delta_\infty^* \sim \mathcal{N}(0, \sigma_\infty^2)$. The implication (i) $\implies$ (iii) in Mammen (1992, Theorem 1) holds not only for $M = N$ but also, after the obvious rescaling, for the general $M(N)$ case as well when $\lim_{N\to\infty} M/N < \infty$. So, together with Eq. (26), we have that

$$d_K(\mathcal{L}(W_N - \mu_\infty), \mathcal{L}(\Delta_N^* \,|\, X_{1:N})) \xrightarrow{P} 0$$

and hence

$$\kappa_N^* := d_K(\mathcal{L}(\Delta_\infty^*), \mathcal{L}(\Delta_N^* \,|\, X_{1:N})) \xrightarrow{P} 0.$$

We can write the bagged posterior probability of model 1 as $Q^*(1 \,|\, X_{1:N}) = \mathbb{E}\{\phi_M(W_N^*) \,|\, X_{1:N}\} = \mathbb{E}\{\phi_M(\Delta_N^* + (M/N)^{1/2} W_N) \,|\, X_{1:N}\}$. Let $I_N = [-\epsilon_N, \epsilon_N]$ for $\epsilon_N = M^{-1/4}$. Since the density of Δ_∞^* is bounded by a constant b, it follows that for any $\alpha \in \mathbb{R}$,

$$\mathbb{P}(\Delta_N^* + \alpha \in I_N) \le \mathbb{P}(\Delta_\infty^* + \alpha \in I_N) + 2\kappa_N^* \le 2b\epsilon_N + 2\kappa_N^*.$$

Since $|\phi_M(t) - \mathbb{1}(t > 0)| \le \exp(-M^{1/2}|t|)$, for all $t \notin I_N$, $|\phi_M(t) - \mathbb{1}(t > 0)| \le \exp(-M^{1/2}\epsilon_N)$. We conclude that

$$\left|\mathbb{E}\{\phi_M(\Delta_N^* + (M/N)^{1/2} W_N) \,|\, X_{1:N}\} - \mathbb{E}\{\mathbb{1}(\Delta_N^* + (M/N)^{1/2} W_N > 0) \,|\, X_{1:N}\}\right|$$
$$\le \exp(-M^{1/2}\epsilon_N) + 2b\epsilon_N + 2\kappa_N^* = o_P(1).$$

Moreover,

$$\left|\mathbb{E}\{\mathbb{1}(\Delta_N^* + (M/N)^{1/2} W_N > 0) \,|\, X_{1:N}\}| - \mathbb{E}\{\mathbb{1}(\Delta_\infty^* + (M/N)^{1/2} W_N > 0) \,|\, X_{1:N}\}\right|$$
$$\le \kappa_N^* = o_P(1).$$

Combining the previous two displays, we have

$$\begin{aligned}
&\mathbb{E}\{\phi_M(\Delta_N^* + (M/N)^{1/2} W_N) \,|\, X_{1:N}\} \\
&= \mathbb{E}\{\mathbb{1}(\Delta_\infty^* + (M/N)^{1/2} W_N > 0) \,|\, X_{1:N}\} + o_P(1) \\
&= \Phi((M/N)^{1/2} W_N / \sigma_\infty) + o_P(1) \\
&\xrightarrow{\mathcal{D}} \Phi(c^{1/2} W_\infty / \sigma_\infty),
\end{aligned}$$

where the second equality follows from the definition of Δ_∞^*, and convergence in distribution follows from the assumption that $M/N \to c$, Eq. (26), and Slutsky's theorem.

If $c > 0$ then the cumulative distribution function of the random variable $U^* := \Phi(c^{1/2} W_\infty / \sigma_\infty)$ is given by $u \mapsto \Phi(c^{-1/2}\Phi^{-1}(u) - \mu_\infty/\sigma_\infty)$ for $u \in (0, 1)$, and differentiating, we find that the density of U^* is $u \mapsto \Phi'\big(c^{-1/2}\Phi^{-1}(u) - \mu_\infty/\sigma_\infty\big) c^{-1/2} / \Phi'(\Phi^{-1}(u))$. If $c = 0$, then we instead have that $Q^*(1 \,|\, X_{1:N}) \xrightarrow{\mathcal{D}} \Phi(0) = 1/2$, which implies convergence in probability.

G.4. Proof of Corollary 4.2. Note that $Q(1 \,|\, X_{1:N}) = \phi(\Lambda_{X_{1:N}})$ and $Q^*(1 \,|\, X_{1:N}) = \mathbb{E}\{\phi(\Lambda_{X^*_{1:M}}) \,|\, X_{1:N}\}$ where $\phi(t) = \{1 + \exp(-t)\}^{-1}$. We have the asymptotic expansion (Clarke and Barron, 1990; Dawid, 2011)

$$\Lambda_{X_{1:N}} = \frac{1}{2}(D_2 - D_1) \log N + \sum_{n=1}^{N} \log \frac{p_{\theta_{1\circ}}(X_n \,|\, 1)}{p_{\theta_{2\circ}}(X_n \,|\, 2)} + O_P(1).$$

Letting $Z_n := \log p_{\theta_{1\circ}}(X_n \,|\, 1) - \log p_{\theta_{2\circ}}(X_n \,|\, 2) = \ell_{1,\theta_{1\circ}}(X_n) - \ell_{2,\theta_{2\circ}}(X_n)$, the conclusions follow as in the proof of Theorem 4.1, although the argument is somewhat simplified by the fact that $X_1, X_2, \ldots$ i.i.d., so we do not need to reason about triangular arrays.