ROBUST AND REPRODUCIBLE MODEL SELECTION USING BAGGED POSTERIORS

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 University, jwmiller@hsph.harvard.edu*

Bayesian model selection is premised on the assumption that the data are generated from one of the postulated models, however, in many applications, all of these models are incorrect. When two or more models provide a nearly equally good fit to the data, Bayesian model selection can be highly unstable, potentially leading to self-contradictory findings. In this paper, we explore using bagging on the posterior distribution ("BayesBag") when performing model selection – that is, averaging the posterior model probabilities over many bootstrapped datasets. We provide theoretical results characterizing the asymptotic behavior of the standard posterior and the BayesBag posterior under misspecification, in the model selection setting. We empirically assess the BayesBag approach on synthetic and real-world data in (i) feature selection for linear regression and (ii) phylogenetic tree reconstruction. Our theory and experiments show that in the presence of misspecification, BayesBag provides (a) greater reproducibility and (b) greater accuracy in selecting the correct model, compared to the standard Bayesian posterior; on the other hand, under correct specification, BayesBag is slightly more conservative than the standard posterior. Overall, our results demonstrate that BayesBag provides an easy-to-use and widely applicable approach that improves upon standard Bayesian model selection by making it more stable and reproducible.

1. Introduction. In Bayesian statistics, the standard method of quantifying uncertainty in the choice of model is simply to use the posterior distribution over models. An implicit assumption of this approach is that one of the assumed models is exactly correct, but it is widely recognized that in practice, this assumption is typically unrealistic. When all of the models are incorrect, it is well known that the posterior concentrates on the model that provides the best fit in terms of Kullback–Leibler divergence. However, when two or more models can explain the data almost equally well, the posterior becomes unstable and can yield contradictory results when seemingly inconsequential changes are made to the models or to the data (Meng and Dunson, 2019; Oelrich et al., 2020; Yang and Zhu, 2018). For instance, as the size of the data set grows, the posterior probability of a given model may oscillate between values very close to 1 and very close to 0, *ad infinitum*. In short, standard Bayesian model selection can be unreliable and irreproducible.

This article develops the theory and practice of *BayesBag* for model selection, a simple and widely applicable approach to stabilizing Bayesian inferences. Originally suggested by Waddell et al. (2002) and Douady et al. (2003) in the context of phylogenetic inference and

Keywords and phrases: Asymptotics, Bagging, Bayesian model averaging, Bootstrap, Model misspecification, Stability

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. Let $Q(\mathfrak{m} \,|\, x) \propto p(x \,|\, \mathfrak{m}) Q_0(\mathfrak{m})$ denote the posterior probability of model $\mathfrak{m} \in \mathfrak{M}$ given data x, where $\mathfrak{M}$ is a finite or countably infinite set of models, $p(x \,|\, \mathfrak{m})$ is the marginal likelihood, and $Q_0(\mathfrak{m})$ is the prior probability. We define the *bagged posterior* $Q^*(\mathfrak{m} \,|\, x)$ by taking bootstrapped copies $x^* := (x_1^*, \ldots, x_M^*)$ of the original dataset $x := (x_1, \ldots, x_N)$ and averaging over the posteriors obtained by treating each bootstrap dataset as the observed data, that is,

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

where the sum is over all possible N^M bootstrap datasets of M samples drawn with replacement from the original dataset. In practice, we approximate $Q^*(\mathfrak{m} \,|\, x)$ 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

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

The BayesBag approach is to use $Q^*(\mathfrak{m} \,|\, x)$ to quantify uncertainty in the model $\mathfrak{m}$. BayesBag is easy to use since the bagged posterior model probability is simply an average over standard Bayesian model probabilities, which means no additional algorithmic tools are needed beyond what a data analyst would normally use for posterior inference. 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.

In previous work, we explored the benefits of using BayesBag for parameter estimation and prediction (Huggins and Miller, 2019). Surprisingly, despite its attractive features, there has been little practical or theoretical investigation of BayesBag prior to this. In the only prior work of which we are aware, in a short discussion paper Bühlmann (2014) presented only a few simulation results in a simple Gaussian location model, while Waddell et al. (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 versus the bootstrap (in the latter).

In this paper, we focus on the use of BayesBag in the model selection setting, which turns out to be significantly different than the parameter inference and prediction setting in terms of both theory and methodology. We find that in the presence of misspecification, model selection with the bagged posterior has appealing statistical properties while also being easy to use and computationally tractable on practical problems.

The paper is organized as follows. Section 2 provides an overview of our theory, methodology, and experiments. In Section 3, we present our theoretical results, illustrate the theory graphically, and discuss the use of BayesBag for model criticism. Section 4 contains a simulation study using BayesBag for feature selection in linear regression. In Section 5, we evaluate BayesBag on real-world data in applications involving (i) feature selection for linear regression and (ii) phylogenetic tree reconstruction.

2. Summary of results.

2.1. *Theory.* As first noted by Berk (1966), when the best fit to the data distribution is attained by more than one model, the posterior typically does not converge on a single model. For instance, consider the case of two distinct models, $\mathfrak{M} = \{1, 2\}$, and suppose the dataset is $X_{1:N} = (X_1, \ldots, X_N)$ where $X_1, X_2, \ldots$ are independent and identically distributed (i.i.d.) random variables. If both models are misspecified and $\lim_{N\to\infty} N^{-1/2}\mathbb{E}\{\log p(X_{1:N} \,|\, 1) - \log p(X_{1:N} \,|\, 2)\} = 0$, then the posterior mass on model 1 converges in distribution to a $\mathsf{Bern}(1/2)$ random variable (Theorem 3.1):

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

In other words, half of the time, model 1 has posterior probability ≈ 1, and the other half of the time, it has posterior probability ≈ 0. Since both models provide equally good approximations of the true data-generating distribution, asymptotically, the ideal behavior would be $Q(1 \,|\, X_{1:N}) \to 1/2$, but Eq. (3) describes the opposite behavior: a single model has posterior probability 1. Yang and Zhu (2018) prove a similar but more limited result.

BayesBag model selection is much better behaved and avoids this pathological behavior. As we show in Theorem 3.1, the bootstrap resampling stabilizes the model probabilities such that when $M = N$, the bagged posterior probability of model 1 converges in distribution to a uniform random variable on the interval from 0 to 1:

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

Moreover, if we choose $M \to \infty$ such that $M/N \to 0$, then the bagged posterior mass on model 1 has the ideal behavior of converging to 1/2:

$$Q^*(1 \,|\, X_{1:N}) \xrightarrow[N\to\infty]{\mathcal{D}} 1/2.$$

It is important to note that this is not simply due to the bagged posterior reverting to the prior; this result holds for any prior giving positive mass to both models. Theorem 3.1 establishes these results as well as the asymptotics of $Q^*(1 \,|\, X_{1:N})$ more generally, showing that the bagged posterior stabilizes model probabilities in the original spirit of bagging (Breiman, 1996; Bühlmann and Yu, 2002).

In practice, it seems implausible that two models would fit the true data-generating distribution *exactly* equally well. However, it turns out that even if model 1 has posterior probability tending to 1 asymptotically, for a finite sample size it can happen that $N^{-1/2}\mathbb{E}\{\log p(X_{1:N} \,|\, 1) - \log p(X_{1:N} \,|\, 2)\} \approx 0$, such that model 2 has posterior probability near 1 roughly half of the time. The analysis of Yang and Zhu (2018) was motivated by widespread observations of this type of phenomenon in Bayesian phylogenetic tree reconstruction (Alfaro et al., 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).

2.2. *Methodology.* BayesBag requires one to choose the bootstrap dataset size M and the number of bootstrap datasets B. In the model selection setting, our theoretical and empirical results indicate that $M = N$ is a good default choice that will still behave fairly well even in the extreme scenario of multiple models explaining the data-generating distribution equally well. Since the standard posterior is recovered in the limit as $M/N \to \infty$ (Theorem 3.1), if one has high confidence in the correctness of one model, then a value of $M > N$ could be used to make the bagged posterior less conservative. Meanwhile, a value of $M < N$ makes the bagged posterior more conservative. If a large amount of data is 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 as $1/4$ may be advisable to stabilize the results. It is interesting to note that $M = N$ provides a good default choice in both the model selection setting and the parameter inference and prediction setting of Huggins and Miller (2019), even though the theory is completely different.

The choice of B controls the accuracy of the Monte Carlo approximation to the bagged posterior; see Eqs. (1) and (2). Thus, it is straightforward to empirically estimate the error using the standard formula for the variance of a Monte Carlo approximation (Huggins and Miller, 2019). We have found $B \approx 50$–100 to be sufficient in all of the applications we have considered.

2.3. *Experiments.* We validate our theory and proposed methods through simulations on feature selection for linear regression, and we evaluate the performance of BayesBag on real-data applications involving feature selection and phylogenetic tree reconstruction. Overall, our empirical results demonstrate that in the presence of significant misspecification, the bagged posterior produces more stable inferences and selects correct models more often than the standard posterior; on the other hand, when one of the models is correctly specified, the bagged posterior is slightly more conservative than the standard posterior. Thus, BayesBag leads to more stable model selection results that are robust to minor changes in the model or representation of the data.

3. BayesBag for model selection. In this section, we present our theoretical results on BayesBag for model selection (Section 3.1), illustrate the theory with plots comparing the asymptotics of BayesBag versus the standard posterior (Section 3.2), and discuss the use of BayesBag for model criticism (Section 3.3).

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

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

where $\theta_\mathfrak{m} \in \Theta_\mathfrak{m}$ is an element of a model-specific parameter space with prior distribution $\Pi_0(\mathrm{d}\theta_\mathfrak{m} \,|\, \mathfrak{m})$. The posterior probability of $\mathfrak{m} \in \mathfrak{M}$ is $Q(\mathfrak{m} \,|\, X_{1:N}) \propto p(X_{1:N} \,|\, \mathfrak{m}) Q_0(\mathfrak{m})$.

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

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

Note that this is equivalent to the informal definition in Eq. (1).

We develop our asymptotic theory in the case of two models, $\mathfrak{M} = \{1, 2\}$. For the moment, we assume that 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 N, 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 3.2.) 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 main 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 3.1. *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* $\varepsilon > 0$,
(iv) $\lim_{N\to\infty} M = \infty$, *with* $M = M(N)$, *and*
(v) $c := \lim_{N\to\infty} M/N \in [0, \infty)$,

then

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

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

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 C.1. See Section 3.2 for a graphical illustration of the results in Theorem 3.1. An extension to the case of three or more models is an interesting direction for future work. We conjecture that the behavior of the standard and bagged posteriors is significantly more complicated in this case because there is dependence on both the correlation structure and the relative variances of the log likelihood ratios between each pair of models.

We now consider extending Theorem 3.1 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 $\theta_{\mathfrak{m}\circ} := \arg\max_{\theta_\mathfrak{m} \in \Theta_\mathfrak{m}} \mathbb{E}\{\ell_{\mathfrak{m},\theta_\mathfrak{m}}(X_1)\}$. The corollary requires further assumptions on the individual models. Toward this end, for arbitrary data x, let $\Lambda_x := \log p(x \mid 1)Q_0(1) - \log p(x \mid 2)Q_0(2)$. Also, let $X_{1:\infty}$ denote the infinite sequence of data $(X_1, X_2, \dots)$. We will assume that conditionally on $X_{1:\infty}$, for almost every $X_{1:\infty}$,

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

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

Denote the standard posterior distribution given $X_{1:N}$ and $\mathfrak{m}$ by

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

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

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

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

COROLLARY 3.2. *Let* $X_1, X_2, \dots$ *i.i.d.* $\sim P_\circ$ *and for* $\mathfrak{m} \in \mathfrak{M}$*, assume that:*

(i) $\theta_\mathfrak{m} \mapsto \ell_{\theta_\mathfrak{m}}(X_1)$ *is differentiable at* $\theta_{\mathfrak{m}\circ}$ *in probability;*
(ii) *there is an open neighborhood* U *of* $\theta_{\mathfrak{m}\circ}$ *and a function* $m_{\theta_{\mathfrak{m}\circ}} : \mathbb{X} \to \mathbb{R}$ *such that* $P_\circ(m^3_{\theta_{\mathfrak{m}\circ}}) < \infty$ *and for all* $\theta_\mathfrak{m}, \theta'_\mathfrak{m} \in U$, $|\ell_{\theta_\mathfrak{m}} - \ell_{\theta'_\mathfrak{m}}| \le m_{\theta_{\mathfrak{m}\circ}} \|\theta_\mathfrak{m} - \theta'_\mathfrak{m}\|_2$ *a.s.*$[P_\circ]$*;*
(iii) $-P_\circ(\ell_{\theta_\mathfrak{m}} - \ell_{\theta_{\mathfrak{m}\circ}}) = \frac{1}{2}(\theta_\mathfrak{m} - \theta_{\mathfrak{m}\circ})^\top J_{\theta_{\mathfrak{m}\circ}}(\theta_\mathfrak{m} - \theta_{\mathfrak{m}\circ}) + o(\|\theta_\mathfrak{m} - \theta_{\mathfrak{m}\circ}\|^2_2)$ *as* $\theta_\mathfrak{m} \to \theta_{\mathfrak{m}\circ}$*;*

(iv) *$J_{\theta_{\mathfrak{m}\circ}}$ is an invertible matrix; and*
(v) *letting $\vartheta^*_{\mathfrak{m}} \sim \Pi^*(\cdot \,|\, X_{1:N}, \mathfrak{m})$, it holds that conditionally on $X_{1:\infty}$, for almost every $X_{1:\infty}$, for every sequence of constants $C_N \to \infty$,*

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

Further assume that Eq. (4) holds, $\lim_{N\to\infty} M = \infty$, $c \in [0, \infty)$, $\mathbb{E}\{\ell_{1,\theta_{1\circ}}(X_1) - \ell_{2,\theta_{2\circ}}(X_1)\} = 0$, 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 3.1 apply.

The proof is in Appendix C.2.

3.2. *Graphical illustration of the theory.* Figure 1 illustrates how Theorem 3.1 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, $\mathbb{P}\{Q(1 \,|\, X_{1:N}) \approx 0\}$ is non-negligible. On the other hand, the probability that the bagged posterior strongly favors model 2 goes to zero rapidly as δ_∞ increases – that is, $\mathbb{P}\{Q^*(1 \,|\, X_{1:N}) \approx 0\} \to 0$ 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, the standard posterior will overwhelmingly favor the "wrong" model in 1 out of 50 experiments, whereas BayesBag will somewhat strongly favor the wrong model in only around 7 out of 100,000 experiments.

3.3. *Model criticism with BayesBag.* In the setting of parameter inference and prediction, Huggins and Miller (2019) developed a measure of misspecification, referred to as the *model–data mismatch index*, based on comparing the bagged posterior versus the standard posterior. Here, we discuss how to use the mismatch index in the setting of model selection.

To describe the mismatch index, we consider parameter inference in a single model, and therefore omit dependence on $\mathfrak{m}$ in our notation. Let $f : \Theta \to \mathbb{R}$ be a real-valued function and suppose the quantity of inferential interest is $f(\theta_\circ)$. Let v_N and v^*_N denote, respectively, the standard and bagged posterior variances of $f(\theta)$ using $M = N$. If the posterior is well-calibrated, then asymptotically, $v^*_N = 2v_N$. The asymptotic version of the mismatch index is defined as

$$\mathcal{I}(f) := \begin{cases} 1 - 2v_N/v^*_N & \text{if } v^*_N > v_N \\ \mathsf{NA} & \text{otherwise.} \end{cases}$$

The interpretation is as follows: $\mathcal{I}(f) \approx 0$ indicates no evidence of mismatch; $\mathcal{I}(f) > 0$ (respectively, $\mathcal{I}(f) < 0$) indicates the standard posterior is overconfident (respectively, underconfident); $\mathcal{I}(f) = \mathsf{NA}$ indicates that either the required asymptotic assumptions do not hold (for example, due to multimodality in the posterior or small sample size) or there is severe model–data mismatch. We refer the interested reader to Huggins and Miller (2019) for more detailed justification and a description of the non-asymptotic version of $\mathcal{I}$.

For a set of functions of interest $\mathcal{F}$, we suggest taking the most pessimistic mismatch index: $\mathcal{I}(\mathcal{F}) := \sup_{f\in\mathcal{F}} \mathcal{I}(f)$. 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

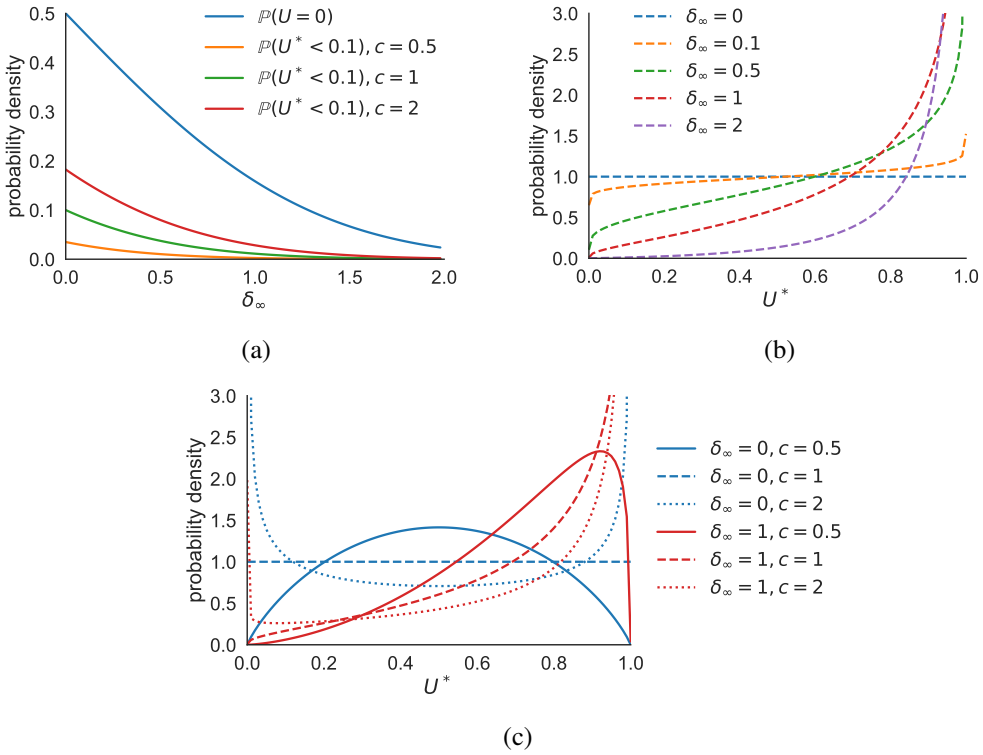


Fig 1: Asymptotic distribution of posterior probability of model 1 under the standard posterior (U) and bagged posterior (U^*). Larger values of the effect size $\delta_\infty = \mu_\infty / \sigma_\infty$ indicate stronger evidence for model 1. **(A)** Probabilities that $U = 0$ and $U^* < 0.1$ as a 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.

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 and therefore adopt the shorthand notation $\mathcal{I} := \mathcal{I}(\mathcal{F}_{\text{proj}})$.

For model selection problems, the use of the mismatch index requires some care. One common case is a partially ordered, finite set of models such that there exists a unique maximal model. Feature selection is an example of this type of problem, which we explore in Section 4, where the maximal model includes all features. In this case it makes sense to apply the mismatch index to the maximal model, since if any model is correctly specified, then the maximal model is correctly specified. Another common situation is when all models have a set of shared, interpretable parameters, in which case we can consider the marginal posterior distribution of the shared parameters across all models. Section 5 considers phylogenetic tree reconstruction, which involves models having this property. The case of an infinite set of models without shared parameters is more delicate. If the models are nested, one possibility would be to apply the mismatch index to the most complex model that has non-trivial posterior probability.

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

Model. The data consist of regressors $Z_n \in \mathbb{R}^D$ and observations $Y_n \in \mathbb{R}$ for $n = 1, \ldots, N$, and the parameter is $\theta = (\theta_0, \ldots, \theta_D) = (\log \sigma^2, \beta_1, \ldots, \beta_D) \in \mathbb{R}^{D+1}$. For each $\gamma \in \{0,1\}^D$, we define a model such that the dth regressor is included in the linear regression if and only if $\gamma_d = 1$. Letting $D_\gamma := \sum_{d=1}^D \gamma_d$ and $k^\star \in \{1, \ldots, D\}$, we consider a collection of models $\mathfrak{M}_{k^\star} := \{\gamma \in \{0,1\}^D \,|\, D_\gamma \leq k^\star\}$. Let $Z \in \mathbb{R}^{N \times D}$ denote the matrix with the nth row equal to Z_n and let Z_γ denote the submatrix of Z that includes the dth column if and only if $\gamma_d = 1$. Conditional on model γ, the parameter space is $\Theta_\gamma = \mathbb{R}^{D_\gamma + 1}$ and the assumed model is

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

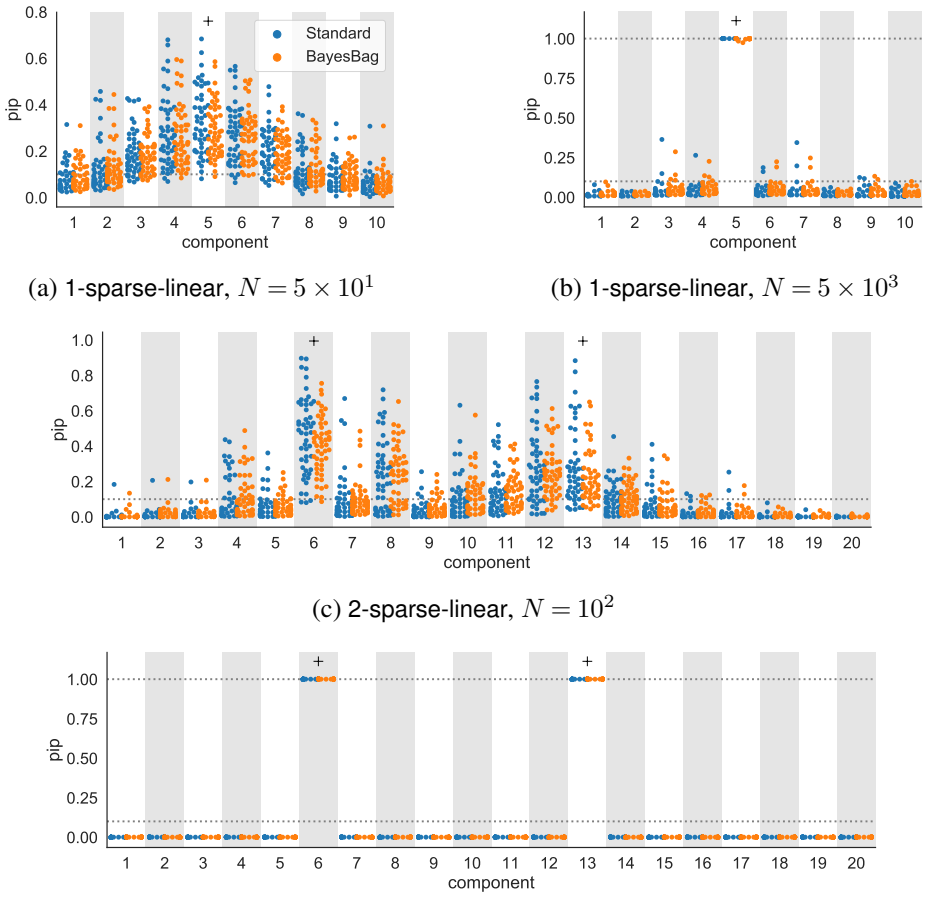
To perform posterior inference for γ, we analytically compute the marginal likelihood for each model γ, integrating out σ^2 and β. For the prior on $\gamma \in \mathfrak{M}_{k^\star}$, define $Q_0(\gamma) \propto q_0^{D_\gamma}(1 - q_0)^{D - D_\gamma}$, where $q_0 \in (0,1)$ is the prior inclusion probability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

Data. 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

$$Y_n = f(Z_n)^\top \beta_\dagger + \epsilon_n \tag{5}$$

for $n = 1, \ldots, N$, with the regressor distribution G, the regression function f, and the coefficient vector $\beta_\dagger \in \mathbb{R}^D$ as described next. We used either the linear regression function $f(z) = z$ or, to generate misspecified data, the nonlinear function $f(z) = (z_1^3, \ldots, z_D^3)^\top$. We chose G and $\beta_\dagger$ the spirit of GWAS fine-mapping (Schaid et al., 2018) to simulate a scenario with many highly correlated regressors of which only a few regressors are truly "causal." We used a k-sparse vector (for $k \in \{1,2\}$) 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 $h = 10$, $Z \sim G$ 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 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.

Experimental conditions. We generated datasets under the k-sparse-linear and k-sparse-nonlinear settings with either (a) $D = 10$, $N = 50$, and $k = 1$, or (b) $D = 20$, $N = 100$, and $k = 2$. We set the prior inclusion probability to $q_0 = k/D$ and the model hyperparameters to $a_0 = 2$, $b_0 = 1$, and $\lambda = 16$, with the latter setting helping to penalize the addition of extraneous features. We set $M = N$ per our default recommendation and set $k^\star = 2$. Each experimental condition was replicated 50 times.



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

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

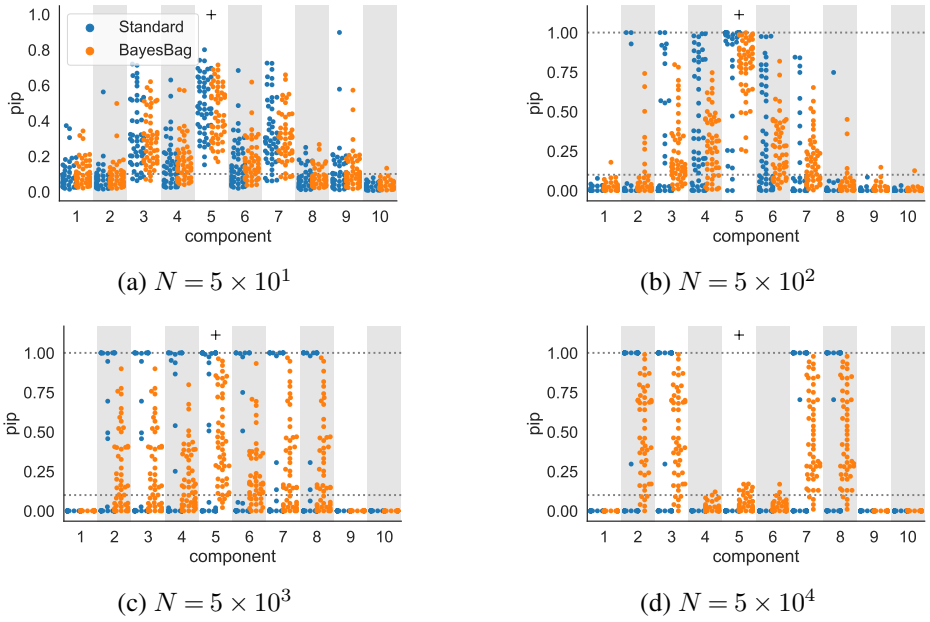
(c) 2-sparse-linear, $N = 10^2$

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

Fig 2: 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 (that is, 1).

Results. We are interested in verifying the theory of Section 3 in the finite-sample regime, which suggests that when the model is misspecified, similar models may be assigned wildly varying probabilities under the standard posterior, while the bagged posterior probabilities will tend to be more balanced. In Figs. 2, 3 and A.2, we plot the standard and bagged posterior inclusion probabilities (pips) for each component, for all 50 replications. First, Fig. 2 shows that when the model is correctly specified, standard Bayes and BayesBag behave 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. 3 and A.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



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

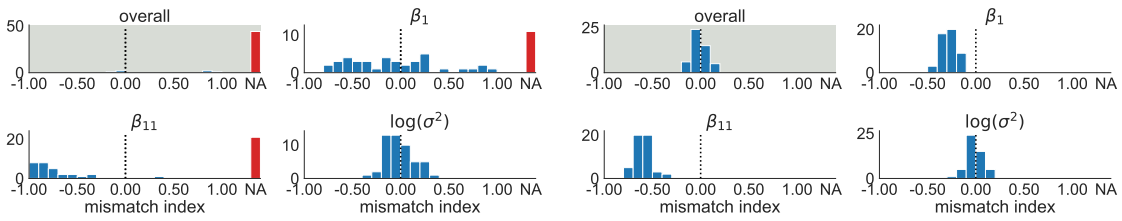
(c) $N = 5 \times 10^3$ (d) $N = 5 \times 10^4$

Fig 3: Posterior inclusion probabilities (pips) for misspecified 1-sparse-nonlinear data. See the caption of Fig. 2 for further explanation. The standard Bayes pips show considerable instability both (i) across datasets with N fixed and (ii) as N increases, while the BayesBag pips are much more stable.

that all components will be selected – however, to maintain sparsity, we chose $k^\star = 2$. See Appendix B for derivations and further discussion; also see Buja et al. (2019a,b).

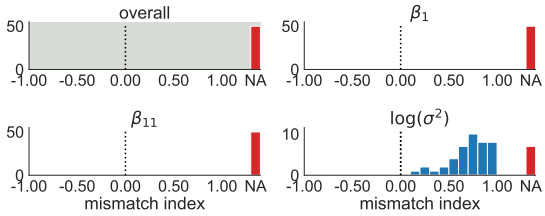
Figure 3 shows the results for the 1-sparse-nonlinear data. The regressor distribution G and coefficient vector $\beta_\dagger$ are such that, by symmetry, components $5 - i$ and $5 + i$ are equivalent, for $i = 0, \ldots, 4$. As $N \to \infty$, it is optimal to use component 3 (and/or component 7) and component 2 (and/or component 8). The standard Bayes pip for either component 3 or 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 on the interval from 0 to 1, as expected. The standard Bayes 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 3.1 and Corollary 3.2. We defer discussion of the results for 1-sparse-nonlinear data (Fig. A.2) to Appendix A.

Figures 4 and A.1 show 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 k-sparse-linear data, the overall mismatch indices were either near zero or were NA, reflecting that the model is correctly specified but there are some issues with poor identifiability. For the k-sparse-nonlinear data, the mismatch indices were nearly all NA, reflecting that the model is misspecified and there may also be identifiability issues.

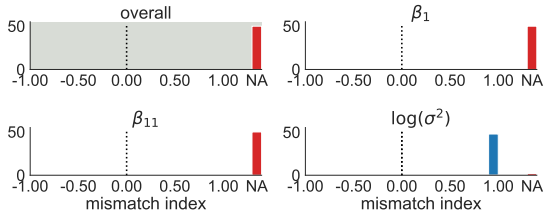


(a) 2-sparse-linear, $N = 10^2$

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



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



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

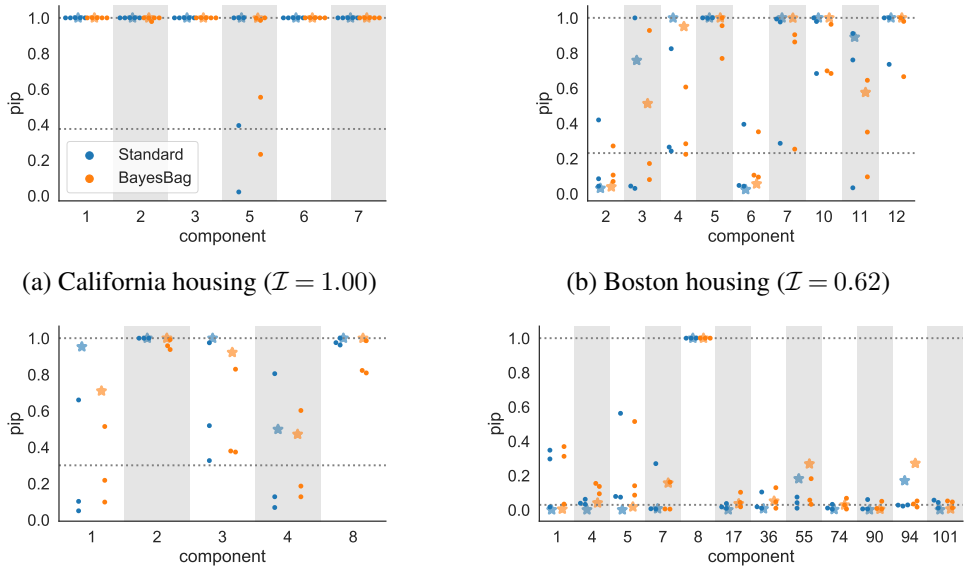
Fig 4: Model–data mismatch indices $\mathcal{I}$ for selected parameters as well as the overall $\mathcal{I}$ value, in the case of 2-sparse data. We only display two components of β since the $\mathcal{I}$ values follow fairly similar distributions for all components.

TABLE 1
Real-world datasets used in experiments. LR = linear regression, 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
Whale mitochondrial coding DNA	PTR	14	10,605
Whale mitochondrial amino acids	PTR	14	3,535

5. Applications. We evaluate the performance of BayesBag for model selection using real-world data in two scenarios: feature selection for linear regression and phylogenetic tree reconstruction. Table 1 summarizes the datasets used.

5.1. *Feature selection for linear regression.* We compared standard Bayesian and BayesBag model selection for linear regression on four real-world datasets. For BayesBag, we set $M = N$ per our default recommendation. We used a prior inclusion probability of $q_0 = 3/D$ and used $k^\star = D$ for the maximum number of nonzero components, except for the residential building dataset, where for computational tractability we used $k^\star = 3$. We expected the parameters to be well-identified for all datasets except the residential building dataset, since the residential building dataset required only 58 out of 104 principle com-



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

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

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

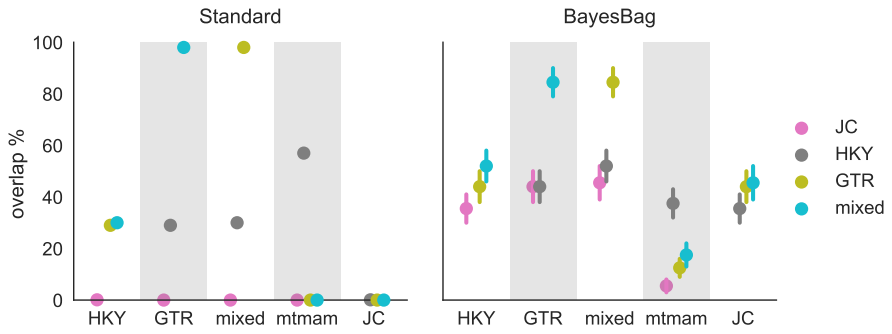
(d) Residential building ($\lambda = 16$, $\mathcal{I} = \mathsf{NA}$)

Fig 5: Posterior inclusion probabilities (pips) for four real-world datasets when the data was 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 (that is, 1).

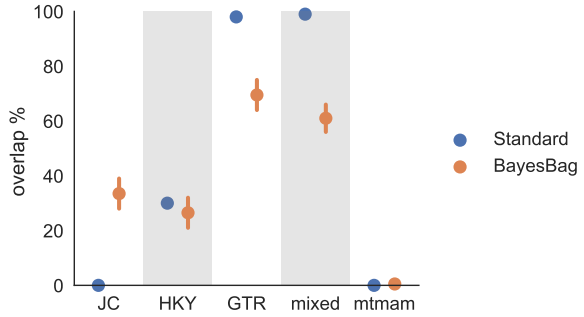
ponents to explain 99% of the variance, whereas for the other three datasets, D out of D principle components were needed to explain 99% of the variance. Therefore, we used $\lambda = 16$ for the residential building dataset and $\lambda = 1$ otherwise. The model mismatch indices (computed with $\gamma_d = 1$ 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 1.00, 0.62, and 0.03, respectively, indicating that the model was misspecified for the two housing datasets.

Figure 5 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. Figure 5 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. These results are consistent with the simulation results in Section 4.

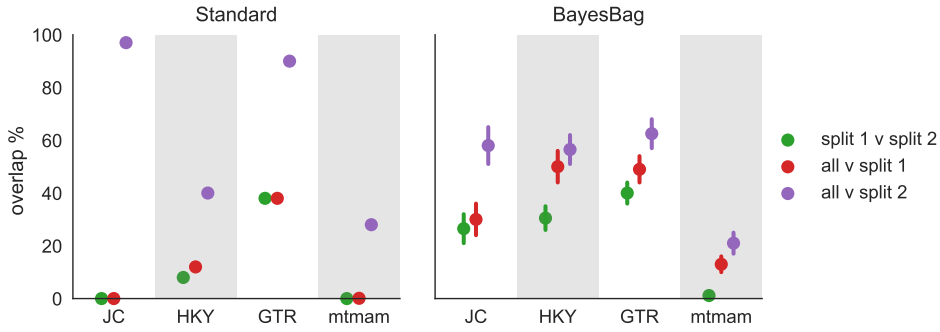
5.2. *Phylogenetic tree reconstruction.* Finally, we investigate the use of BayesBag for reconstructing the phylogenetic tree of a collection of species based on their observed



(a) Inter-model comparison



(b) Comparison to mixed using standard Bayesian model selection



(c) Intra-model comparison

Fig 6: Comparison of standard Bayesian and BayesBag model selection on the whale dataset in terms of overlapping probability mass of 99% highest posterior density regions. (a) Overlap between each pair of models; (b) overlap between each model using BayesBag and the mixed model using standard Bayes; (c) overlap between each pair of data splits, for each model. To quantify uncertainty in the overlap due to Monte Carlo error, 80% confidence intervals are shown for the overlaps involving BayesBag.

characteristics. This is an important model selection problem due to the widespread use of phylogeny reconstruction algorithms. Systematists have exhaustively documented that Bayesian model selection of phylogenetic trees can behave poorly. In particular, the standard posterior 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 et al., 2003; Buckley, 2002; Douady et al., 2003; Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004; Waddell et al., 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 assumed 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. 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 built-in convergence diagnostics for MrBayes. 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% highest posterior density (HPD) regions for each ⟨data, model, inference method⟩ configuration. For selected pairs of configurations, we then computed the overlap of the two 99% HPD regions in terms of (a) probability mass and (b) number of trees. Since the bagged posterior is approximated via Monte Carlo as in Eq. (2), we quantify the uncertainty in each overlap by reporting an 80% confidence interval for the overlapping probability mass. (We computed these confidence intervals using standard bootstrap methodology for a Monte Carlo estimate.)

First, we looked at the overlap between pairs of models when using the standard posterior and bagged posterior. As shown in Fig. 6(a) and Table A.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. Thus, these pairs of models produced contradictory results when using standard Bayes. Meanwhile, when using BayesBag, all pairs of models had nonzero overlap, with typical amounts ranging from 30% to 50%. Hence, BayesBag provided results that were more consistent across models, compared to standard Bayes.

However, the good overlap provided 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. 6(b) and

Table A.2, all the bagged posteriors (with the exception of mtmam) put substantial posterior probability on the 99% HPD region 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% HPD region 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 each model, we would hope to see substantial overlap when comparing the results across these three datasets (all, S1, and S2). 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. 6(c) and Table A.3. Meanwhile, 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.

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(2): 255–266.

Berk, R. H. (1966). "Limiting Behavior of Posterior Distributions when the Model is Incorrect." *The Annals of Mathematical Statistics*, 37(1): 51–58.

Breiman, L. (1996). "Bagging Predictors." *Machine Learning*, 24(2): 123–140.

Buckley, T. R. (2002). "Model Misspecification and Probabilistic Tests of Topology: Evidence from Empirical Data Sets." *Systematic Biology*, 51(3): 509–523.

Bühlmann, P. (2014). "Discussion of Big Bayes Stories and BayesBag." *Statistical Science*, 29(1): 91–94.

Bühlmann, P. and Yu, B. (2002). "Analyzing bagging." *The Annals of Statistics*, 30(4): 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(4): 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(4): 545–565.

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

Clarke, B. S. and Barron, A. R. (1990). "Information-theoretic asymptotics of Bayes methods." *Information Theory, IEEE Transactions on*, 36(3): 453–471.

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

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(2): 248–254.

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(6): 904–913.

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

Kallenberg, O. (2002). *Foundations of Modern Probability*. New York, NY: Springer, 2nd edition.

Lemmon, A. R. and Moriarty, E. C. (2004). "The Importance of Proper Model Assumption in Bayesian Phylogenetics." *Systematic Biology*, 53(2): 265–277.

Mammen, E. (1992). "Bootstrap, wild bootstrap, and asymptotic normality." *Probability Theory and Related Fields*, 93(4): 439–455.

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

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].

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(3): 539–542.

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(8): 1–14.

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.

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(2): 361–371.

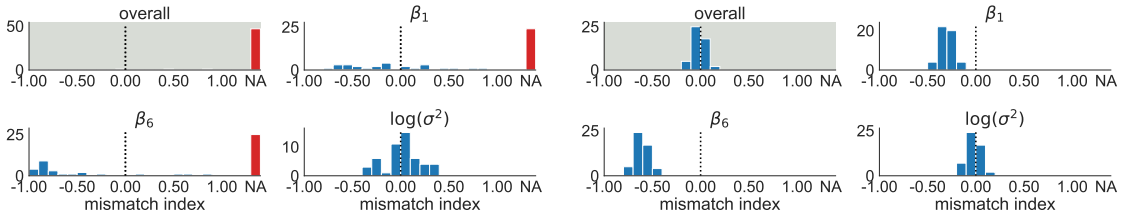
Yang, Z. (2007). "Fair-Balance Paradox, Star-tree Paradox, and Bayesian Phylogenetics." *Molecular Biology and Evolution*, 24(8): 1639–1655.

— (2008). "Empirical evaluation of a prior for Bayesian phylogenetic inference." *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1512): 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(8): 1854–1859.

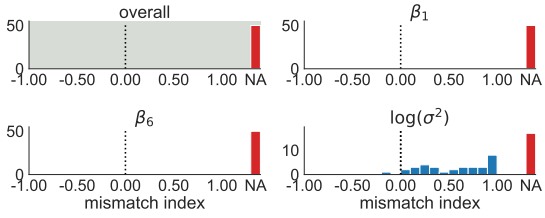
APPENDIX A: ADDITIONAL FIGURES AND TABLES

Figure A.2 shows results for 2-sparse-nonlinear data, under the setup of the simulation study in Section 4. The results are similar to the 1-sparse-nonlinear case in Fig. 3, however, note 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 are more reliable and do not display this instability.

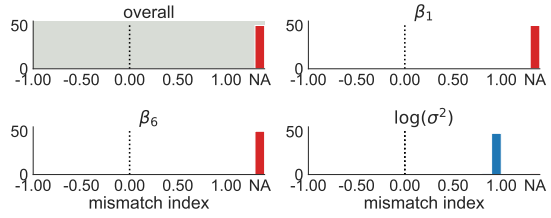


(a) 1-sparse-linear, $N = 5 \times 10^1$

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

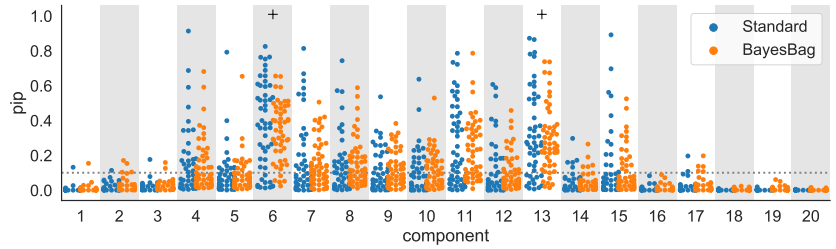


(c) 1-sparse-nonlinear, $N = 5 \times 10^1$

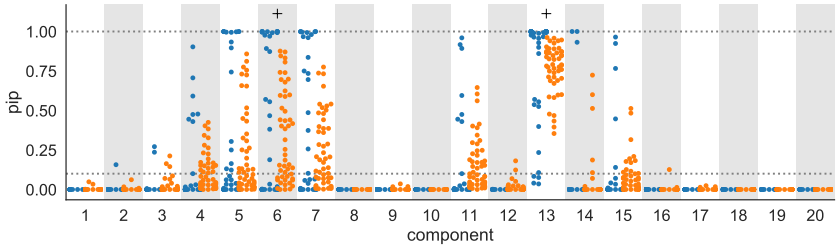


(d) 1-sparse-nonlinear, $N = 5 \times 10^3$

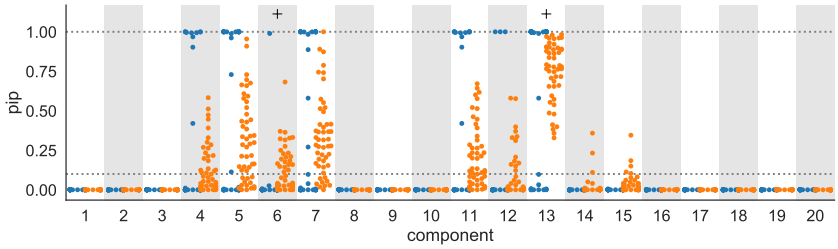
Fig A.1: Model–data mismatch indices $\mathcal{I}$ for selected parameters as well as the overall $\mathcal{I}$ value, in the case of 1-sparse data. We only display two components of β since the $\mathcal{I}$ values follow fairly similar distributions for all components.



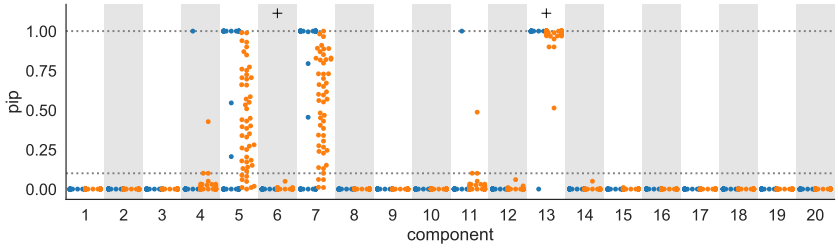
(a) $N = 10^2$



(b) $N = 10^3$



(c) $N = 10^4$



(d) $N = 10^5$

Fig A.2: Posterior inclusion probabilities (pips) for misspecified 2-sparse-nonlinear data. See the caption of Fig. 2 for further explanation.

TABLE A.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 two 99% HPD density regions and "# trees" shows the number of trees in the intersection of the two regions. For BayesBag, "mass" shows an 80% confidence interval for the overlap and "# trees" shows the median number of trees in the intersection.

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 A.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 A.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 A.3

Comparison of self-consistency on whale dataset. The form of each data entry and the BayesBag parameters are the same as Table A.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 B: FEATURE SELECTION IN LINEAR REGRESSION

We derive the KL-optimal linear regression parameters for data generated as in our simulation studies (Section 4). 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 sparse. Assuming a linear regression model and assuming the data follow Eq. (5), we have

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

Thus,

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

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

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

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

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

APPENDIX C: PROOFS

NOTATION. We use $\xrightarrow{P}$ to denote convergence in probability and $\xrightarrow{P_+}$ to denote convergence in outer probability.

C.1. Proof of Theorem 3.1. We first prove a simple uniform central limit theorem that is needed for our proof of Theorem 3.1. 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 C.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 et al. (2010, Thm. 3.2, Thm. 3.3, Eq. 3.14), for any $\alpha \in (0, 1)$,

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

Further,

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

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

$$\lim_{N\to\infty} d_K(\mathcal{L}(\tilde{W}_N), \mathcal{N}(0, \sigma^2)) = 0. \tag{6}$$

Since the cumulative distribution function of $Z \sim \mathcal{N}(\mu, \sigma^2)$ is Lipschitz for some constant $C > 0$, $|\mathbb{P}(Z < t) - \mathbb{P}(Z < s)| \le C|t - s|$. 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

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

By Eq. (6), 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 3.1, 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 C.1 with $\xi_{Nn} := Z_{Nn} + Z_{N0}/N$ that

$$\lim_{N\to\infty} d_K(\mathcal{L}(W_N), \mathcal{L}(W_\infty)) = 0, \tag{7}$$

where the Minkowski inequality and assumption (iii) of Theorem 3.1 verify assumption (iii) of Proposition C.1. In particular, Eq. (7) 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 3.1, parts (2) and (3). Let

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

where $K_{N,1:N} \sim \mathsf{Multi}(M, 1/N)$ 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. (7), we have that

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

and hence

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

We can write the bagged posterior probability of model 1 as

$$Q^*(1 \mid X_{1:N}) = \mathbb{E}\{\phi_M(W_N^*) \mid X_{1:N}\} = \mathbb{E}\{\phi_M(\Delta_N^* + (M/N)^{1/2} W_N) \mid X_{1:N}\}.$$

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 $\alpha \in \mathbb{R}$,

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

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

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

Moreover,

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

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. (7), 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.

C.2. Proof of Corollary 3.2. 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 $\phi(t) = \{1 + \exp(-t)\}^{-1}$. Under assumptions (i)-(iv), 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 3.1, 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.