Reproducible Parameter Inference Using Bagged Posteriors

Jonathan H. Huggins[1] **and Jeffrey W. Miller**[2]

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

[2]*Department of Biostatistics, Harvard University, e-mail:* jwmiller@hsph.harvard.edu

Abstract: Under model misspecification, it is known that Bayesian posteriors often do not properly quantify uncertainty about true or pseudo-true parameters. Even more fundamentally, misspecification leads to a lack of reproducibility in the sense that the same model will yield contradictory posteriors on independent data sets from the true distribution. To define a criterion for reproducible uncertainty quantification under misspecification, we consider the probability that two credible sets constructed from independent data sets have nonempty overlap, and we establish a lower bound on this overlap probability that holds whenever the credible sets are valid confidence sets. We prove that credible sets from the standard posterior can strongly violate this bound, indicating that it is not internally coherent under misspecification. To improve reproducibility in an easy-to-use and widely applicable way, we propose to apply bagging to the Bayesian posterior ("BayesBag"); that is, to use the average of posterior distributions conditioned on bootstrapped datasets. We motivate BayesBag from first principles based on Jeffrey conditionalization and show that the bagged posterior typically satisfies the overlap lower bound. Further, we prove a Bernstein–Von Mises theorem for the bagged posterior, establishing its asymptotic normal distribution. We demonstrate the benefits of BayesBag via simulation experiments and an application to crime rate prediction.

MSC2020 subject classifications: Primary 62F15, 62F40; secondary 62A01, 62F35.
Keywords and phrases: Bagging, Bernstein–Von Mises theorem, Bootstrap, Jeffrey conditionalization, Model misspecification, Overlap probability, Uncertainty calibration.

1. Introduction

It is widely acknowledged that statistical models are usually not exactly correct in practice [4, 5, 10, 32]. This model misspecification is known to lead to unreliable inferences, and in particular, Bayesian posteriors can be unstable and poorly calibrated under misspecification [18, 26, 28]. Unfortunately, this leads to a lack of reproducibility, even when using the same method on a replicate data set from the same distribution [23, 48]. In this paper, we propose a criterion for reproducible uncertainty quantification and a general technique for achieving it.

Defining valid uncertainty quantification in misspecified models presents a conceptual problem since there is no ""true parameter" that indexes a model with distribution equal to the data-generating distribution. The usual solution

---

is to focus on a pseudo-true parameter, typically defined as the asymptotically optimal parameter in terms of Kullback–Leibler (KL) divergence [19, 28, 36, 46]. However, depending on the objectives of the analysis, it might not be desirable to concentrate at the KL-optimal parameter [3, 26, 35]. Thus, rather than adopting a particular definition of pseudo-truth, we introduce a truth-agnostic approach to assessing reproducibility. Specifically, we consider the probability that two credible sets constructed from independent data sets have nonempty overlap. We establish a simple lower bound on this overlap probability that holds whenever the sets do have coverage probability of $1 - \alpha$ for the quantity of interest. Under misspecification, we show that credible sets from the standard posterior can strongly violate this bound – particularly when the dimension grows with the number of observations – indicating that it exhibits poor reproducibility.

To improve the reproducibility of Bayesian inference under misspecification, we propose to use *BayesBag* [7, 15, 44]. The idea of BayesBag is to apply bagging [6] to the Bayesian posterior. More precisely, 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 N^M possible bootstrap datasets of M samples drawn with replacement from the original dataset. In this work, we focus on parameter inference and prediction, complementing our work on BayesBag for model selection [23]. In theory and experiments, we consider both the case of fixed finite-dimensional parameters as well as cases where the parameter dimension grows with the sample size.

We motivate the bagged posterior from first principles using Jeffrey conditionalization and show that bagged posterior credible sets typically satisfy our lower bound on the overlap probability, indicating that the bagged posterior quantifies uncertainty in a more reproducible way. These results illustrate how the bagged posterior integrates the attractive features of Bayesian inference— such as flexible hierarchical modeling, automatic integration over nuisance parameters, and the use of prior information—with the distributional robustness of frequentist methods, nonparametrically accounting for sampling variability and model misspecification. Simulation experiments validate our theory and demonstrate the bagged posterior is particularly important for stability in high-dimensional settings. An application to crime rate prediction using a Poisson regression model with a horseshoe prior to induce approximate sparsity demonstrate that BayesBag-based analysis can also lead to different conclusions – and better predictions – than a standard Bayesian analysis.

In practice, we suggest approximating $\pi^*(\theta \,|\, x)$ by generating B independent bootstrap datasets $x_{(1)}^*, \ldots, x_{(B)}^*$, where $x_{(b)}^*$ consists of M samples drawn with

replacement from x, yielding the approximation

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

Since the bagged posterior is just the average of standard Bayesian posteriors, one can use any algorithm for standard posterior inference to compute each of the B posteriors, and then aggregate across them. While this requires B times as much computation as a single posterior, it is trivial to parallelize the computation of the B posteriors. Since Eq. (2) is a simple Monte Carlo approximation, the error of this approximation can easily be estimated in order to choose B appropriately [23].

Despite its many attractive features, there has been little practical or theoretical investigation of bagged posteriors prior to Huggins and Miller [23]. In the only previous work of which we are aware, Bühlmann [7] presented some simulation results for a simple Gaussian location model, while Waddell, Kishino and Ota [44] and Douady et al. [15] used bagged posteriors for phylogenetic tree inference in papers focused primarily on speeding up model selection and comparing Bayesian inference versus the bootstrap.

The article is organized as follows. In Section 2, we motivate the use of Bayes-Bag for reproducible uncertainty quantification in terms of our overlap criterion as well as a Jeffrey conditionalization derivation. In Section 3, we prove that the standard posterior often fails to satisfy the overlap criterion, whereas the bagged posterior typically satisfies it, focusing on Gaussian location models, regular finite-dimensional models, and linear regression. In Section 4, we prove a general Bernstein–Von Mises theorem establishing the asymptotic normal distribution of the bagged posterior, which is employed in the overlap theory of the preceding section. Section 5 evaluates the performance of the bagged posterior in simulation studies, and Section 6 illustrates with an application to crime rate prediction using Poisson regression. We close with a discussion in Section 7.

2. Motivation

When misspecified, a Bayesian model can be so unstable that it contradicts itself. Specifically, given two independent data sets from the same distribution, the resulting two posteriors—for the same model—can place nearly all their mass on disjoint sets. Figure 1a provides a simple illustration of the problem. Intuitively, it seems clear that this must violate some principle of coherent uncertainty quantification. But if there is no true parameter for which the model is correct, then what is a posterior quantifying uncertainty about? In most previous work, this question is dealt with by focusing on the pseudo-true parameter—that is, the model parameter value that is closest in Kullback–Leibler divergence to the true distribution [11, 21, 28, 45]. However, this choice—or any choice of pseudo-truth—is somewhat arbitrary and entails implicit assumptions about the goal of the analysis, such as minimizing a certain loss function.
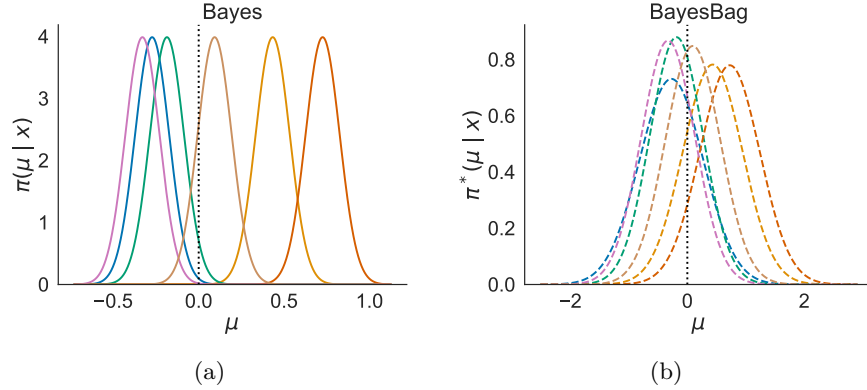
Fig 1: Standard and bagged posterior distributions of the mean for a Gaussian location model assuming the data are i.i.d. $\mathcal{N}(\mu, 1)$, when the data are actually i.i.d. $\mathcal{N}(0, 5^2)$. Posteriors for six randomly generated data sets of size $N = 100$ are shown. **(a)** Many pairs of posterior distributions have essentially no overlap with each other, and 5 out 6 do not contain the true mean in their 95% central credible sets. **(b)** All pairs of bagged posterior distributions have significant overlap and 6 out of 6 contain the true mean in their 95% central credible sets.

In this section, we instead formulate a criterion for reproducible uncertainty quantification that does not require any assumptions of what is true in terms of models or parameters. The basic idea is that two valid confidence sets constructed from independent data sets must intersect with a certain minimal probability. We prove a simple lower bound on this overlap probability that holds for any valid confidence sets, for any definition of pseudo-truth, and for any data distribution. We then use this criterion to motivate the use of BayesBag via Jeffrey conditionalization. Figure 1b illustrates how the bagged posterior does not suffer from the instability exhibited by the standard posterior.

2.1. Overlap criterion for reproducible uncertainty quantification

Suppose we have data $X \sim P_\circ$ and we are interested in estimating an unknown quantity $\eta(P_\circ) \in \Omega$ that is a functional of the unknown data-generating distribution $P_\circ$. Here, $\eta(P_\circ)$ may or may not correspond to a parameter in some model. Suppose $x \mapsto A_x$ is a procedure for constructing sets $A_X \subseteq \Omega$ intended to contain $\eta(P_\circ)$ with probability $1 - \alpha$.

Definition 2.1. We say $x \mapsto A_x$ *has coverage* $1 - \alpha$ *with respect to* $P_\circ$ *if* $\mathbb{P}(\eta(P_\circ) \in A_X) \geq 1 - \alpha$ when $X \sim P_\circ$.

This definition is agnostic to making any assumptions of what is true in terms of models or parameters.

Proposition 2.2. *Let $X, Y \sim P_\circ$ independently. If $x \mapsto A_x$ and $y \mapsto B_y$ have, respectively, coverage $1 - \alpha$ and $1 - \alpha'$ with respect to $P_\circ$, then*

$$\mathbb{P}(A_X \cap B_Y \neq \varnothing) \geq (1 - \alpha)(1 - \alpha'). \tag{3}$$

This provides a lower bound on the probability that two credible sets with valid coverage intersect. For example, if the coverage is $1 - \alpha = 1 - \alpha' = 0.95$, then the lower bound on the probability of intersection is 0.9025. We refer to $\mathbb{P}(A_X \cap B_Y \neq \varnothing)$ as the *overlap probability*, and satisfying the bound is referred to as the *overlap criterion*. Failing to satisfy this criterion indicates a lack of stability and reproducibility across plausible datasets. While satisfying this bound is a necessary condition for coherent uncertainty quantification, it is not sufficient. For example, choosing $A_X = A$, a constant, would satisfy the bound but would clearly be an ineffective method for inference.

2.2. Jeffrey conditionalization for reproducibility leads to BayesBag

For reproducibility, one needs to represent uncertainty across data sets from the true distribution. A natural way to do this is via Jeffrey conditionalization, which turns out to lead to the bagged posterior. This interpretation elegantly unifies the Bayesian and frequentist elements of the bagged posterior that might otherwise seem challenging to interpret together in a principled way.

To explain, 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 $p(x) = \int p(x \mid y)p(y)\mathrm{d}y$ and $p(y) = \int p(y \mid x)p(x)\mathrm{d}x$, 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 $q(y) := \int p(y \mid x)p_\circ(x)\mathrm{d}x$. The idea is that $q(x, y) := p(y \mid x)p_\circ(x)$ 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 [13, 24, 25].

Suppose $x = x_{1:N} := (x_1, \ldots, x_N)$ 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 is to quantify our uncertainty in θ by

$$q(\theta) = \int p(\theta \mid x_{1:N})p_{\circ N}(x_{1:N})\mathrm{d}x_{1:N}. \tag{4}$$

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 $\prod_{n=1}^{N} \mathbb{P}_N$ to approximate $p_{\circ N}$ in Eq. (4). Doing so, we arrive at the bagged posterior $\pi^*(\theta \mid x)$ from Eq. (1), in the case of $M = N$:

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

where $X_1^*, \ldots, X_N^*$ i.i.d. $\sim \mathbb{P}_N$ given $X_{1:N}$. Thus, the bagged posterior represents uncertainty in θ, integrating over data sets drawn from an approximation to the true distribution. Hence, it is natural to expect the bagged posterior to improve reproducibility across data sets.

2.3. BayesBag combines Bayesian and frequentist uncertainty

In Eq. (4), $p(\theta \,|\, x_{1:N})$ represents Bayesian model-based uncertainty and integrating with respect to $p_{\circ N}(x_{1:N})$ represents frequentist sampling uncertainty. Remarkably, these two sources of uncertainty combine additively in the bagged posterior whenever $\theta \in \mathbb{R}^D$.

To see this, let $X^* \,|\, x$ be a random bootstrap dataset given data x, and let $\vartheta^* \,|\, X^* \sim \pi(\theta \,|\, X^*)$ be distributed according to the standard posterior given data X^*. 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

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

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 a Bayesian model-based point estimate.

3. Reproducibility using overlap probability

We now investigate if and when the standard and bagged posteriors satisfy the overlap criterion for reproducible uncertainty quantification. We focus on Gaussian location models, regular finite-dimensional models, and linear regression as representative cases, and consider settings where the dimension is fixed or growing with the sample size. We show that under misspecification, the bagged posterior typically satisfies the overlap criterion whereas the standard posterior does not. But, for correctly specified models, both the standard and bagged posteriors usually satisfy the criterion.

First, however, as a check on the reasonableness of our criterion, we establish that for any correctly specified Bayesian model, the overlap criterion is satisfied in expectation with respect to the prior.

Theorem 3.1. *Consider any model for data $X|\vartheta$ and any prior π on ϑ. Suppose $x \mapsto A_x$ is a $100(1-\alpha)\%$ posterior credible set for ϑ under this model and prior, that is, $\mathbb{P}(\vartheta \in A_x \mid x) \geq 1 - \alpha$ for all x. If $\vartheta \sim \pi$, and $X \mid \vartheta$, $\tilde{X} \mid \vartheta$ are independent data from the assumed model, then $\mathbb{E}\{\mathbb{P}(A_X \cap A_{\tilde{X}} \neq \varnothing \mid \vartheta)\} \geq (1-\alpha)^2$.*

Theorem 3.1 is a direct analogue of the classical result that posterior credible sets have correct frequentist coverage in expectation under the assumed prior. All proofs are in Appendix D.

3.1. *Gaussian location model*

We first consider the simple Gaussian location model in which observations x_n are modeled as i.i.d. $\mathcal{N}(\theta, V)$ with fixed positive definite covariance matrix V, and assume a conjugate prior, $\theta \sim \mathcal{N}(0, V_0)$. Given data $x = (x_1, \ldots, x_N)$, the posterior is $\theta \mid x \sim \mathcal{N}(R_N \bar{x}_N, V_N)$, where $\bar{x}_N := N^{-1} \sum_{n=1}^{N} x_n$, $R_N := (V_0^{-1} V/N + I)^{-1}$, and $V_N := (V_0^{-1} + NV^{-1})^{-1}$. For intuition, one can think of $R_N \approx I$ since $\|R_N - I\| = O(N^{-1})$. The bagged posterior mean and covariance are

$$\mathbb{E}(\vartheta^* \mid x) = \mathbb{E}(R_M \bar{X}_M^* \mid x) = R_M \bar{x}_N$$

$$\mathrm{Cov}(\vartheta^* \mid x) = \mathbb{E}(V_M \mid x) + \mathrm{Cov}(R_M \bar{X}_M^* \mid x) = V_M + M^{-1} R_M \hat{\Sigma}_N R_M,$$

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 $\mathbb{E}(\vartheta^* \mid x) = \mathbb{E}(\vartheta \mid x)$ and $\mathrm{Cov}(\vartheta^* \mid x) = \mathrm{Cov}(\vartheta \mid x) + N^{-1} R_N \hat{\Sigma}_N R_N$. 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 $\hat{\Sigma}_N$.

3.1.1. *Overlap probability for Gaussian location model with fixed dimension*

Consider the Gaussian location model above. Fix $\alpha \in (0,1)$ and $u \in \mathbb{R}^D \setminus \{0\}$, and let $A_{x_{1:N}}$ be a $100(1-\alpha)\%$ central credible interval for $u^\top \theta$ given data $x_{1:N}$. For BayesBag, let $A_{x_{1:N}}^*$ denote the $100(1-\alpha)\%$ central interval for the normal distribution matching the mean and variance of the bagged posterior distribution of $u^\top \theta$ given $x_{1:N}$. For readability, we abbreviate $p(\text{overlap}) := \mathbb{P}(A_{X_{1:N}} \cap A_{Y_{1:N}} \neq \varnothing)$ and $p^*(\text{overlap}) := \mathbb{P}(A_{X_{1:N}}^* \cap A_{Y_{1:N}}^* \neq \varnothing)$.

Theorem 3.2. *Suppose the true data distribution $P_\circ$ has positive definite covariance $\Sigma_\circ$. Let $X_1, X_2, \ldots$ i.i.d. $\sim P_\circ$ and $Y_1, Y_2, \ldots$ i.i.d. $\sim P_\circ$ independently. Define $W \sim \mathcal{N}(0,1)$ and $z_{\alpha/2} \in \mathbb{R}$ such that $\mathbb{P}(|W| > z_{\alpha/2}) = \alpha$. Then as $N \to \infty$, for the standard posterior,*

$$p(\text{overlap}) \longrightarrow \mathbb{P}\left(|W| \leq z_{\alpha/2}\sqrt{2}\Big(\frac{u^\top V u}{u^\top \Sigma_\circ u}\Big)^{1/2}\right);$$

and, assuming $M = M(N)$ satisfies $\lim_{N \to \infty} M/N = c > 0$, for the bagged posterior,

$$p^*(\text{overlap}) \longrightarrow \mathbb{P}\left(|W| \le z_{\alpha/2}\sqrt{2}\Big(\frac{u^\top ((V + \Sigma_\circ)/c)u}{u^\top \Sigma_\circ u} \Big)^{1/2} \right),$$

If the model is correct then $V = \Sigma_\circ$, so the standard and bagged posteriors have the same asymptotic behavior when $M = 2N$, specifically, the overlap probability converges to $\mathbb{P}(|W| \le z_{\alpha/2}\sqrt{2})$. However, in misspecified cases where $u^\top V u < u^\top \Sigma_\circ u$, the overlap probability for the standard posterior can be arbitrarily small. On the other hand, the bagged posterior satisfies $\lim_{N\to\infty} p^*(\text{overlap}) \ge \mathbb{P}(|W| \le z_{\alpha/2}) = 1 - \alpha$ when $0 < c \le 2$ since $u^\top V u \ge 0$. Thus, BayesBag is guaranteed to satisfy the overlap criterion necessary for reproducible uncertainty quantification (Eq. (3)) when $0 < c \le 2$, while standard Bayes is not.

3.1.2. Overlap probability for Gaussian location model with growing dimension

To study the case of growing dimension D in the Gaussian location model, we establish finite sample expressions for the overlap probability in the special case of $V = I$ and a flat prior ($V_0^{-1} = 0$), assuming Gaussian data.

Theorem 3.3. *Consider the same setup as in Theorem 3.2. Suppose $P_\circ = \mathcal{N}(0, \Sigma_\circ)$, $V = I$, $V_0^{-1} = 0$, and $\|u\| = 1$. Then for the standard posterior,*

$$p(\text{overlap}) = \mathbb{P}\left(|W| \le \frac{z_{\alpha/2}\sqrt{2}}{(u^\top \Sigma_\circ u)^{1/2}} \right) \tag{5}$$

where $W \sim \mathcal{N}(0,1)$, and for the BayesBag posterior, when $N \ge 2$,

$$p^*(\text{overlap}) \ge \mathbb{P}\left(|T_{2N-2}| \le z_{\alpha/2}\sqrt{(N-1)/M} \right) \tag{6}$$

where T_{2N-2} is t-distributed with $2N - 2$ degrees of freedom.

Note that the right-hand side of Eq. (5) does not depend on N, and the right-hand side of Eq. (6) does not depend on D. Eq. (5) can be arbitrarily small as D grows, since $u^\top \Sigma_\circ u$ can be arbitrarily large. For instance, this will often be the case when $\Sigma_\circ$ has order D^2 nonnegligible entries. Thus, as the dimension D grows, the standard posterior can severely violate the overlap criterion. Meanwhile, if $M/N \to 1$ as $N \to \infty$, then the lower bound in Eq. (6) converges to $\mathbb{P}(|W| \le z_{\alpha/2}) = 1 - \alpha > (1 - \alpha)^2$ since $T_{2N-2} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$ as $N \to \infty$. Therefore, for all N sufficiently large, for all D, BayesBag satisfies the overlap criterion.

3.2. Regular finite-dimensional models

Asymptotically, sufficiently regular finite-dimensional models behave like the Gaussian location model. We have $N^{1/2}(\vartheta - \hat{\theta}_N) \,|\, X_{1:N} \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_{\theta_\circ}^{-1})$ by the

Bernstein–Von Mises theorem, and $N^{1/2}(\hat{\theta}_N - \theta_\circ) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_{\theta_\circ}^{-1} I_{\theta_\circ} J_{\theta_\circ}^{-1})$ by classical theory, where $\hat{\theta}_N$ is the maximum likelihood estimator, $\theta_\circ$ is the Kullback–Leibler optimal parameter, and $J_{\theta_\circ}$, $I_{\theta_\circ}$ are information matrices; see Section 4 for details. In Section 4, we prove that for the bagged posterior, $N^{1/2}(\vartheta^* - \hat{\theta}_N) \mid X_{1:N} \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_{\theta_\circ}^{-1}/c + J_{\theta_\circ}^{-1} I_{\theta_\circ} J_{\theta_\circ}^{-1}/c)$ where $c := \lim_{N \to \infty} M/N$.

Fix $u \in \mathbb{R}^D \setminus \{0\}$ and $\alpha \in (0, 1)$. Let $p_\infty(\text{overlap})$ and $p_\infty^*(\text{overlap})$ denote the asymptotic overlap probabilities of $100(1-\alpha)\%$ central credible intervals for $u^\top \theta$ under these asymptotic normal distributions for the standard and bagged posteriors, respectively, assuming $J_{\theta_\circ}$ and $I_{\theta_\circ}$ are positive definite.

Theorem 3.4. *Let $W \sim \mathcal{N}(0, 1)$. For the standard posterior,*

$$p_\infty(\text{overlap}) = \mathbb{P}\left(|W| \leq z_{\alpha/2}\sqrt{2}\left(\frac{u^\top J_{\theta_\circ}^{-1} u}{u^\top J_{\theta_\circ}^{-1} I_{\theta_\circ} J_{\theta_\circ}^{-1} u} \right)^{1/2} \right),$$

and for the bagged posterior,

$$p_\infty^*(\text{overlap}) = \mathbb{P}\left(|W| \leq z_{\alpha/2}\sqrt{2}\left(\frac{u^\top (J_{\theta_\circ}^{-1}/c + J_{\theta_\circ}^{-1} I_{\theta_\circ} J_{\theta_\circ}^{-1}/c) u}{u^\top J_{\theta_\circ}^{-1} I_{\theta_\circ} J_{\theta_\circ}^{-1} u} \right)^{1/2} \right)$$

$$\geq \mathbb{P}(|W| \leq z_{\alpha/2}\sqrt{2/c}).$$

In general, the ratio $(u^\top J_{\theta_\circ}^{-1} u)/(u^\top J_{\theta_\circ}^{-1} I_{\theta_\circ} J_{\theta_\circ}^{-1} u)$ can be arbitrarily large or small. In particular, $p_\infty(\text{overlap})$ can be arbitrarily small, implying that the asymptotic standard posterior can strongly violate the overlap criterion in Eq. (3). On the other hand, as long as $c \leq 2$, we have $p_\infty^*(\text{overlap}) \geq 1 - \alpha$, implying that the asymptotic bagged posterior satisfies the overlap criterion.

3.3. Linear regression model

Consider data consisting of regressors $Z_n \in \mathbb{R}^D$ and outcomes $Y_n \in \mathbb{R}$ ($n = 1, \ldots, N$), and let $Z \in \mathbb{R}^{N \times D}$ denote the complete design matrix, and $Y \in \mathbb{R}^N$ the vector of outcomes. We analyze the linear regression model

$$Y \mid Z, \beta \sim \mathcal{N}(Z\beta, \sigma^2 I),$$

where $\beta \in \mathbb{R}^D$ is the vector of coefficients, $\sigma^2 > 0$ is the outcome variance, To simplify the analysis, assume $Z^\top Z$ is invertible, σ^2 is fixed but possibly unknown, and β is given a flat prior. For any $u \in \mathbb{R}^D \setminus \{0\}$, the resulting posterior on $u^\top \beta$ is

$$u^\top \beta \mid Z, Y \sim \mathcal{N}(v^\top Y, \sigma^2 \|v\|^2)$$

where $v := Z(Z^\top Z)^{-1} u$. Now, suppose the true distribution is $Y \mid Z \sim \mathcal{N}(\mu_\dagger, \Sigma_\dagger)$ where $\mu_\dagger$ and $\Sigma_\dagger$ are functions of Z, say, $\mu_\dagger = m(Z)$ and $\Sigma_\dagger = K(Z)$. Note that the model is correctly specified when $m(Z) = Z\beta$ and $K(Z) = \sigma^2 I$.

Consider two replicate experiments with data $Y \mid Z \sim \mathcal{N}(m(Z), K(Z))$ and $\tilde{Y} \mid \tilde{Z} \sim \mathcal{N}(m(\tilde{Z}), K(\tilde{Z}))$, respectively, where $Z^\top Z$ and $\tilde{Z}^\top \tilde{Z}$ are invertible and the model variances are σ^2 and $\tilde{\sigma}^2$. Letting $A = v^\top Y \pm z_{\alpha/2}\sigma\|v\|$ and $\tilde{A} = \tilde{v}^\top \tilde{Y} \pm z_{\alpha/2}\tilde{\sigma}\|\tilde{v}\|$ be the corresponding $100(1-\alpha)\%$ central credible sets for $u^\top \beta$, the overlap probability is $p(\text{overlap} \mid Z, \tilde{Z}) = \mathbb{P}(A \cap \tilde{A} \neq \varnothing \mid Z, \tilde{Z})$.

Theorem 3.5. *Consider the linear regression model above and let $W \sim \mathcal{N}(0,1)$.*

1. If $m(Z) = Z\beta_\dagger$ and $K(Z) = \sigma_\dagger^2 I$, then

$$p(\text{overlap} \mid Z, \tilde{Z}) = \mathbb{P}\left(|W| \leq \frac{z_{\alpha/2}(\sigma\|v\| + \tilde{\sigma}\|\tilde{v}\|)}{\sigma_\dagger(\|v\|^2 + \|\tilde{v}\|^2)^{1/2}}\right). \tag{7}$$

2. If $Z = \tilde{Z}$, but we make no assumptions on the form of $m(Z)$ or $K(Z)$, then

$$p(\text{overlap} \mid Z, \tilde{Z}) = \mathbb{P}\left(|W| \leq \frac{z_{\alpha/2}\,(\sigma + \tilde{\sigma})\|v\|}{\sqrt{2}(v^\top K(Z)v)^{1/2}}\right). \tag{8}$$

3. If $K(Z) = \sigma_\dagger^2 I$, but we make no assumptions on the form of $m(Z)$, then

$$p(\text{overlap} \mid Z, \tilde{Z}) \leq \mathbb{P}\left(\left|W + \frac{v^\top m(Z) - \tilde{v}^\top m(\tilde{Z})}{\sigma_\dagger\,(\|v\|^2 + \|\tilde{v}\|^2)^{1/2}}\right| \leq \frac{z_{\alpha/2}\sqrt{\sigma^2 + \tilde{\sigma}^2}}{\sigma_\dagger}\right). \tag{9}$$

Eq. (7) shows that if the linear regression model is correctly specified, then the standard posterior satisfies the overlap criterion (Eq. (3)), since $\sigma = \tilde{\sigma} = \sigma_\dagger$) and therefore $p(\text{overlap} \mid Z, \tilde{Z}) \geq \mathbb{P}(|W| \leq z_{\alpha/2}) = 1 - \alpha > (1-\alpha)^2$, by the fact that $\|v\| + \|\tilde{v}\| \geq (\|v\|^2 + \|\tilde{v}\|^2)^{1/2}$. If the model is correct but the variance is unknown, and consistent estimators of $\sigma_\dagger^2$ are plugged in for σ^2 and $\tilde{\sigma}^2$, then the overlap criterion is satisfied for all N sufficiently large.

However, when either the covariance $K(Z)$ or the mean function $m(Z)$ is misspecified, the standard posterior can violate the overlap criterion. Consider the case where $Z = \tilde{Z}$, that is, the design matrix is the same across replicates; we refer to this as a fixed design setting. Eq. (8) shows that $p(\text{overlap} \mid Z, \tilde{Z})$ does not depend on $m(Z)$, so misspecification of the mean function has no effect on the overlap probability in this case. Nonetheless, Eq. (8) shows that $p(\text{overlap} \mid Z, \tilde{Z})$ can be arbitrarily small when the covariance is misspecified, because the ratio $(\sigma + \tilde{\sigma})\|v\|/(v^\top K(Z)v)^{1/2}$ can be arbitrarily small. Clearly, this ratio will be small if σ^2 and $\tilde{\sigma}^2$ are blindly set too low, but it can also be small if these variances are estimated from the data. For instance, if the true distribution exhibits heteroskedasticity (that is, $K(Z)$ has a nonconstant diagonal), then $p(\text{overlap} \mid Z, \tilde{Z})$ can violate the overlap criterion even when σ^2 and $\tilde{\sigma}^2$ are estimated; see Section 5.

Finally, consider the case where Z and $\tilde{Z}$ are not necessarily equal and we make no assumptions on $m(Z)$. To avoid trivial failure modes in which the choice of Z and $\tilde{Z}$ leads to a nonnegligible differential bias $v^\top m(Z) - \tilde{v}^\top m(\tilde{Z})$ as N

grows, assume a random design setting where the rows of Z and $\tilde{Z}$ are independent and identically distributed. Then even if $K(Z) = \sigma_\dagger^2 I$, so that there is no heteroskedasticity and no correlation among outcomes, the overlap criterion can still be violated. As before, $p(\text{overlap} \mid Z, \tilde{Z})$ can be arbitrarily small if σ^2 and $\tilde{\sigma}^2$ are blindly set too low, but it can also be small if these variances are estimated. By Eq. (9), $p(\text{overlap} \mid Z, \tilde{Z})$ will be small if the magnitude of

$$\frac{v^\top m(Z) - \tilde{v}^\top m(\tilde{Z})}{(\|v\|^2 + \|\tilde{v}\|^2)^{1/2}} = \frac{u^\top Z^+ f(Z)\beta_\dagger - u^\top \tilde{Z}^+ f(\tilde{Z})\beta_\dagger}{\left(u^\top (Z^\top Z)^{-1}u + u^\top (\tilde{Z}^\top \tilde{Z})^{-1}u\right)^{1/2}} \tag{10}$$

is large relative to $\sqrt{\sigma^2 + \tilde{\sigma}^2}$, where $Z^+ = (Z^\top Z)^{-1}Z^\top$ is the pseudoinverse. A trivial way this can occur is if the entries of $\beta_\dagger$ are large. More interestingly, however, Eq. (10) can be large if the dimension D grows with N, even if each entry of $\beta_\dagger$ has fixed magnitude. Specifically, in Section 5 we present experiments demonstrating this when $\beta_\dagger$ consists of the first D entries of a fixed sequence $\beta_{\dagger,1}, \beta_{\dagger,2}, \dots$ such that $\sum_{d=1}^{D} \beta_{\dagger,d}^2 \to \infty$ as $D \to \infty$.

4. Asymptotic normality of the bagged posterior

In this section, we establish a Bernstein–Von Mises theorem for the bagged posterior under sufficiently regular finite-dimensional models (Theorem 4.1). In particular, we show that while the standard posterior may be arbitrarily under- or over-confident when the model is misspecified, the bagged posterior avoids overconfident uncertainty quantification by accounting for sampling variability.

More formally, consider a model $\{P_\theta : \theta \in \Theta\}$ for independent and identically distributed (i.i.d.) data $x_1, \dots, x_N$, where $x_n \in \mathbb{X}$ and $\Theta \subset \mathbb{R}^D$ is open. Suppose p_θ is the density of P_θ with respect to some reference measure. The standard Bayesian posterior distribution given $x_{1:N}$ is

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

where $\Pi_0(d\theta)$ is the prior distribution and $p(x_{1:N}) := \int \{\prod_{n=1}^{N} p_\theta(x_n)\}\Pi_0(d\theta)$ is the marginal likelihood. We often use the shorthand notation $\Pi_N := \Pi(\cdot \mid x_{1:N})$.

Assume the observed data $X_1, \dots, X_N$ is 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, $\theta_\circ = \arg\max_{\theta \in \Theta} \mathbb{E}\{\log p_\theta(X_1)\}$. Under regularity conditions, the maximum likelihood estimator $\hat{\theta}_N := \arg\max_\theta \prod_{n=1}^{N} p_\theta(X_n)$ is asymptotically normal in the sense that

$$N^{1/2}(\hat{\theta}_N - \theta_\circ) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_{\theta_\circ}^{-1} I_{\theta_\circ} J_{\theta_\circ}^{-1}), \tag{11}$$

where $J_\theta := -\mathbb{E}\{\nabla_\theta^2 \log p_\theta(X_1)\}$, $I_\theta := \text{Cov}\{\nabla_\theta \log p_\theta(X_1)\}$, and $J_{\theta_\circ}^{-1} I_{\theta_\circ} J_{\theta_\circ}^{-1}$ is known as the sandwich covariance [47]. Under mild conditions, the Bernstein–Von Mises theorem (42, Ch. 10 and 28) guarantees that for $\vartheta \sim \Pi_N$,

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

Hence, the standard posterior is correctly calibrated, asymptotically, if the covariance matrices of the Gaussian distributions in Eqs. (11) and (12) coincide – that is, if $J_{\theta_\circ}^{-1} I_{\theta_\circ} J_{\theta_\circ}^{-1} = J_{\theta_\circ}^{-1}$, which is implied by $I_{\theta_\circ} = J_{\theta_\circ}$. 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. (12) 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.

Our Bernstein–von Mises theorem shows that the bagged posterior does not suffer from the overconfidence of the standard posterior. 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. We formally define the bagged posterior $\Pi^*(\cdot \,|\, X_{1:N})$ as

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

for all measurable $A \subseteq \Theta$; this is equivalent to the informal definition in Eq. (1). To avoid notational clutter, we suppress the dependence of $\Pi^*(\cdot \,|\, X_{1:N})$ on M. We 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 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.

For a measure ν and function f, we use the shorthand $\nu(f) := \int f \mathrm{d}\nu$. Let $X_{1:\infty}$ denote the infinite sequence $(X_1, X_2, \dots)$, and abbreviate $\ell_\theta := \log p_\theta$.

Theorem 4.1. *Suppose $X_1, X_2, \dots$ i.i.d. $\sim P_\circ$ and assume that:*

(i) *$(\theta, x) \mapsto \ell_\theta(x)$ is measurable and $\theta \mapsto \ell_\theta(X_1)$ is differentiable at $\theta_\circ$ with probability 1;*

(ii) *there is an open neighborhood U of $\theta_\circ$ and a function $m_{\theta_\circ} : \mathbb{X} \to \mathbb{R}$ such that for some $\delta > 0$ $P_\circ(m_{\theta_\circ}^{2+\delta}) < \infty$ 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) *conditionally on $X_{1:\infty}$, for almost every $X_{1:\infty}$, for every sequence of constants $C_N \to \infty$,*

$$\mathbb{E}\big[\Pi(\{\theta \in \Theta \,:\, \|\theta - \theta_\circ\|_2 > C_N / M^{1/2}\} \,|\, X_{1:M}^*) \,\big|\, X_{1:N}\big] \to 0;$$

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

The result also holds in 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, so $\ell_\theta(x) := \log p_\theta(y \,|\, z)$.

The proof of Theorem 4.1 is in Appendix D. Theorem C.1 is a simpler version of the same result for the univariate Gaussian location model, for which the statement and our proof technique are more transparent. Our technical assumptions are essentially the same as those used by Kleijn and van der Vaart [28] to prove the Bernstein–Von Mises theorem under misspecification for the standard posterior. Of particular note, Kleijn and van der Vaart [28] require that (and give conditions under which) for every sequence of constants $C_N \to \infty$,

$$\mathbb{E}\big[\Pi(\{\theta \in \Theta \,:\, \|\theta - \theta_\circ\|_2 > C_N/N^{1/2}\} \,|\, X_{1:N})\big] \to 0.$$

We conjecture that under reasonable regularity assumptions, this expected posterior concentration condition implies our condition (v).

To interpret this result, it is helpful to compare it to the behavior of the standard posterior. Under the conditions of Theorem 4.1, if $\vartheta \sim \Pi_N$, then $N^{1/2}(\vartheta - \theta_\circ) - \Delta_N \,|\, X_{1:N} \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_{\theta_\circ}^{-1})$ in probability by Kleijn and van der Vaart [28, 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 $J_{\theta_\circ}^{-1}$. Hence, asymptotically, the bagged posterior is never overconfident if $c = 1$ (for instance, if $M = N$) and by Theorem 3.4, we expect $100(1 - \alpha)\%$ credible sets of the bagged posteriors to have overlap probability of at least $1 - \alpha$ when $0 < c \leq 2$.

5. Simulations

In this section, we validate our theoretical results through a simulation study with a linear regression model, which is ideal for investigating the properties of BayesBag since all computations of posterior quantities can be done in closed form. The setup is similar to the linear regression model from Section 3.3 except we place proper priors on the regression coefficients and the outcome variance σ^2. The data consist of regressors $Z_n \in \mathbb{R}^D$ and outcomes $Y_n \in \mathbb{R}$ ($n = 1, \ldots, N$), and the parameter is $\theta = (\theta_0, \ldots, \theta_D) = (\log \sigma^2, \beta_1, \ldots, \beta_D) \in \mathbb{R}^{D+1}$. Using conjugate priors, the assumed model is

$$\sigma^2 \sim \Gamma^{-1}(a_0, b_0)$$
$$\beta_d \,|\, \sigma^2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2/\lambda) \qquad\qquad d = 1, \ldots, D,$$
$$Y_n \,|\, Z_n, \beta, \sigma^2 \overset{\text{indep}}{\sim} \mathcal{N}(Z_n^\top \beta, \sigma^2) \qquad\qquad n = 1, \ldots, N,$$

where $a_0 = 2, b_0 = 1$, and $\lambda = 1$ are fixed hyperparameters.

Data generating distribution. We simulated data for a random design scenario by generating $Z_n \overset{\text{i.i.d.}}{\sim} G$, $\epsilon_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, and

$$Y_n = f(Z_n)^\top \beta_\dagger + \epsilon_n \tag{13}$$

for $n = 1, \ldots, N$, where $\beta_{\dagger d} = 4/\sqrt{d}$ for $d = 1, \ldots, D$ and we used two settings for each of f and G.

- **Regression function f.** By default, we used a linear function $f(z) = z$ to simulate data for the well-specified setting. Alternatively, we used the nonlinear function $f(z) = (z_1^3, \ldots, z_D^3)^\top$ for a misspecified setting.
- **Regressor distribution G.** By default, we used $G = \mathcal{N}(0, I)$ to simulate data; we refer to this as the uncorrelated setting. Alternatively, we used a correlated-κ setting, where, for $h = 10$, $Z \sim G$ was defined by generating $\xi \sim \chi^2(h)$ and then $Z \mid \xi \sim \mathcal{N}(0, \Sigma)$ where $\Sigma_{dd'} = \exp\{-(d - d')^2/\kappa^2\}/(\xi_d \xi_{d'})$ and $\xi_d = \sqrt{\xi/(h-2)}^{\mathbb{1}(d \text{ is odd})}$. The motivation for the correlated-κ sampling procedure is to generate correlated regressors that have different tail behaviors while still having the same first two moments, since regressors are typically standardized to have mean 0 and variance 1. Note that, marginally, $Z_1, Z_3, \ldots$ are each rescaled t-distributed random variables with h degrees of freedom such that $\text{Var}(Z_1) = 1$, and $Z_2, Z_4, \ldots$ are standard normal.

Overlap probabilities. The primary objective in these experiments is to validate that the BayesBag posterior does not violate the probability of overlap lower bounds while the Bayesian posterior sometimes does. Thus, for each data-generating distribution of interest, we estimate overlap probabilities by generating R pairs of datasets $\{(Z_{1:N}^{(r,1)}, Y_{1:N}^{(r,1)}, Z_{1:N}^{(r,2)}, Y_{1:N}^{(r,2)})\}_{r=1}^R$ plus an additional 100 test points $Z_1^{\text{test}}, \ldots, Z_{100}^{\text{test}} \sim G$. If the $1-\alpha$ posterior credible intervals for $(Z_i^{\text{test}})^\top \beta$ conditioned on $(Z_{1:N}^{(r,1)}, Y_{1:N}^{(r,1)})$ and $(Z_{1:N}^{(r,2)}, Y_{1:N}^{(r,2)})$ overlap, set the overlap indicator variable $O_\alpha^{(r)}(Z_i^{\text{test}}) = 1$. Otherwise set $O_\alpha^{(r)}(Z_i^{\text{test}}) = 0$. For each $i \in \{1, \ldots, 100\}$, we estimate the probability of overlap for Z_i^{test} as

$$\mathbb{P}\left(\text{overlap of } (Z_i^{\text{test}})^\top \beta \text{ at level } 1 - \alpha\right) \approx R^{-1} \sum_{r=1}^R O_\alpha^{(r)}(Z_i^{\text{test}}).$$

For all experiments we use $R = 100$. Figure 2 shows that for nonlinear-uncorrelated data, BayesBag never violates the overlap lower bounds while Bayes always or often violates the lower bounds, depending on the value of $1 - \alpha$ (larger $1 - \alpha$ leads to more violations). Figures 8 to 10 in the Supplementary Material show similar results for nonlinear-correlated-κ data, although the problem with Bayes is less severe as the correlation increases. Moreover, as shown in Fig. 3, the problem becomes more severe as N and D jointly increase, but improves or stays the
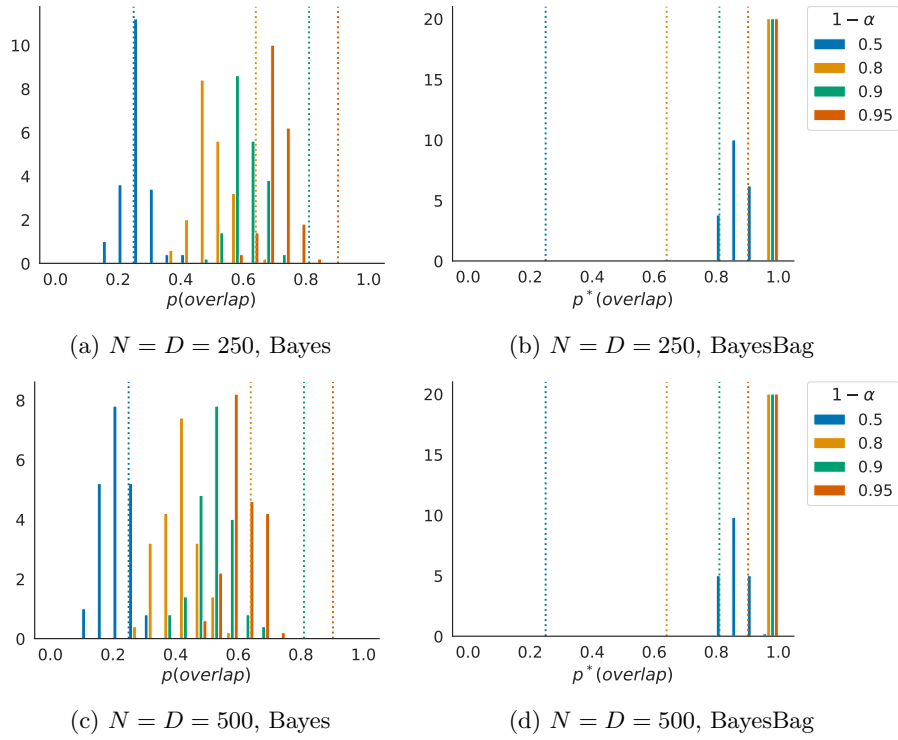
(a) $N = D = 250$, Bayes

(b) $N = D = 250$, BayesBag

(c) $N = D = 500$, Bayes

(d) $N = D = 500$, BayesBag

Fig 2: Histograms of the probability of overlap of $1-\alpha$ credible sets for $(Z_i^{\text{test}})^\top \beta$ $(i = 1, \ldots, 100)$ for the linear regression model with nonlinear-uncorrelated data. Vertical dotted lines indicate the overlap lower bounds $(1 - \alpha)^2$.
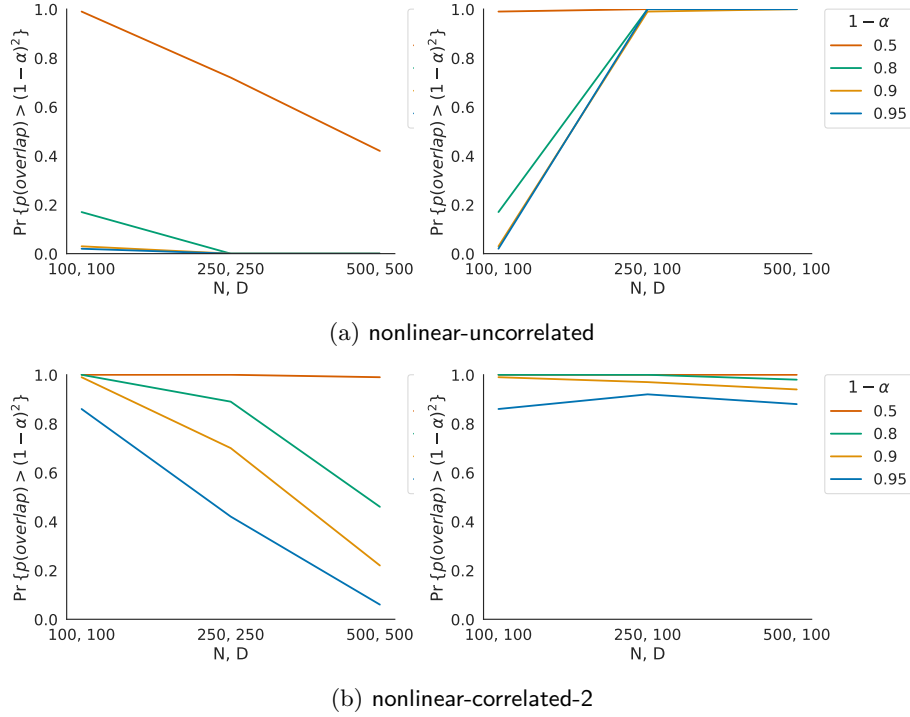
(a) nonlinear-uncorrelated

(b) nonlinear-correlated-2

Fig 3: Proportion of test points Z_i^{test} for which the Bayes overlap probability satisfies the lower bound. For BayesBag, the proportion is 1 in all cases.

same if D is fixed and N increases. These results emphasize how the misspecified high-dimensional regime is particularly problematic for the reproducibility of the standard posterior. We find similar results in the case of a fixed design matrix with heteroskedastic noise (see Appendix B.2 in the Supplementary Material).

Predictive performance. To complement our overlap probability analysis, we also computed the mean log predictive densities at the same test points. Figure 4 shows that while in well-specified linear settings the standard posterior can slightly outperform BayesBag (by roughly 0.2 nats or less), in the misspecified nonlinear settings BayesBag can be far superior (by 0.2 to nearly 10 nats).

6. Application

We next consider an application to community-level crime data from the United States using a Poisson regression model with log link function and the spike-and-slab prior proposed by Piironen and Vehtari [38]. The data consist of $N = 1994$ observations containing 100 community-level covariates such as demographic
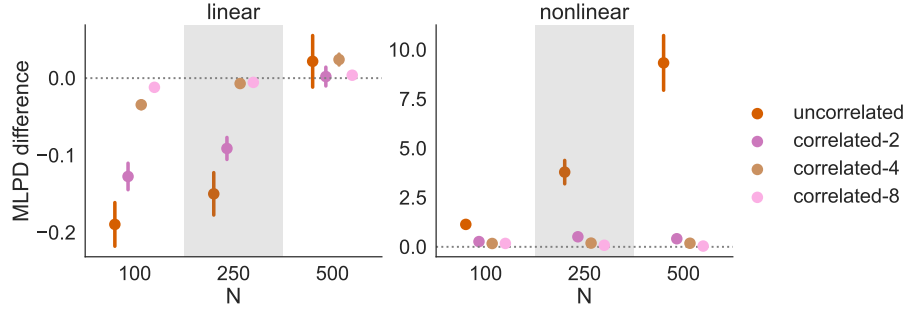
Fig 4: 99% confidence intervals for difference in the mean log predictive densities of the standard and BayesBag posteriors (paired t intervals), with values greater than zero indicating superior performance by BayesBag. Note the different scales for linear versus nonlinear.

summaries and local law enforcement statistics such as the number of police officers per capita. The goal is to predict the number of violent crimes per 100,000 persons in the population. We chose $M = N$ and used $B = 50$ bootstrap samples to approximate the bagged posterior. Nearly identical results were obtained with $B = 25$, indicating that $B = 50$ was sufficiently large.

To compute overlap probabilities, we held out 20% of the observations as test points and randomly split the remaining observations into two equally sized data sets, from which we computed two posteriors to compare. We generated $R = 50$ replicate experiments in this way, and followed the procedure in Section 5 to approximate the overlap probability for each replicate.

Figure 5 validates our theoretical results: the standard posterior is unstable across datasets, with overlap probabilities below $(1 - \alpha)^2$ for $1 - \alpha \in \{0.8, 0.9, 0.95\}$ in the vast majority of replicates. The bagged posteriors, on the other hand, have overlap greater than $(1 - \alpha)^2$ in all replicates. Moreover, BayesBag has superior predictive performance: the mean log predictive densities for the standard and bagged posteriors are, respectively, -5.4 and -4.3 with a 99% confidence interval for difference of $(1.043, 1.093)$ (paired t interval).

To explore how using the bagged rather than the standard posterior might result in different conclusions, we compared the posterior marginals of the regression coefficients, with some representative results shown in Figs. 6 and 7. In all cases, the bagged posteriors were more diffuse, as would be expected. In several cases, however, the BayesBag results are qualitatively different from the standard posterior results. The standard posterior for the coefficient of `Upper Quartile Rent` is symmetric and concentrated below zero while for the bagged posterior it has a sharp peak at zero and is skewed left (Fig. 6). Similarly, the standard posteriors are symmetric for the coefficients of covariates related to percent of different racial and ethnic groups (Fig. 7). Meanwhile, the bagged posterior for the coefficients of `Percent Asian` and `Percent Hispanic` are mul-
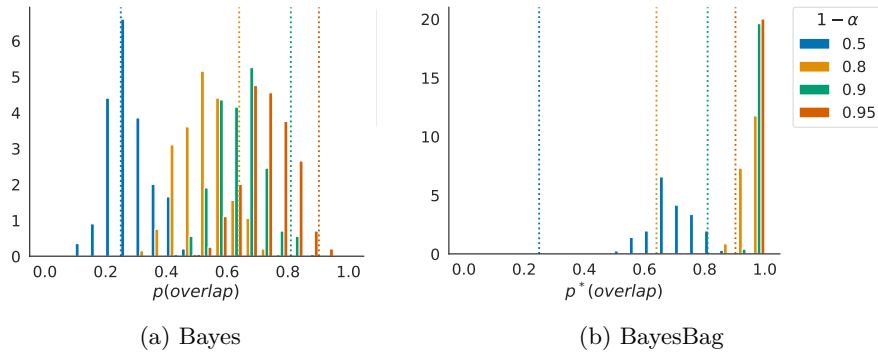
(a) Bayes   (b) BayesBag

Fig 5: For crime data using a sparse Poisson regression model, shown are histograms of the overlap probability for $Z^\top \beta$ where Z is drawn from a held-out test set. For most replicates, the overlap probabilities for the standard posteriors are below $(1 - \alpha)^2$ for $1 - \alpha \in \{0.8, 0.9, 0.95\}$. Meanwhile, for all replicates, the overlap probabilities for the bagged posteriors are greater than $(1 - \alpha)^2$.

timodal and have significantly more mass centered at zero. These examples illustrate how the bagged and standard posteriors may yield substantively different results in practice – BayesBag is not merely inflating the posterior uncertainty.

7. Discussion

We conclude by first situating BayesBag in the wider literature on robust Bayesian inference, and then, with that additional context in place, highlighting the strengths of our approach and suggest fruitful directions for future development.

7.1. Bayesian bagging

Despite the similar sounding names, BayesBag is very different than Bayesian bagging [9, 31]. Bayesian bagging consists of applying the Bayesian bootstrap to a point estimator of a classification or regression model, such as ordinary least squares. In other words, it is a slight variant of traditional bagging where, instead of multinomial weights, one uses continuous weights drawn uniformly from the probability simplex. In contrast, BayesBag uses traditional bagging on the posterior of an arbitrary Bayesian model. In short, Bayesian bagging performs bagging using Bayes, whereas BayesBag performs Bayes using bagging. Relatedly, in the same way that bagging expands the model space for a classification or regression method [14], BayesBag expands the posterior space for a Bayesian model.
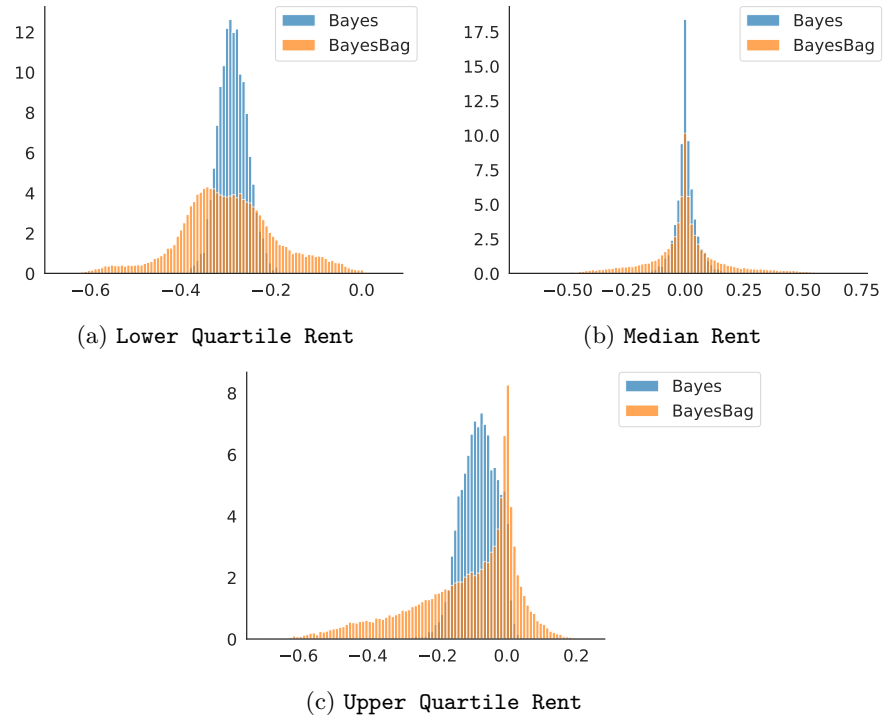
(a) `Lower Quartile Rent`

(b) `Median Rent`

(c) `Upper Quartile Rent`

Fig 6: The standard and bagged posterior marginals for three coefficients related to rental cost for the data and model from Section 6.

(a) `Percent Black`

(b) `Percent White`
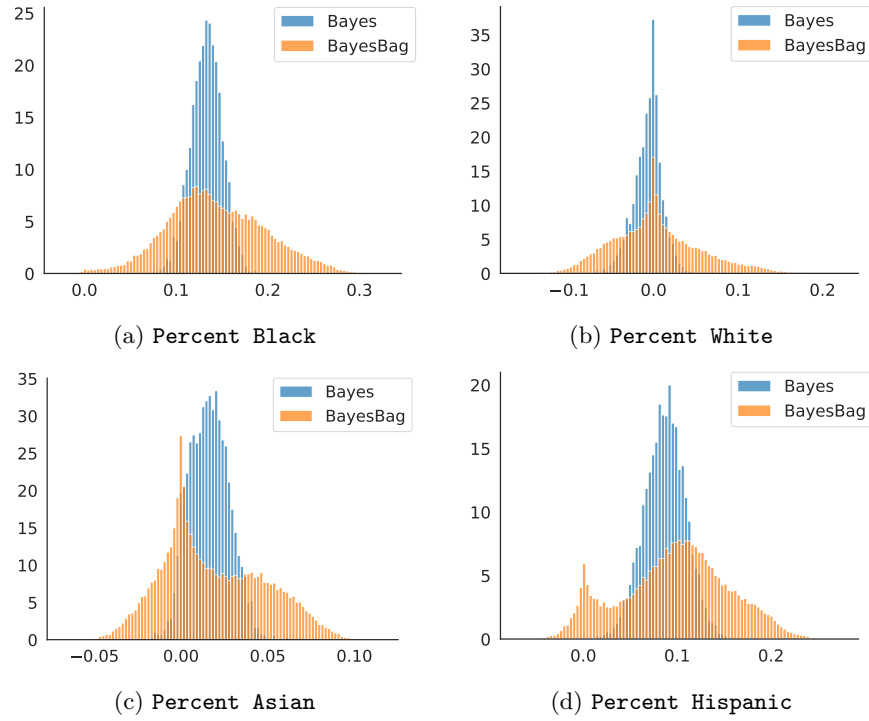
(c) `Percent Asian`

(d) `Percent Hispanic`

Fig 7: The standard and bagged posterior marginals for three coefficients related to race for the data and model from Section 6.

7.2. *Bayesian uncertainty quantification with the bootstrap*

The bootstrap has previously been employed to perform uncertainty quantification in Bayesian settings. See Laird and Louis [30] and references therein for uses of the bootstrap to adjust for underestimated uncertainties when using empirical Bayesian methods. Similar in spirit to the present work, Efron [17] develops a variety of methods for obtaining frequentist uncertainty quantification of Bayesian point estimates, including some that rely on bootstrapping.

7.3. *Robust Bayesian inference*

Two common themes emerge when surveying existing methods for robust Bayesian inference. First, many methods require choosing a free parameter, and the proposals for choosing free parameters tend to be 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 lose key parts of what makes 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 [19, 20, 22, 33, 35, 41]. 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 $\zeta \in (0, 1)$ [2, 19, 35, 39, 46]. 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 [1, 35]. Second, choosing an appropriate value of ζ can be difficult. Grünwald [19] proposes SafeBayes, a theoretically sound method which is evaluated empirically in Grünwald and van Ommen [20] and de Heide et al. [12]. 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 [22, 33, 35, 39, 41].

More in the spirit of BayesBag are a number of bootstrapped point estimation approaches [8, 33, 34, 37, 40]. 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 [37] and a generalization proposed by Lyddon, Holmes and Walker [33] do not incorporate a prior, and therefore lose many of the benefits of Bayesian inference. The related approach of Lyddon, Walker and Holmes [34], 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 [36] suggests replacing the standard posterior by a Gaussian distribution with covariance proportional to a plug-in estimate of the sandwich covariance. A benefit of our approach is that it does not rely on a Gaussian approximation and does not require estimation of the sandwich covariance, making it suitable for small-sample settings. While our theory does focus on Gaussian or asymptotically Gaussian posteriors, in practice BayesBag is applicable in non-asymptotic regimes where the posterior is highly non-Gaussian, as shown by the application in Section 6.

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 added requirement is the design of a bootstrapping scheme. Although this paper focuses on using BayesBag with independent observations, future work can draw on the large literature devoted to adapting the bootstrap to more complex models such as those involving time-series and spatial data. 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; see Huggins and Miller [23] for how to use BayesBag for model selection.

Another appeal of BayesBag as a methodology is that the only hyperparameter – the bootstrap dataset size M – is straightforward to set. Specifically, $M = N$ is a natural, theoretically well-justified choice that, while slightly conservative, yields reproducible inferences.

In terms of computation, when using the approximation in Eq. (2), there is an additional cost due to the need to compute the posterior for each bootstrapped dataset. However, it is trivial to compute the bootstrapped posteriors in parallel. As described in Appendix A.1, 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. Nonetheless, speeding up BayesBag with more specialized computational methods could be worthwhile in some applications. For example, in Appendix A.2, we suggest one simple approach to speeding up Markov chain Monte Carlo (MCMC) runs when using BayesBag. Pierre Jacob has proposed using more advanced unbiased MCMC techniques for potentially

even greater computational efficiency.[1]

Another 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, straightforward integration over nuisance parameters, and the use of prior information. Further, our Jeffrey conditionalization interpretation establishes solid epistemological foundations for using BayesBag. Thus, it provides an appealing and philosophically coherent synthesis of Bayesian and frequentist approaches without introducing difficult-to-choose tuning parameters and without sacrificing the most useful parts of Bayesian inference.

Appendix A: Computation

A.1. Choosing the number of bootstrap datasets for BayesBag

If we wish to use Eq. (2) to approximate the bagged posterior, then we must 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., 29]. 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 it may be desirable to reduce computational expense by balancing the number of bootstrap samples B versus the accuracy of the approximation to $\pi(\theta \,|\, x^*_{(b)})$ (e.g., in terms of the length of Markov chain Monte Carlo runs). We discuss this computational trade-off next.

A.2. A BayesBag sampling algorithm

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.

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

---

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

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 it is usually desirable to obtain a high-quality approximation to the standard posterior anyway.

Algorithm 1 Basic BayesBag Sampler

Require: A Markov chain Monte Carlo procedure $\text{MCMC}(x, T, \theta_{\text{init}}, \beta_{\text{init}})$ 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 samples M, number of bootstrap datasets B, initial hyperparameters β_{init}
1: $\beta, \theta_{1:T} \leftarrow \text{MCMC}(x, T, \beta_{\text{init}})$
2: **for** $b = 1, \ldots, B$ **do**
3:      Generate a new bootstrap dataset $x^*_{(b)}$ of size M from x
4:      Sample $\theta^*_{(b)\text{init}}$ uniformly from $\theta_{1:T}$
5:      $\beta_{(b)}, \theta^*_{(b)1:T^*} \leftarrow \text{MCMC}(x^*_{(b)}, T^*, \theta^*_{(b)\text{init}}, \beta)$
6: **end for**
7: $\theta^*_{1:BT^*} \leftarrow \text{concatenate}(\theta^*_{(1)1:T^*}, \ldots, \theta^*_{(B)1:T^*})$
8: **return** posterior samples $\theta_{1:T}$ and BayesBag samples $\theta^*_{1:BT^*}$

Appendix B: Additional experimental results

B.1. Additional linear regression simulations

Figures 8 to 10 show similar results for nonlinear-correlated-κ data to what Fig. 2 shows for nonlinear-uncorrelated data, although the problem with Bayes is less severe as the correlation increases.

B.2. Fixed design linear regression simulations

To simulate data for a fixed design scenario, we set $z_{n0} = 1$ to include an intercept, set covariates z_{n1} and z_{n2} to be a uniform grid on $[-2, 2] \times [-2, 2]$, and generate the remaining covariates as i.i.d. $\mathcal{N}(0, 1)$. We use the (well-specified) linear regression function $f(z) = z$ and to introduce misspecification, we generate the outcomes as in Eq. (13) but with heteroskedastic noise given by $\epsilon_n \,|\, z_n \overset{\text{indep}}{\sim} \mathcal{N}(0, 1 + z_{n1}^2 + z_{n2}^2)$. Figure 11 shows that standard Bayes exhibits poor overlap behavior, similar to the case of nonlinear-correlated-2 data (Fig. 8), whereas BayesBag has overlap probability very close to 1 at every test point. BayesBag also has superior predictive performance, with 99% confidence intervals for the difference in mean log predictive densities of $(0.49, 0.69)$ and $(1.00, 1.38)$ for, respectively, $N = D = 256$ and 400.
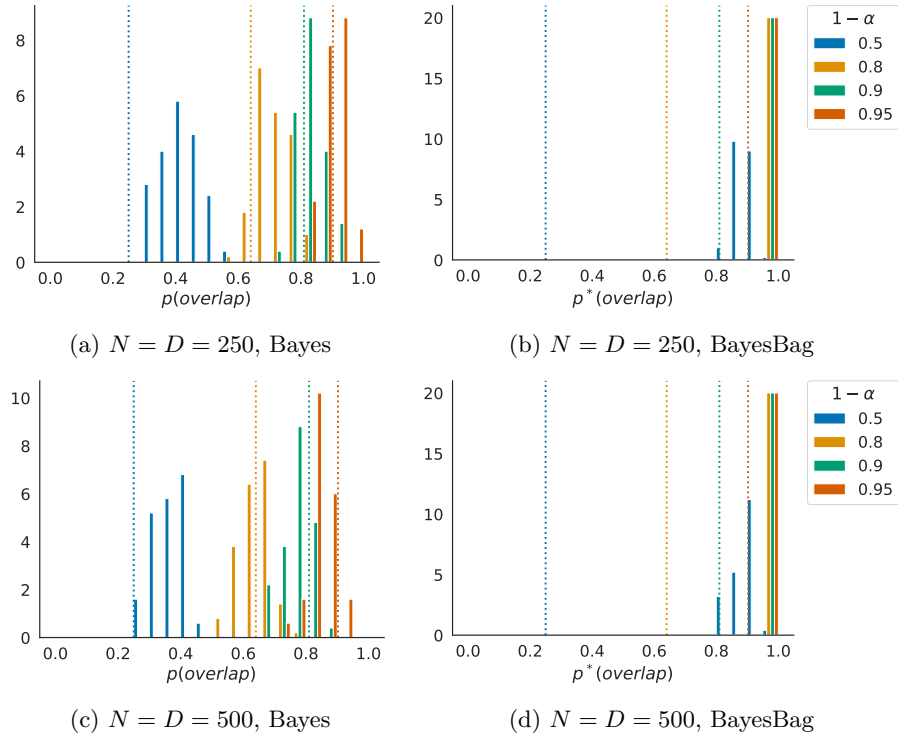
(a) $N = D = 250$, Bayes

(b) $N = D = 250$, BayesBag

(c) $N = D = 500$, Bayes
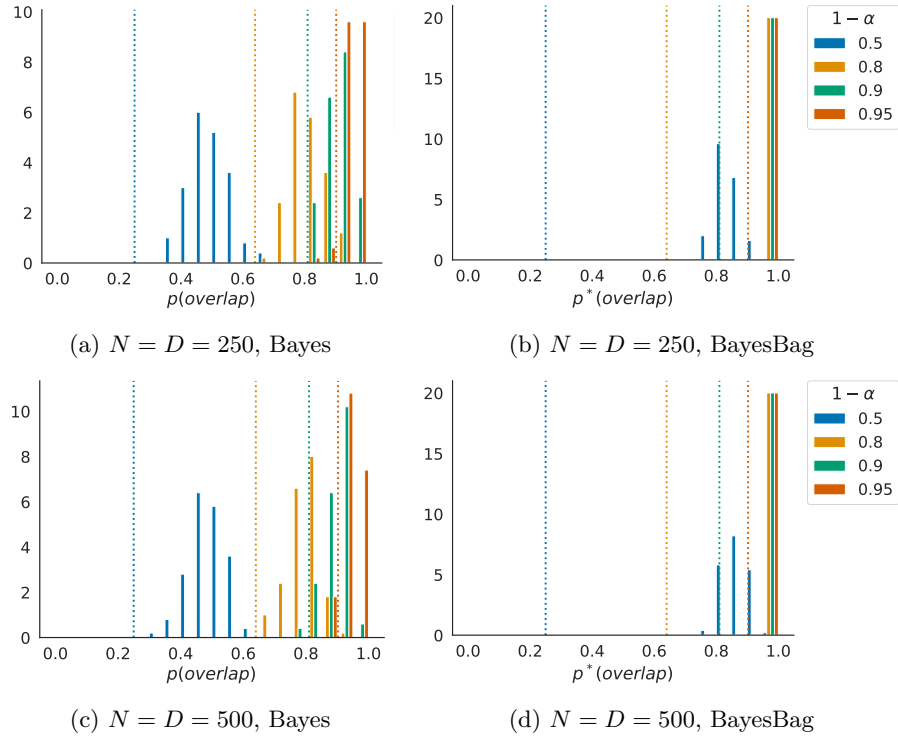
(d) $N = D = 500$, BayesBag

Fig 8: Histograms of the probability of overlap of $1-\alpha$ credible sets for $(Z_i^{\text{test}})^\top \beta$ $(i = 1, \ldots, 100)$ for linear regression with nonlinear-correlated-2 data.
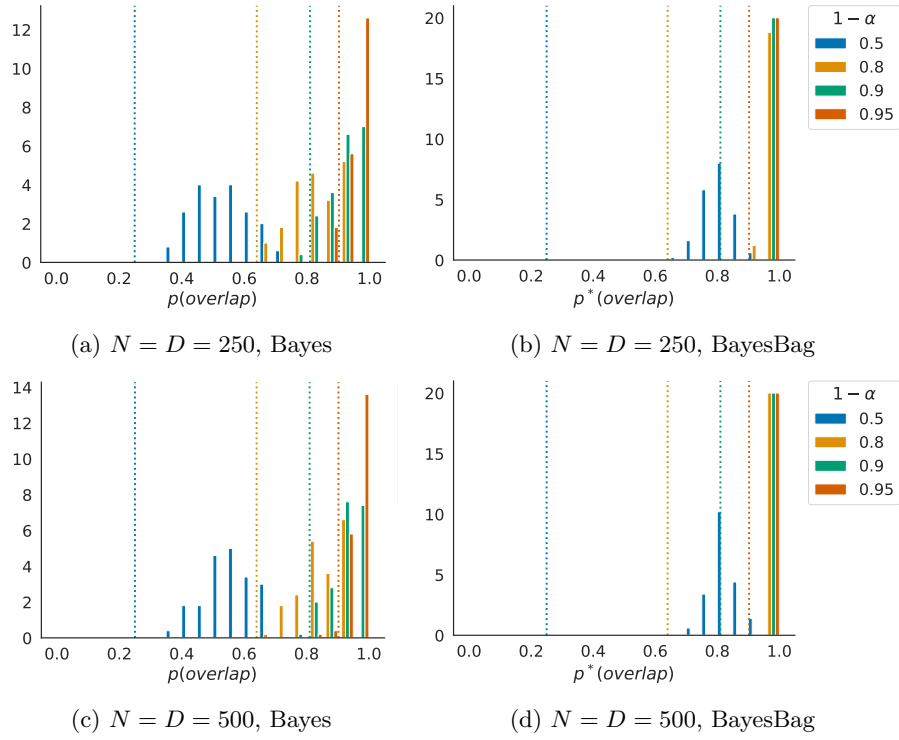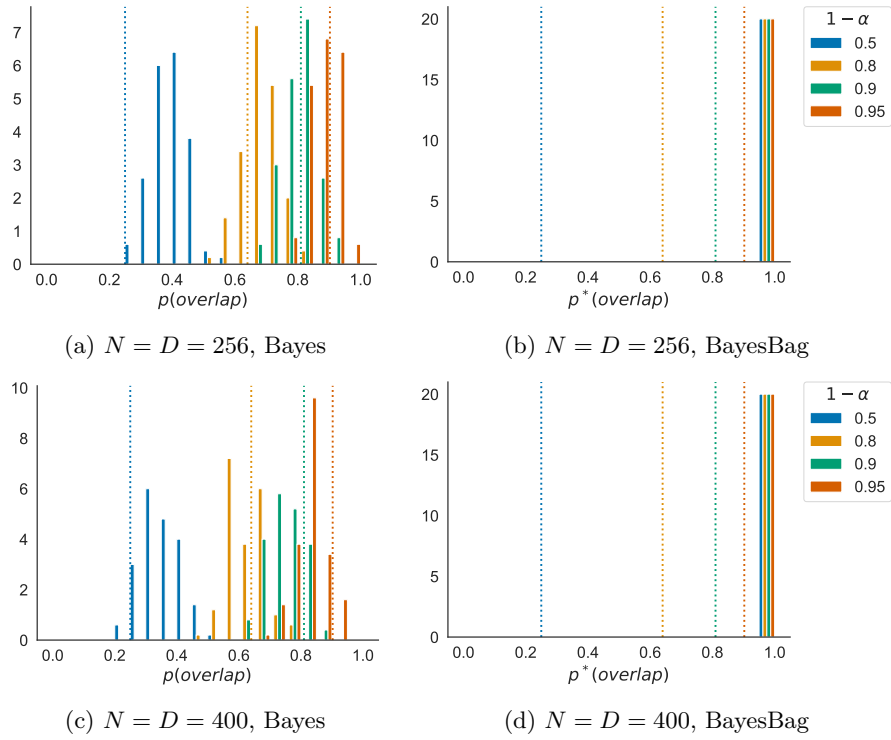
(a) $N = D = 250$, Bayes

(b) $N = D = 250$, BayesBag

(c) $N = D = 500$, Bayes

(d) $N = D = 500$, BayesBag

Fig 9: Histograms of the probability of overlap of $1-\alpha$ credible sets for $(Z_i^{\text{test}})^\top \beta$ $(i = 1, \ldots, 100)$ for linear regression with nonlinear-correlated-4 data.

(a) $N = D = 250$, Bayes

(b) $N = D = 250$, BayesBag

(c) $N = D = 500$, Bayes

(d) $N = D = 500$, BayesBag

Fig 10: Histograms of the probability of overlap of $1 - \alpha$ credible sets for $(Z_i^{\text{test}})^\top \beta$ $(i = 1, \ldots, 100)$ for linear regression with nonlinear-correlated-8 data.

(a) $N = D = 256$, Bayes

(b) $N = D = 256$, BayesBag

(c) $N = D = 400$, Bayes

(d) $N = D = 400$, BayesBag

Fig 11: Histograms of the probability of overlap of $1 - \alpha$ credible sets for $(Z_i^{\text{test}})^\top \beta$ $(i = 1, \dots, 100)$ for linear regression with linear mean function f, fixed design, and heteroskedastic error.

Appendix C: BayesBag Bernstein–Von Mises Theorem for Gaussian Location Model

Theorem C.1. *Let $X_1, X_2, \ldots \in \mathbb{R}$ i.i.d. such that for some $\delta > 0$, $\mathbb{E}(|X_1|^{2+\delta}) < \infty$. Consider the Gaussian location model from Section 3.1. Let $\vartheta^* \mid X_{1:N} \sim \Pi_N^*$ and suppose $c := \lim_{N \to \infty} M/N \in (0, \infty)$ for $M = M(N)$. Then for almost every $(X_1, X_2, \ldots)$,*

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

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 $N^{1/2}$.

Proof of Theorem C.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

$$\psi_{\mathcal{N}(\mu,\sigma^2)}(t) = \exp(i\mu t - \sigma^2 t^2/2), \qquad t \in \mathbb{R}. \tag{14}$$

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

$$\psi_{\mathsf{Multi}(L,p)}(t) = \left(\sum_{k=1}^K p_k e^{it_k}\right)^L, \qquad t \in \mathbb{R}^K. \tag{15}$$

Let $\widetilde{\Pi}(\cdot \mid X_{1:M}^*) := \mathcal{N}(N^{1/2}R_M(\bar{X}_M^* - \bar{X}_N), NV_M)$, noting that this is the distribution of $N^{1/2}\{\vartheta^* - \mathbb{E}(\vartheta^* \mid X_{1:N})\} \mid X_{1:M}^*$. Similarly, let $\widetilde{\Pi}^*(\cdot \mid X_{1:N})$ denote the distribution of $N^{1/2}\{\vartheta^* - \mathbb{E}(\vartheta^* \mid X_{1:N})\} \mid X_{1:N}$. Let $Y_{Nn} := N^{1/2}R_M(X_n - \bar{X}_N)$ and let $K_{1:N} \sim \mathsf{Multi}(M, 1/N)$. Using Eqs. (14) and (15), we have

$$
\begin{aligned}
\psi_{\widetilde{\Pi}^*(\cdot \mid X_{1:N})}(t) &= \mathbb{E}\{\psi_{\widetilde{\Pi}(\cdot \mid X_{1:M}^*)}(t) \mid X_{1:N}\} \\
&= \mathbb{E}\big[\exp\{itM^{-1}\textstyle\sum_{n=1}^N K_n Y_{Nn} - NV_M t^2/2\} \mid X_{1:N}\big] \\
&= \left\{\frac{1}{N}\sum_{n=1}^N \exp(itM^{-1}Y_{Nn})\right\}^M \exp(-NV_M t^2/2). \tag{16}
\end{aligned}
$$

Let $\hat{V}_N := N^{-1}\sum_{n=1}^N (X_n - \bar{X}_N)^2$. By Durrett [16, Lemma 3.3.19], $e^{is} = 1 + is - s^2/2 + \mathcal{R}(s)$ where $\mathcal{R}(s) \leq \min(|s|^3, 2|s|^2)$. Since $N^{-1}\sum_{n=1}^N Y_{Nn} = 0$, the first factor of Eq. (16) can be expanded as

$$
\begin{aligned}
&\left\{\frac{1}{N}\sum_{n=1}^N \left(1 + itM^{-1}Y_{Nn} - \tfrac{1}{2}t^2 M^{-2}Y_{Nn}^2 + \mathcal{R}(tM^{-1}Y_{Nn})\right)\right\}^M \\
&= \left\{1 - \tfrac{1}{2}t^2 \frac{NR_M^2}{M^2}\hat{V}_N + \frac{1}{N}\sum_{n=1}^N \mathcal{R}(tM^{-1}Y_{Nn})\right\}^M. \tag{17}
\end{aligned}
$$

The remainder term $\sum_{n=1}^{N} \mathcal{R}(tM^{-1}Y_{Nn})$ is bounded by

$$\frac{t^2 R_M^2 N}{M^2} \max_{n\in\{1,\dots,N\}} \min(N^{1/2}|t||X_n - \bar{X}_N|/M, 2) \sum_{n=1}^{N}(X_n - \bar{X}_N)^2$$

and by Lemma C.2, $\limsup_{N\to\infty}\max_{n=1,\dots,N} X_n/N^{1/2} \overset{a.s.}{\to} 0$. By the strong law of large numbers

$$\limsup_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} |X_n - \bar{X}_N|^2 \overset{a.s.}{<} \infty.$$

Combining these bounds with the that fact that $M/N \to c$ and $R_M \to 1$, we conclude that 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 $\hat{V}_N \overset{a.s.}{\to} \mathrm{Var}(X_1)$ 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. (17) converges to $\exp(-\frac{1}{2}t^2 \mathrm{Var}(X_1)/c)$. Combining this with Eq. (16), and noting that $NV_M \to V/c$, we have that almost surely, for all $t \in \mathbb{R}$,

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

The result follows by Lévy's continuity theorem [27, Theorem 5.3]. $\qquad\square$

Lemma C.2. *Suppose $X_1, X_2, \dots \in \mathbb{R}$ i.i.d. such that for some $\delta > 0$, $\mathbb{E}(|X_1|^{2+\delta}) < \infty$. Then $\max_{n\in\{1,\dots,N\}} |X_n - \bar{X}_N|/N^{1/2} \overset{a.s.}{\to} 0$ as $N \to \infty$.*

Proof. Define $Y_n := |X_n|/n^{1/2}$. By Markov's inequality, for all $\varepsilon > 0$,

$$\mathbb{P}(Y_n \geq \varepsilon) \leq \frac{\mathbb{E}(|X_1|^{2+\delta})}{n^{1+\delta/2}\varepsilon^{2+\delta}}.$$

Hence, $\sum_{n=1}^{\infty}\mathbb{P}(Y_n \geq \varepsilon) < \infty$, so by the Borel–Cantelli lemma, $\limsup_n Y_n \leq \varepsilon$ almost surely. Since $\varepsilon > 0$ is arbitrary, this implies that $\limsup_n Y_n = 0$ almost surely. Now, let $\varepsilon' > 0$ and define N_* (depending on $Y_1, Y_2, \dots$) such that $\sup_{n>N_*} Y_n \leq \varepsilon'$. Letting $M = \max_{n\leq N_*} Y_n$, we have that almost surely,

$$\begin{aligned}
\max_{n\leq N} |X_n - \bar{X}_N|/N^{1/2} &\leq \max_{n\leq N} 2|X_n|/N^{1/2} \\
&\leq \max_{n\leq N_*} 2|X_n|/N^{1/2} + \max_{N_*<n\leq N} 2|X_n|/N^{1/2} \\
&\leq 2MN_*^{1/2}/N^{1/2} + \max_{N_*<n\leq N} 2Y_n \\
&\leq 3\varepsilon'
\end{aligned}$$

for all N sufficiently large. Therefore, since $\varepsilon' > 0$ is arbitrary,

$$\limsup_{N\to\infty} \max_{n\leq N} |X_n - \bar{X}_N|/N^{1/2} = 0$$

almost surely. $\qquad\square$

Appendix D: Proofs

Proof of Proposition 2.2. Denote $\eta_\circ = \eta(P_\circ)$. Since X and Y are independent, it follows that $\mathbb{1}(\eta_\circ \in A_X)$ and $\mathbb{1}(\eta_\circ \in B_Y)$ are independent. Thus,

$$\begin{aligned}
\mathbb{P}(A_X \cap B_Y \neq \varnothing) &\geq \mathbb{P}(\eta_\circ \in A_X \cap B_Y) \\
&= \mathbb{P}(\eta_\circ \in A_X, \, \eta_\circ \in B_Y) \\
&= \mathbb{P}(\eta_\circ \in A_X)\mathbb{P}(\eta_\circ \in B_Y) \\
&\geq (1-\alpha)(1-\alpha').
\end{aligned}$$

$\square$

Proof of Theorem 3.1. Since X and $\tilde{X}$ are independent and identically distributed given ϑ,

$$\begin{aligned}
\mathbb{E}\big(\mathbb{P}(A_X \cap A_{\tilde{X}} \neq \varnothing \mid \vartheta)\big) &\geq \mathbb{E}\big(\mathbb{P}(\vartheta \in A_X, \, \vartheta \in A_{\tilde{X}} \mid \vartheta)\big) \\
&= \mathbb{E}\big(\mathbb{P}(\vartheta \in A_X \mid \vartheta)\, \mathbb{P}(\vartheta \in A_{\tilde{X}} \mid \vartheta)\big) \\
&= \mathbb{E}\big(\mathbb{P}(\vartheta \in A_X \mid \vartheta)^2\big) \\
&\geq \mathbb{E}\big(\mathbb{P}(\vartheta \in A_X \mid \vartheta)\big)^2 \\
&= \mathbb{E}\big(\mathbb{P}(\vartheta \in A_X \mid X)\big)^2 \\
&\geq (1-\alpha)^2,
\end{aligned}$$

where in the last step we use that $\mathbb{P}(\vartheta \in A_x \mid x) \geq 1 - \alpha$ for all x. $\square$

Proof of Theorem 3.2. To handle both the standard Bayes and BayesBag cases simultaneously, consider a multivariate normal posterior on θ with mean $R_M \bar{X}_N \in \mathbb{R}^D$ and covariance matrix $V_M + bM^{-1}R_M\hat{\Sigma}_N R_M$; then standard Bayes is the case of $M = N$ and $b = 0$, while BayesBag is the case of $b = 1$. The posterior of $u^\top \theta$ is then

$$\mathcal{N}(u^\top R_M \bar{X}_N, \, \sigma^2_{X_{1:N}})$$

where $\sigma^2_{X_{1:N}} := u^\top V_M u + bM^{-1}u^\top R_M \hat{\Sigma}_N R_M u$. Thus, a $100(1-\alpha)\%$ credible interval for $u^\top \theta$ is given by

$$A^b_{X_{1:N}} = u^\top R_M \bar{X}_N \pm z_{\alpha/2}\, \sigma_{X_{1:N}}.$$

Letting $X_{1:N}$ and $Y_{1:N}$ be independent data sets drawn i.i.d. from $P_\circ$, we have

$$\mathbb{P}(A^b_{X_{1:N}} \cap A^b_{Y_{1:N}} \neq \varnothing) = \mathbb{P}\Big(\big|u^\top R_M \bar{X}_N - u^\top R_M \bar{Y}_N\big| \leq z_{\alpha/2}(\sigma_{X_{1:N}} + \sigma_{Y_{1:N}})\Big). \tag{18}$$

By the central limit theorem, $N^{1/2}(\bar{X}_N - \bar{Y}_N) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2\Sigma_\circ)$. By assumption, $M/N \to c > 0$ as $N \to \infty$, which implies that $M \to \infty$. Recalling that $R_M = (V_0^{-1}V/M + I)^{-1}$ and $V_M = (V_0^{-1} + MV^{-1})^{-1}$, we have $R_M \to I$ and $NV_M \to$

V/c as $N \to \infty$. Thus, by the strong law of large numbers, $N^{1/2}\sigma_{X_{1:N}} \to (u^\top V u/c + b u^\top \Sigma_\circ u/c)^{1/2}$ almost surely as $N \to \infty$, and likewise for $N^{1/2}\sigma_{Y_{1:N}}$. Therefore, by Slutsky's theorem, as $N \to \infty$,

$$\mathbb{P}(A^b_{X_{1:N}} \cap A^b_{Y_{1:N}} \neq \varnothing) \longrightarrow \mathbb{P}\Big(\big|\mathcal{N}(0, 2u^\top\Sigma_\circ u)\big| \leq z_{\alpha/2} 2(u^\top V u/c + b u^\top \Sigma_\circ u/c)^{1/2}\Big)$$

$$= \mathbb{P}\bigg(|W| \leq z_{\alpha/2}\sqrt{2}\Big(\frac{u^\top((V + b\Sigma_\circ)/c)u}{u^\top\Sigma_\circ u}\Big)^{1/2}\bigg)$$

where $W \sim \mathcal{N}(0,1)$. This proves the theorem. $\qquad\square$

Proof of Theorem 3.3. Since $X_n, Y_n \sim \mathcal{N}(0, \Sigma_\circ)$ i.i.d., we have $u^\top \bar{X}_N - u^\top \bar{Y}_N \sim \mathcal{N}(0, 2u^\top\Sigma_\circ u/N)$. Thus, for the standard posterior, by setting $V = I$, $V_0^{-1} = 0$, $\|u\| = 1$, $M = N$, and $b = 0$ in Eq. (18), we have

$$\mathbb{P}(A_{X_{1:N}} \cap A_{Y_{1:N}} \neq \varnothing) = \mathbb{P}\Big(\big|u^\top \bar{X}_N - u^\top \bar{Y}_N\big| \leq z_{\alpha/2}\, 2/\sqrt{N}\Big)$$

$$= \mathbb{P}\Big(\big|(2u^\top\Sigma_\circ u/N)^{1/2}W\big| \leq z_{\alpha/2}\, 2/\sqrt{N}\Big)$$

$$= \mathbb{P}\big(|W| \leq z_{\alpha/2}\sqrt{2}/(u^\top\Sigma_\circ u)^{1/2}\big)$$

since $R_M = I$ and $\sigma^2_{X_{1:N}} = 1/N$. This proves the first part.

For the bagged posterior, define $X'_n = u^\top X_n$ and $Y'_n = u^\top Y_n$. Letting $s^2_{X'}$ denote the sample variance of X'_n, we have

$$\sigma^2_{X_{1:N}} = M^{-1} + M^{-1}u^\top\hat{\Sigma}_N u \geq M^{-1}N^{-1}\sum_{n=1}^{N}(X'_n - \bar{X}'_N)^2 = M^{-1}s^2_{X'},$$

and likewise for $\sigma^2_{Y_{1:N}}$, since $V = I$, $V_0^{-1} = 0$, $\|u\| = 1$, and $b = 1$. Thus,

$$\sigma_{X_{1:N}} + \sigma_{Y_{1:N}} \geq (\sigma^2_{X_{1:N}} + \sigma^2_{Y_{1:N}})^{1/2} \geq M^{-1/2}(s^2_{X'} + s^2_{Y'})^{1/2}.$$

Hence, by Eq. (18),

$$\mathbb{P}(A^*_{X_{1:N}} \cap A^*_{Y_{1:N}} \neq \varnothing) = \mathbb{P}\Big(\big|u^\top \bar{X}_N - u^\top \bar{Y}_N\big| \leq z_{\alpha/2}(\sigma_{X_{1:N}} + \sigma_{Y_{1:N}})\Big) \qquad (19)$$

$$\geq \mathbb{P}\Big(\big|\bar{X}'_N - \bar{Y}'_N\big| \leq z_{\alpha/2}\, M^{-1/2}(s^2_{X'} + s^2_{Y'})^{1/2}\Big).$$

Letting $s^2 = u^\top\Sigma_\circ u$, we have that $\sqrt{N}\bar{X}'_N/s = N^{-1/2}\sum_{n=1}^{N}X'_n/s \sim \mathcal{N}(0,1)$ and $Ns^2_{X'}/s^2 \sim \chi^2_{N-1}$ independently, by Cochran's theorem. Since X'_n and Y'_n are independent, $\sqrt{N/(2s^2)}(\bar{X}'_N - \bar{Y}'_N) \sim \mathcal{N}(0,1)$ and $(N/s^2)(s^2_{X'} + s^2_{Y'}) \sim \chi^2_{2N-2}$ independently. Hence,

$$\sqrt{N-1}\,\frac{\bar{X}'_N - \bar{Y}'_N}{(s^2_{X'} + s^2_{Y'})^{1/2}} \overset{\mathcal{D}}{=} \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_{2N-2}/(2N-2)}} \overset{\mathcal{D}}{=} T_{2N-2}. \qquad (20)$$

Combining Eqs. (19) and (20) yields

$$\mathbb{P}(A^*_{X_{1:N}} \cap A^*_{Y_{1:N}} \neq \varnothing) \geq \mathbb{P}\left(\frac{|\bar{X}'_N - \bar{Y}'_N|}{(s^2_{X'} + s^2_{Y'})^{1/2}} \leq z_{\alpha/2}\, M^{-1/2}\right)$$
$$= \mathbb{P}\left(|T_{2N-2}| \leq z_{\alpha/2}\sqrt{(N-1)/M}\right),$$

as claimed. $\qquad\square$

Proof of Theorem 3.4. The proof is similar to the Gaussian location model, but simpler in some ways since we are already in the asymptotic regime. First, consider the standard posterior. Denote $V = J^{-1}_{\theta_\circ}$ and $\Sigma = J^{-1}_{\theta_\circ} I_{\theta_\circ} J^{-1}_{\theta_\circ}$. Since $N^{1/2}(\vartheta - \hat{\theta}_N)\,|\,X_{1:N} \xrightarrow{\mathcal{D}} \mathcal{N}(0, V)$ by assumption, a $100(1 - \alpha)\%$ credible interval for $u^\top\theta$ based on the asymptotic normal distribution is $A_{X_{1:N}} = u^\top\hat{\theta}_N \pm z_{\alpha/2}(u^\top V u/N)^{1/2}$. Likewise, $N^{1/2}(\hat{\theta}_N - \theta_\circ) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma)$ by assumption. Thus, letting $\hat{\theta}_X$ and $\hat{\theta}_{\tilde{X}}$ be the maximum likelihood estimators based on independent data sets $X_{1:N}$ and $\tilde{X}_{1:N}$, we have $N^{1/2}(u^\top\hat{\theta}_X - u^\top\hat{\theta}_{\tilde{X}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 2u^\top\Sigma u)$ by the continuity theorem. Thus, letting $W \sim \mathcal{N}(0, 1)$, the asymptotic overlap probability for the standard posterior is

$$p_\infty(\text{overlap}) = \lim_{N\to\infty} \mathbb{P}\left(|u^\top\hat{\theta}_X - u^\top\hat{\theta}_{\tilde{X}}| \leq 2z_{\alpha/2}(u^\top V u/N)^{1/2}\right)$$
$$= \lim_{N\to\infty} \mathbb{P}\left(|N^{1/2}(u^\top\hat{\theta}_X - u^\top\hat{\theta}_{\tilde{X}})| \leq 2z_{\alpha/2}(u^\top V u)^{1/2}\right)$$
$$= \mathbb{P}\left(|(2u^\top\Sigma u)^{1/2}W| \leq 2z_{\alpha/2}(u^\top V u)^{1/2}\right)$$
$$= \mathbb{P}\left(|W| \leq \sqrt{2}z_{\alpha/2}\left(\frac{u^\top V u}{u^\top\Sigma u}\right)^{1/2}\right),$$

as claimed. For the bagged posterior, we have $N^{1/2}(\vartheta^* - \hat{\theta}_N)\,|\,X_{1:N} \xrightarrow{\mathcal{D}} \mathcal{N}(0, V/c + \Sigma/c)$ by assumption. Hence, the proof is the same but with $V/c + \Sigma/c$ in place of V. The claimed inequality for the bagged posterior holds because $u^\top V u \geq 0$, due to the fact that V is positive semi-definite. $\qquad\square$

Proof of Theorem 3.5. Let $\mu_\dagger = m(Z)$, $\tilde{\mu}_\dagger = m(\tilde{Z})$, $\Sigma_\dagger = K(Z)$, and $\tilde{\Sigma}_\dagger = K(\tilde{Z})$. We have

$$p(\text{overlap} \mid Z, \tilde{Z}) = \mathbb{P}(A \cap \tilde{A} \neq \varnothing \mid Z, \tilde{Z})$$
$$= \mathbb{P}\left(|v^\top Y - \tilde{v}^\top\tilde{Y}| \leq z_{\alpha/2}(\sigma\|v\| + \tilde{\sigma}\|\tilde{v}\|)\,\Big|\,Z, \tilde{Z}\right) \qquad (21)$$
$$= \mathbb{P}\left(\left|W + \frac{v^\top\mu_\dagger - \tilde{v}^\top\tilde{\mu}_\dagger}{(v^\top\Sigma_\dagger v + \tilde{v}^\top\tilde{\Sigma}_\dagger\tilde{v})^{1/2}}\right| \leq \frac{z_{\alpha/2}(\sigma\|v\| + \tilde{\sigma}\|\tilde{v}\|)}{(v^\top\Sigma_\dagger v + \tilde{v}^\top\tilde{\Sigma}_\dagger\tilde{v})^{1/2}}\right).$$

where $W \sim \mathcal{N}(0, 1)$. If $m(Z) = Z\beta_\dagger$ and $K(Z) = \sigma^2_\dagger I$, then

$$v^\top\mu_\dagger - \tilde{v}^\top\tilde{\mu}_\dagger = u^\top Z^+ Z\beta_\dagger - u^\top\tilde{Z}^+\tilde{Z}\beta_\dagger = u^\top\beta_\dagger - u^\top\beta_\dagger = 0 \qquad (22)$$

where $Z^+ := (Z^\top Z)^{-1} Z^\top$ denotes the Moore–Penrose pseudoinverse of Z. Therefore, by Eqs. (21) and (22),

$$p(\text{overlap} \mid Z, \tilde{Z}) = \mathbb{P}\left(|W| \leq \frac{z_{\alpha/2}(\sigma\|v\| + \tilde{\sigma}\|\tilde{v}\|)}{\sigma_\dagger(\|v\|^2 + \|\tilde{v}\|^2)^{1/2}}\right),$$

which proves Eq. (7).

Suppose $Z = \tilde{Z}$, but we make no assumptions on the form of $m(Z)$ or $K(Z)$. Then $v = \tilde{v}$, $\mu_\dagger = \tilde{\mu}_\dagger$, and $\Sigma_\dagger = \tilde{\Sigma}_\dagger$. Therefore, by Eq. (21),

$$p(\text{overlap} \mid Z, \tilde{Z}) = \mathbb{P}\left(|W| \leq \frac{z_{\alpha/2}(\sigma + \tilde{\sigma})\|v\|}{\sqrt{2}(v^\top \Sigma_\dagger v)^{1/2}}\right),$$

proving Eq. (8).

Suppose $K(Z) = \sigma_\dagger^2 I$, but we make no assumptions on the form of $m(Z)$. Then by the Cauchy–Schwarz inequality,

$$\sigma\|v\| + \tilde{\sigma}\|\tilde{v}\| \leq \sqrt{\sigma^2 + \tilde{\sigma}^2}(\|v\|^2 + \|\tilde{v}\|^2)^{1/2} = \frac{\sqrt{\sigma^2 + \tilde{\sigma}^2}}{\sigma_\dagger}(v^\top \Sigma_\dagger v + \tilde{v}^\top \tilde{\Sigma}_\dagger \tilde{v})^{1/2},$$

so by Eq. (21),

$$p(\text{overlap} \mid Z, \tilde{Z}) \leq \mathbb{P}\left(\left|W + \frac{v^\top \mu_\dagger - \tilde{v}^\top \tilde{\mu}_\dagger}{\sigma_\dagger(\|v\|^2 + \|\tilde{v}\|^2)^{1/2}}\right| \leq \frac{z_{\alpha/2}\sqrt{\sigma^2 + \tilde{\sigma}^2}}{\sigma_\dagger}\right),$$

which proves Eq. (9). $\qquad \square$

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.

Proof of Theorem 4.1. We use the shorthand notation $\ell_\theta := \log p_\theta$, and denote the gradient and Hessian by $\dot{\ell}_\theta := \nabla_\theta \ell_\theta$ and $\ddot{\ell}_\theta := \nabla_\theta^2 \ell_\theta$, respectively. 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

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

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 [42, 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 [43, 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

$$\mathbb{G}_N^* \{N^{1/2} \lambda_N - h_N^\top \dot{\ell}_\circ\} \overset{P_+}{\to} 0, \tag{23}$$

where we recall that $\overset{P_+}{\to}$ denotes convergence in outer probability. By the proof of Kleijn and van der Vaart [28, Lemma 2.1],

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

and, following the same reasoning, we can expand the lefthand side of Eq. (23) and multiply though by $c^{1/2}$ to get

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

and hence

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

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

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

We apply Kleijn and van der Vaart [28, Theorem 2.1] to conclude that, letting $\vartheta^{*\prime} | X_{1:M}^* \sim \Pi(\cdot \mid X_{1:M}^*)$, the total variation distance between the distribution of $N^{1/2}(\vartheta^{*\prime} - \theta_\circ) \mid X_{1:M}^*$ 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 [28, Theorem 2.1], we have $X_{1:M}^*$ in place of $X^{(n)}$, $\mathbb{P}_N^M$ in place of $P_0^{(n)}$, $c J_\circ$ 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 $N^{1/2}(\vartheta^* - \theta_\circ) - \Delta_N \mid X_{1:N}$ evaluated at $t \in \mathbb{R}^D$ as

$$\begin{aligned}
&\mathbb{E} \left[\exp\left\{ i \Delta_N^{*\top} t - t^\top J_\circ^{-1} t/(2c) \right\} \mid X_{1:N} \right] + \epsilon_N(t) \\
&= \mathbb{E} \left[\exp\left\{ i N^{1/2} \mathbb{P}_N^* \dot{\ell}_\circ^\top J_\circ^{-1} t \right\} \mid X_{1:N} \right] \exp\left\{ -i N^{1/2} \mathbb{P}_N \dot{\ell}_\circ^\top J_\circ^{-1} t \right\} \\
&\quad \times \exp\left\{ -t^\top J_\circ^{-1} t/(2c) \right\} + \epsilon_N(t).
\end{aligned} \tag{24}$$

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. (24) to get

$$
\mathbb{E}\left[\exp\left\{iN^{1/2}M^{-1}\sum_{n=1}^{N}K_n\dot{\ell}_\circ(X_n)^\top J_\circ^{-1}t\right\}\,\middle|\,X_{1:N}\right]\exp\left\{-iN^{1/2}\mathbb{P}_N\dot{\ell}_\circ^\top J_\circ^{-1}t\right\}
$$

$$
=\left[\frac{1}{N}\sum_{n=1}^{N}\exp\left\{\frac{iN^{1/2}\dot{\ell}_\circ(X_n)^\top J_\circ^{-1}t}{M}\right\}\right]^M\exp\left\{-iN^{1/2}\mathbb{P}_N\dot{\ell}_\circ^\top J_\circ^{-1}t\right\}
$$

$$
=\left[\frac{1}{N}\sum_{n=1}^{N}\exp\left\{\frac{iN^{1/2}\delta\dot{\ell}_\circ(X_n)^\top J_\circ^{-1}t}{M}\right\}\right]^M
$$

$$
=\left[\frac{1}{N}\sum_{n=1}^{N}\left\{1+\frac{iN^{1/2}\delta\dot{\ell}_\circ(X_n)^\top J_\circ^{-1}t}{M}-\frac{N(\delta\dot{\ell}_\circ(X_n)^\top J_\circ^{-1}t)^2}{2M^2}+\mathcal{R}_n\right\}\right]^M
$$

$$
=\left\{1-\frac{Nt^\top J_\circ^{-1}\mathbb{P}_N(\delta\dot{\ell}_\circ\delta\dot{\ell}_\circ^\top)J_\circ^{-1}t}{2M^2}+\mathcal{R}_n\right\}^M, \tag{25}
$$

where (recalling the notation from the proof of Theorem C.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 C.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. (24) and (25), 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 [27, Theorem 5.3]. □

Acknowledgments

Funding

Supplementary Material

Experiment code
Python code and data to recreate all numerical experiments.

References

[1] Antoniano-Villalobos, I. and Walker, S. G. (2013). Bayesian Nonparametric Inference for the Power Likelihood. *Journal of Computational and Graphical Statistics* **22** 801–813.

[2] Bhattacharya, A., Pati, D. and Yang, Y. (2019). Bayesian fractional posteriors. *The Annals of Statistics* **47** 39–66.

[3] 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.

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

[5] 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.

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

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

[8] Chamberlain, G. and Imbens, G. (2003). Nonparametric applications of Bayesian inference. *Journal of Business Economic Statistics* **21** 12–18.

[9] Clyde, M. and Lee, H. (2001). Bagging and the Bayesian bootstrap. In *International Workshop on Artificial Intelligence and Statistics* 57–62. PMLR.

[10] Cox, D. R. (1990). Role of Models in Statistical Analysis. *Statistical Science* **5** 169–174.

[11] De Blasi, P. and Walker, S. G. (2013). Bayesian asymptotics with misspecified models. *Statistica Sinica* 1–19.

[12] de Heide, R., Kirichenko, A., Mehta, N. and Grünwald, P. D. (2019). Safe-Bayesian Generalized Linear Regression. *arXiv.org* **arXiv:1910.09227 [math.ST]**.

[13] Diaconis, P. and Zabell, S. L. (1982). Updating subjective probability. *Journal of the American Statistical Association* **77** 822–830.

[14] DOMINGOS, P. M. (1997). Why Does Bagging Work? A Bayesian Account and its Implications. In *KDD* 155–158.

[15] 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.

[16] DURRETT, R. (2019). *Probability: Theory and Examples. Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press.

[17] EFRON, B. (2015). Frequentist accuracy of Bayesian estimates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **77** 617–646.

[18] GRECO, L., RACUGNO, W. and VENTURA, L. (2008). Robust likelihood functions in Bayesian inference. *Journal of Statistical Planning and Inference* **138** 1258 – 1270. https://doi.org/10.1016/j.jspi.2007.05.001

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

[20] 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.

[21] HOFF, P. and WAKEFIELD, J. (2012). Bayesian sandwich posteriors for pseudo-true parameters. *arXiv preprint arXiv:1211.0087.*

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

[23] HUGGINS, J. H. and MILLER, J. W. (2023). Reproducible Model Selection Using Bagged Posteriors. *Bayesian Analysis* **18** 79–104.

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

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

[26] JEWSON, J., SMITH, J. Q. and HOLMES, C. (2018). Principles of Bayesian Inference Using General Divergence Criteria. *Entropy* **20** 442.

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

[28] 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.

[29] 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.

[30] LAIRD, N. M. and LOUIS, T. A. (1987). Empirical Bayes Confidence Intervals Based on Bootstrap Samples. *Journal of the American Statistical Association* **82** 739–750.

[31] LEE, H. K. and CLYDE, M. A. (2004). Lossless online Bayesian bagging. *Journal of Machine Learning Research* **5** 143–151.

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

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

[34] 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*.

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

[36] 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.

[37] 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.

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

[39] 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.

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

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

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

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

[44] 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.

[45] WALKER, S. G. (2013). Bayesian inference with misspecified models. *Journal of statistical planning and inference* **143** 1621–1633.

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

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

[48] 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.