Fast approximate BayesBag model selection via Taylor expansions

Neil A. Spencer and Jeffrey W. Miller
Department of Biostatistics,
Harvard School of Public Health, Boston, MA 15213
nspencer@hsph.harvard.edu, jwmiller@hsph.harvard.edu

Abstract

BayesBag has been established as a useful tool for robust Bayesian model selection. However, computing BayesBag can be prohibitively expensive for large datasets. Here, we propose a fast approximation of BayesBag model selection. This approximation—based on Taylor approximations of the log marginal likelihood—can achieve results comparable to BayesBag in a fraction of the time.

1 Motivation

A classical result by Berk [1966] established that Bayesian model selection concentrates on whichever model is closest to the truth in Kullback-Leibler divergence, even if all the models are incorrect. However, when multiple models are equally incorrect, there is no guarantee that the posterior will distribute itself in a reasonable manner. Huggins and Miller [2020] showed that the posterior shifts back-and-forth between models as additional data is considered, placing almost all mass on whichever model has higher observed likelihood. Similar behavior can occur even if one model is truly better, so long as the models are sufficiently close [Huggins and Miller, 2020, Giordano, 2020]. Therefore, Bayesian model selection can be both unreliable and unstable whenever the models are misspecified.

In recent years, BayesBag [Bühlmann, 2014] has emerged as a flexible tool for robust Bayesian inference. By averaging posteriors across many bootstrapped datasets, BayesBag posteriors account for both Bayesian uncertainty and frequentist sampling variation. BayesBag thus serves as an effective safeguard against overconfidence due to model mis-specification, in both model selection [Huggins and Miller, 2020] and parameter inference [Huggins and Miller, 2019] settings. For model selection, the averaging across bootstrapped datasets tends to stabilize the posterior, distributing it more evenly among the plausible models. Given at least 100 bootstrap samples, this approach can be far more reliable than standard Bayes [Huggins and Miller, 2020].

On the other hand, naïvely computing BayesBag model selection can be much more expensive. If each bootstrap posterior is computed independently, the difference amounts to roughly two orders of magnitude. Even with parallelization, this can be prohibitive for large problems. At the very least, it can slow down one's workflow considerably. In this work, our goal is to make BayesBag model selection more accessible and convenient by developing a fast Taylor expansion-based approximation.

2 Problem Statement

We consider Bayesian model selection under the same exchangeable data setting as Huggins and Miller [2020]. Given N independent draws $Y = (y_1, \dots, y_N)$ from an unknown distribution, the goal is to choose among K candidate models $(\mathcal{M}_k)_{k\in[K]}$ defined by

$$\mathcal{M}_k: \text{ each } y_n \text{ has distribution } f_k(y_n \mid \theta_k) \text{ for } k \in [K],$$

NeurIPS 2021 Workshop Your Model Is Wrong: Robustness and Misspecification in Probabilistic Modeling.

where θ_k represents any unknown parameters. Using $p_k(\theta_k)$ to denote the prior distribution on θ_k and π_k to denote the prior probability of $\mathcal{M}_k$, the standard Bayesian model selection posterior for model $\mathcal{M}_k$ is given by $\mathbb{P}_{\vec{1}}(\mathcal{M}_k \mid Y = y)$. Here, $\vec{1}$ denotes a vector composed of N ones, and

$$\mathbb{P}_w(\mathcal{M}_k \mid Y = y) := \frac{\pi_k Z_k(y \mid w)}{\sum_{\ell=1}^{K} \pi_\ell Z_\ell(y \mid w)} \tag{1}$$

with $Z_k(y \mid w)$ denoting the marginal w-weighted likelihood of y under model $\mathcal{M}_k$. That is,

$$Z_k(y \mid w) := \int \left(\prod_{n=1}^{N} f_k(y_n \mid \theta_k)^{w_n} \right) p(\theta_k) \mathrm{d}\theta_k. \tag{2}$$

For BayesBag model selection, the posterior probabilities are instead given by

$$\mathbb{P}_{r(w)}(\mathcal{M}_k \mid Y = y) := \sum_{w} \mathbb{P}_w(\mathcal{M}_k \mid Y = y) r(w), \tag{3}$$

where any bootstrap distribution $r(w)$ can be specified by the user. The standard choice for $r(w)$ is $r(w) = \text{Multinomial}(M, \vec{1}/N)$, with $M \approx N^{0.95}$ providing stable results [Huggins and Miller, 2020]. The entries in w thus encode the number of copies of each observation in a bootstrap sample.

Note, however, that (3) is typically an intractable sum. One can instead Monte Carlo approximate $r(w)$ using a $\hat{r}(w)$ composed of $B \approx 100$ Monte Carlo draws $w^1, \ldots, w^B \sim r(w)$. Thus, we can focus on developing a fast approximation technique for the target

$$\mathbb{P}_{\hat{r}(w)}(\mathcal{M}_k \mid Y = y) := B^{-1} \sum_{b=1}^{B} \mathbb{P}_{w^b}(\mathcal{M}_k \mid Y = y). \tag{4}$$

The challenge in computing (4) stems from the fact that $\mathbb{P}_{w^b}(\mathcal{M}_k \mid Y = y)$—as defined in (1)—usually lacks a closed form. Instead, each summand must be approximated using techniques that can be very expensive in high-dimensional settings, such as reversible jump MCMC [Green, 1995] or direct marginal likelihood approximation [Llorente et al., 2020] for each $k \in [K]$. Accordingly, our goal is to reduce the number of b's for which $\mathbb{P}_{w^b}(\mathcal{M}_k \mid Y = y)$ must be explicitly computed.

Before proceeding, it is important to recognize that many strategies for approximating (1) (e.g. [Green, 1995, Geyer, 1994, Dai et al., 2020]) rely on generating posterior draws from the distribution of $\theta_k \mid Y = y$ as a byproduct. Access to these draws turn out to be very helpful going forward, as they allow us to evaluate useful posterior expectations without any additional posterior sampling.

3 Method: Taylor Expansions

Before computing BayesBag, suppose that we have already computed standard Bayesian model selection over the K candidate models. We thus have access to accurate estimates of $\mathbb{P}_{\vec{1}}(\mathcal{M}_k \mid Y = y)$ for all $k \in [K]$. Suppose we also have access to requisite posterior draws for computing expectations $\mathbb{E}_{\vec{1}}\left(\alpha^T \ell_k(\theta_k)\right)$ and $\text{Var}_{\vec{1}}\left(\alpha^T \ell_k(\theta_k)\right)$—for any $\alpha \in \mathbb{R}^N$—where $\ell_k(\theta_k)$ denotes the vector $(\log(f_k(y_1 \mid \theta_k)), \ldots, \log(f_k(y_N \mid \theta_k)))$ of log likelihoods under model $\mathcal{M}_k$. $\mathbb{E}_{\vec{1}}(\cdot)$ and $\text{Var}_{\vec{1}}(\cdot)$ denote the expectation and variance respectively under the standard posterior given $Y = y$.

Borrowing an insight from Campbell and Beronov [2019], we can interpret $\log(Z_k(y \mid w))$ as the log partition function of an exponential family distribution for θ_k with parameter vector w and sufficient statistics $\ell_k(\theta_k)$. It immediately follows that

$$\left.\frac{\mathrm{d}\log(Z_k(y \mid w)))}{\mathrm{d}w}\right|_{w=\vec{1}} = \mathbb{E}_{\vec{1}}\left(\ell_k(\theta_k)\right) \text{ and } \left.\frac{\mathrm{d}^2\log(Z_k(y \mid w)))}{\mathrm{d}w^2}\right|_{w=\vec{1}} = \text{Cov}_{\vec{1}}\left(\ell_k(\theta_k)\right).$$

Recalling that each $\mathbb{P}_{\vec{1}}(\mathcal{M}_k \mid Y = y)$ is proportional to $Z_k(y \mid w = \vec{1})$ via (1), we can compute second order Taylor approximations of the marginal likelihoods $(\log(Z_k(y \mid w^b)))_{k \in [K]}$ up to a common additive constant for each bootstrap draw. That is, for some $C_b \in \mathbb{R}$,

$$\log(Z_k(y \mid w^b)) \approx \log\left(\mathbb{P}_{\vec{1}}(\mathcal{M}_k \mid Y = y)\right) - \log(\pi_k) + t_1^k(\vec{1}, w^b) + \frac{1}{2} t_2^k(\vec{1}, w^b) + C_b,$$

$$\text{where } t_1^k(\vec{1}, w) := \mathbb{E}_{\vec{1}}\left((w - \vec{1})^T \ell_k(\theta_k)\right) \text{ and } t_2^k(\vec{1}, w) := \text{Var}_{\vec{1}}\left((w - \vec{1})^T \ell_k(\theta_k)\right).$$

These approximations can in turn be plugged into (1) to derive the approximation

$$\hat{\mathbb{P}}_{w^b}(\mathcal{M}_k \mid Y = y) = \frac{\mathbb{P}_{\vec{1}}(\mathcal{M}_k \mid Y = y) \exp\left(t_1^k(\vec{1}, w^b) + \frac{1}{2} t_2^k(\vec{1}, w^b)\right)}{\sum_{r=1}^K \mathbb{P}_{\vec{1}}(\mathcal{M}_r \mid Y = y) \exp\left(t_1^r(\vec{1}, w^b) + \frac{1}{2} t_2^r(\vec{1}, w^b)\right)}. \tag{5}$$

We illustrate in Section 4 that the above approximation can be quite accurate. Therefore, using it for each summand in (4) is often enough to accurately approximate BayesBag model selection without doing any explicit posterior computation beyond that of standard Bayes.

To assess the accuracy of this BayesBag approximation, we propose a workflow in which one explicitly performs posterior computation for a few bootstrap samples, then checks if the results agree with the approximation in (5). When choosing which b's to check, we prioritize those with highest Δ_b scores, where Δ_b denotes the absolute difference between (5) and an analogous approximation based on a first order Taylor expansion. A large Δ_b suggests that $\log(Z_k(y \mid w^b))$ is not well approximated by a Taylor expansion centered at $\log(Z_k(y \mid w = \vec{1}))$. We briefly elaborate on this in Section 5.

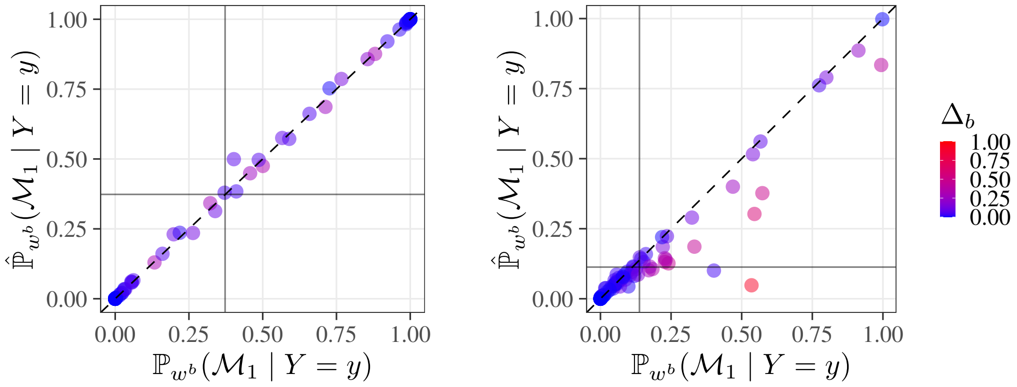
Finally, it is worth acknowledging that our approach resembles Laplace's method [Tierney and Kadane, 1986] in that we use a Taylor expansion to approximate the marginal likelihood. However, our approach expands $\log(Z_k(y \mid w))$ around $w = \vec{1}$ rather $\ell_k(\theta_k)$ around the MAP of θ_k.

4 Simulations

For our first experiment, we consider a variable selection task for linear regression. We generate $N = 1000$ synthetic observations $y_{n \in [N]}$ from a linear regression model with ten covariates $x_1, \ldots, x_{10}$. These covariate values are all generated independently from a standard normal distribution. Likewise, the values for the intercept, the regression coefficients for $\beta_1, \ldots, \beta_9$, and the noise terms are also generated from a standard normal. We manually set $\beta_{10} = 1$ to ensure a nontrivial value.

To define the selection task, we treat x_{10} as unobserved. Instead, we generate new covariates x_{11} and x_{12} such that $\text{Corr}(x_{10}, x_{11}) = \text{Corr}(x_{10}, x_{12}) = \text{Corr}(x_{11}, x_{12}) = 0.9$. Our task is then to choose between two conjugate linear regression models: $\mathcal{M}_1$ containing x_{11}, or $\mathcal{M}_2$ containing x_{12}. Each model also contains an intercept, coefficients for $x_1, \ldots, x_9$, and an unknown noise variance σ^2. To facilitate ground truth comparisons, we employ conjugate priors: $\sigma^2 \sim \text{InverseGamma}(1, 1)$ and $\beta_d \sim N(0, \sigma^2)$ for all d. Figure 1(a) demonstrates the accuracy of our approximation for $B = 100$.

For a second experiment, we generate data using the same set-up as above, except with the regression noise following a t-distribution with df $= 3$. Figure 1(b) depicts the results of this experiment.



(a) Experiment 1

(b) Experiment 2

Figure 1: Plots comparing the Taylor expansion-based probability approximation (vertical axis) to the ground truth (horizontal axis) where each point depicts a single bootstrap sample. The solid lines mark the average along each axis, and the dashed diagonal line marks the realm of perfect approximation. To illustrate the heuristic discussed in Section 3, each point is colored by Δ_b.

For a final experiment, we consider model selection in logistic regression. We generate $N = 1000$ binary observations $y_{n\in[N]}$ according to a contaminated logistic regression model with nine unit variance covariates $x_1, \ldots, x_9$. The covariates are multivariate normal distributed—the first six are independent, while x_7, x_8, and x_9 share pairwise correlations of 0.9. The true regression coefficients for $x_1, \ldots, x_6$ are also normally distributed. We manually set $\beta_0 = -5$, $\beta_7 = \beta_8 = 1$, and $\beta_9 = 0.5$. Finally, we randomly select one percent of the y observations to zero out as contamination.

Our corresponding model selection task concerns four candidate logistic regression models: $\mathcal{M}_1$ contains the intercept and covariates x_1 through x_7, $\mathcal{M}_2$ includes the intercept and x_1 through x_8, $\mathcal{M}_3$ includes the intercept and all covariates but x_8, and $\mathcal{M}_4$ includes everything. All parameters have standard normal priors, except the intercept's prior has variance of 3. Each model has prior probability of $\pi_k = 0.25$. Figure 2 depicts the approximation accuracy results[1] for this experiment.

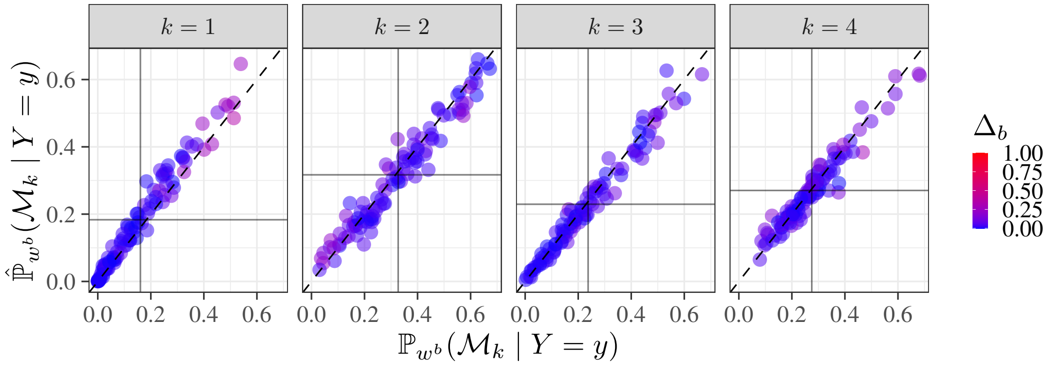


Figure 2: Plots comparing the Taylor expansion-based probability approximation (vertical axis) to the ground truth (horizontal axis) for each bootstrap sample in Experiment 3. The four potential models are separated by panel. Otherwise, the format follows that of Figures 1(a) and 1(b).

5 Conclusions and Remarks

Our Taylor expansion-based approximation achieved accurate results in all three experiments. Some b's in Experiment 2 exhibited downward bias, but not enough to drastically impact the mean. Moreover, the offending b's mostly exhibited larger values of Δ_b, and could thus be flagged for full computation. Our diagnostic Δ_b is closely related to the Fisher information metric [Amari, 2016].

Our technique may not be practical in all cases, such as when the $\mathbb{E}_{\vec{1}}(\ell_k(\theta_k))$ terms are not directly computable from the MCMC draws. This can occur if a Gibbs update is used to impute observation-specific auxiliary variables, leaving evaluation of $\ell_k(\theta_k)$ intractable.

Furthermore, our technique can have problems if the standard posterior probabilities are computed using a MCMC algorithm over all models (e.g. high dimensional variable selection). If the chain takes a prohibitively long time to visit all low-probability models, some will be left with "zero" estimated probability in the standard posterior. This would guarantee zero probabilities in their Taylor approximations, even if they truly exhibit high probability under a bootstrapped dataset. A possible workaround would be to expand around a power posterior [Miller and Dunson, 2018] instead.

Finally, it is worth noting that the exact speed-up factor of our strategy depends on the relative expense of (A): computing Equation (1), versus (B): computing the expectations required for the Taylor expansions. Our approximation will provide the most savings when (B)'s cost is negligible relative to (A), such as those involving high dimensional non-conugate posteriors [Linderman et al., 2016, Oaks et al., 2019]. Our approximation has less utility for simple problems where conjugate priors are available. Note that in order to demonstrate accuracy by comparing against the ground truth, our simulations in Section 4 considered cases where the marginal likelihoods are available in closed form. As such, we did not include direct time comparisons.

[1] To compute our ground truth posterior probabilities for Experiment 3, we implemented a Pólya-Gamma [Polson et al., 2013] augmented Gibbs sampler over the models

Acknowledgements

We would like to thank Ryan Giordano, Jonathan Huggins, and Max Russo for helpful conversations.

References

Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

Robert H Berk. Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics*, pages 51–58, 1966.

Peter Bühlmann. Discussion of big Bayes stories and BayesBag. *Statistical Science*, 29(1):91–94, 2014.

Trevor Campbell and Boyan Beronov. Sparse variational inference: Bayesian coresets from scratch. *Advances in Neural Information Processing Systems*, 32:11461–11472, 2019.

Chenguang Dai, Jeremy Heng, Pierre E Jacob, and Nick Whiteley. An invitation to sequential Monte Carlo samplers. *arXiv preprint arXiv:2007.11936*, 2020.

Charles J Geyer. Estimating normalizing constants and reweighting mixtures in Markov Chain Monte Carlo. *Technical Report No. 568 School of Statistics University of Minnesota*, 1994.

Ryan Giordano. Asymptotics of the log likelihood ratio and a Bayesian model selection "paradox"., Jan 2020. URL `https://rgiordan.github.io/bayes/2020/01/15/likelihood_ratio_model_selection.html`.

Peter J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Jonathan H Huggins and Jeffrey W Miller. Robust inference and model criticism using bagged posteriors. *arXiv preprint arXiv:1912.07104*, 2019.

Jonathan H Huggins and Jeffrey W Miller. Robust and reproducible model selection using bagged posteriors. *arXiv preprint arXiv:2007.14845*, 2020.

Scott W Linderman, Ryan P Adams, and Jonathan William Pillow. Bayesian latent structure discovery from multi-neuron recordings. *Advances in Neural Information Processing Systems*, pages 2010–2018, 2016.

Fernando Llorente, Luca Martino, David Delgado, and Javier Lopez-Santiago. Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *arXiv preprint arXiv:2005.08334*, 2020.

Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 2018.

Jamie R Oaks, Kerry A Cobb, Vladimir N Minin, and Adam D Leaché. Marginal likelihoods in phylogenetics: A review of methods and applications. *Systematic Biology*, 68(5):681, 2019.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504): 1339–1349, 2013.

Luke Tierney and Joseph B Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.