

# Robust inference and model selection using bagged posteriors

Jeff Miller

Joint work with Jonathan Huggins

Harvard T.H. Chan School of Public Health  
Department of Biostatistics

University of Washington || Biostatistics Seminar || Jan 13, 2022

Slides: <http://jwmi.github.io/talks/uw2022.pdf>

Preprint 1: <https://arxiv.org/abs/1912.07104>

Preprint 2: <https://arxiv.org/abs/2007.14845>

# Outline

- 1 Motivation
- 2 Background
- 3 Methodology (Bagged posteriors)
- 4 Theory
- 5 Applications
  - Variable selection
  - Phylogenetic tree inference
  - Hierarchical mixed effects logistic regression

# Outline

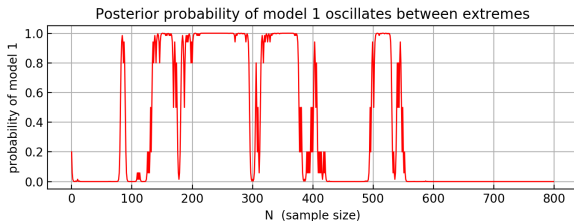
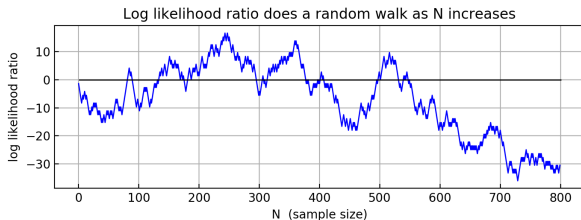
- 1 Motivation
- 2 Background
- 3 Methodology (Bagged posteriors)
- 4 Theory
- 5 Applications
  - Variable selection
  - Phylogenetic tree inference
  - Hierarchical mixed effects logistic regression

# Motivation

- Standard Bayesian inference is known to be sensitive to model misspecification.
- This leads to unreliable uncertainty quantification and poor predictive performance.
- Several methods exist for robust Bayesian inference under misspecification.
- However, finding generally applicable and computationally feasible methods is a difficult challenge.

# Toy Bernoulli example

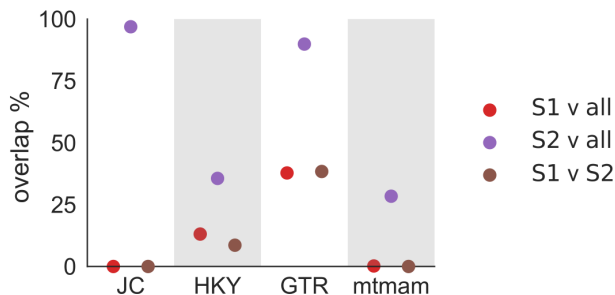
- Suppose  $X_1, \dots, X_N \sim \text{Bernoulli}(p)$  i.i.d.
- Consider the (yes, contrived!) situation in which we only consider two models: (1)  $p = 0.2$  and (2)  $p = 0.8$ , but the true value is  $p = 0.501$ .



## Example: Phylogenetic tree inference for whale species

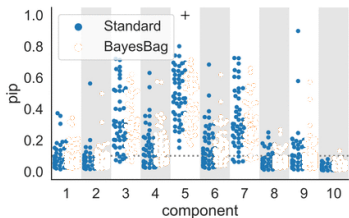
- This is not just a contrived issue – it frequently occurs in practice in phylogenetic inference.
  - ▶ Alfaro et al. (2003), Douady et al. (2003), Wilcox et al. (2002).
- Bayesian phylogenetic inference is very widely used, however, it often yields self-contradictory results due to misspecification.

Overlap between posteriors from two subsets of a whale genetics data set

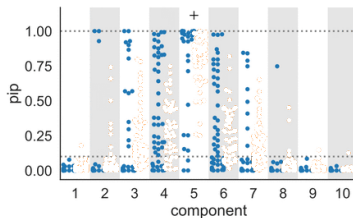


## Example: Variable selection in linear regression

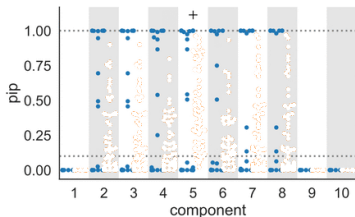
- Similarly, variable selection is unstable when there is misspecification.
- Posterior inclusion probabilities (pips) often flip-flop as  $N$  grows.



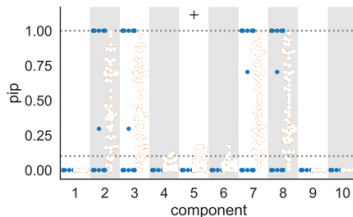
(a)  $N = 5 \times 10^1$



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



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



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

# Outline

- 1 Motivation
- 2 Background
- 3 Methodology (Bagged posteriors)
- 4 Theory
- 5 Applications
  - Variable selection
  - Phylogenetic tree inference
  - Hierarchical mixed effects logistic regression



# Background

- $P_0$  = true distribution of the observed data.
- $\{P_\theta : \theta \in \Theta\}$  is the assumed model.
- Suppose  $P_0$  is not in the assumed model.
- The pseudo-true parameter  $\theta^*$  is the nearest point to  $P_0$  in terms of Kullback–Leibler divergence.
- In this talk, we take the usual perspective that  $\theta^*$  is of interest.
- The posterior concentrates at  $\theta^*$  (under regularity conditions), but ...
  - ▶ It is typically miscalibrated: credible sets do not have correct coverage.
    - ★ Kleijn & van der Vaart (2012)
    - ★ Can recalibrate using sandwich covariance (Müller, 2013, and others)
  - ▶ Slow concentration can occur, causing poor prediction performance.
    - ★ Grünwald & van Ommen (2014)
    - ★ Can fix this using a power posterior  $\propto p(x|\theta)^\zeta p(\theta)$  for certain  $\zeta \in (0, 1)$

# Background

Many methods have been proposed for improving robustness to model misspecification.

- Fitting/prediction, focus on pseudo-true parameter  $\theta^*$ .
  - ▶ Robust adjusted likelihood (Royall & Tsou, 2003)
  - ▶ SafeBayes (Grünwald & van Ommen, 2014)
  - ▶ Modular posteriors (Jacob et al., 2017)
  - ▶ Sandwich covariance adjustment (Müller, 2013)
  - ▶ Holmes & Walker (2017)
  - ... and many others.
- Inference/understanding, focus on ideal parameter  $\theta_I$ .
  - ▶ Coarsened posterior (M. & Dunson, 2019)
  - ▶ Nonparametric perturbation models (M., forthcoming)

# Outline

- 1 Motivation
- 2 Background
- 3 Methodology (Bagged posteriors)
- 4 Theory
- 5 Applications
  - Variable selection
  - Phylogenetic tree inference
  - Hierarchical mixed effects logistic regression

## Bagged posterior (BayesBag)

- Basic idea: Use bagging on the posterior, that is, average the posterior over many bootstrapped datasets.
- More precisely:
  - ▶ Original data set:  $x = (x_1, \dots, x_N)$ .
  - ▶ Bootstrapped copy of original data set:  $x^* = (x_1^*, \dots, x_M^*)$ .
  - ▶ Posterior obtained by treating  $x^*$  as the original data set:

$$\pi(\theta \mid x^*) \propto \pi_0(\theta) \prod_{m=1}^M p_\theta(x_m^*).$$

- ▶ The *bagged posterior* is defined by averaging these posteriors:

$$\pi^*(\theta \mid x) := \frac{1}{NM} \sum_{x^*} \pi(\theta \mid x^*),$$

where the sum is over all  $N^M$  possible bootstrap datasets of  $M$  samples drawn with replacement from the original dataset.

## Bagged posterior (BayesBag): Practical considerations

- In practice, we approximate  $\pi^*(\theta \mid x)$  by generating  $B$  bootstrap datasets  $x_{(1)}^*, \dots, x_{(B)}^*$  and forming the simple Monte Carlo approximation

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

- Any posterior computation technique for the standard posterior can be used to compute each term  $\pi(\theta \mid x_{(b)}^*)$ .
  - ▶ For example, a closed-form solution, MCMC, or quadrature.
- How to choose the number of bootstrap datasets  $B$ ?
  - ▶ As a default,  $B \approx 50$  to 100 often suffices.
  - ▶ Formally, the Monte Carlo error can easily be estimated, since the bootstrap datasets  $x_{(b)}^*$  are i.i.d. given the original dataset.

# Bagged posterior (BayesBag): Practical considerations

- How to choose the bootstrap dataset size  $M$ ?
  - ▶ The choice of  $M$  is connected to calibration of uncertainty.
  - ▶ As  $M/N$  increases, the bagged posterior becomes more concentrated.
- Recommended choice of  $M$  for model selection:
  - ▶ Our theory suggests choosing  $M = o(N)$  or  $M = cN$  with  $c \in (0, 1]$ .
  - ▶ As a default,  $M = N^{0.95}$  works well in theory and practice.
  - ▶ When  $M/N$  is large, the bagged posterior behaves like the standard posterior.
- Recommended choice of  $M$  for parameter inference:
  - ▶ As a default,  $M = N$  is a conservative choice that is robust to misspecification.
  - ▶ If the model is correct, then  $M = 2N$  coincides with the standard posterior, asymptotically.

## Previous work on bagged posteriors (BayesBag)

- Suggested by Waddell et al. (2002) and Douady et al. (2003).
  - ▶ Limited empirical study of BayesBag on phylogenetic inference.
- Independently proposed by Bühlmann (2014).
  - ▶ Limited empirical/theoretical study on a simple univariate Gaussian location model.
  - ▶ Coined the name “BayesBag”, which we adopt here.
- Surprisingly, there seems to have been little empirical or theoretical investigation of bagged posteriors.
- Bagging the posterior is very different than Bayesian Bagging (Clyde & Lee, 2001) and the Bayesian Bootstrap (Rubin, 1981), which are Bayesian ways of doing bagging and bootstrap, respectively.

## Principled justification via Jeffrey conditionalization

- Jeffrey conditionalization (Diaconis & Zabell, 1982; Jeffrey, 1968):
  - ▶ Assume we have a model  $p(x, y)$  for some variables  $x$  and  $y$ .
  - ▶ Suppose we are informed that  $p_0(x)$  is the true distribution of  $x$ .
  - ▶ Then, Jeffrey says to quantify uncertainty in  $y$  using

$$q(y) := \int p(y|x)p_0(x)dx.$$

- Now, to connect this to the bagged posterior:
  - ▶ Take  $x = x_{1:N}$  and  $y = \theta$ .
  - ▶ If we are informed that the true distribution is  $p_0^{(N)}(x_{1:N})$ , then

$$q(\theta) := \int p(\theta | x_{1:N})p_0^{(N)}(x_{1:N})dx_{1:N}.$$

- ▶ Plugging in the empirical distribution  $\frac{1}{N} \sum_{i=1}^N \delta_{x_i}$  for  $p_0$ , we obtain

$$q(\theta) \approx \frac{1}{N^N} \sum_{x_{1:N}^*} p(\theta | x_{1:N}^*),$$

which is precisely the bagged posterior with  $M = N$ .



# Outline

- 1 Motivation
- 2 Background
- 3 Methodology (Bagged posteriors)
- 4 Theory**
- 5 Applications
  - Variable selection
  - Phylogenetic tree inference
  - Hierarchical mixed effects logistic regression

# Overview of theoretical results

- We consider the setting of i.i.d. data  $X_1, \dots, X_N \sim P_0$ .
- **Model selection.** We show that if two models provide a nearly equally good fit to the data distribution  $P_0$ , then:
  - ▶ the standard posterior oscillates randomly, strongly favoring one model or the other at random.
  - ▶ the bagged posterior stabilizes the probabilities probabilities of the two models, improving reproducibility.
- **Parameter inference.** We derive the mean and covariance of the bagged posterior, and prove a Bernstein–von Mises result characterizing the asymptotic normal distribution.

# Theoretical results: Model selection

- Asymptotically, we know the posterior concentrates on the model that is nearest in Kullback–Leibler (KL) divergence to the true distribution.
- To study the non-asymptotic regime via an asymptotic analysis, we consider sequences of models  $\mathbf{m}_{1,N}$  and  $\mathbf{m}_{2,N}$ .
- Letting  $\Lambda_N = \log \frac{p(X_{1:N}|\mathbf{m}_{1,N})}{p(X_{1:N}|\mathbf{m}_{2,N})}$  (the log-likelihood ratio), suppose:
  - ①  $\mathbf{m}_{1,N}$  and  $\mathbf{m}_{2,N}$  are asymptotically comparable in the sense that

$$\lim_{N \rightarrow \infty} E_{P_0}(\Lambda_N / \sqrt{N}) = \mu_\infty \in \mathbb{R},$$

- ②  $\text{Var}_{P_0}(\Lambda_N / \sqrt{N}) = \sigma_\infty^2 \in (0, \infty)$  for all  $N$ , and
  - ③  $M/N \rightarrow c \in [0, \infty)$  as  $N \rightarrow \infty$ , where  $M = M(N) \rightarrow \infty$ .
- The effect size  $\mu_\infty / \sigma_\infty$  is the evidence in favor of model 1.

## Theoretical results: Model selection

- Then as  $N \rightarrow \infty$ , the standard posterior probability of model 1 concentrates at 0 and 1, that is, it converges to a Bernoulli r.v.:

$$\pi(\mathbf{m}_{1,N} \mid X_{1:N}) \xrightarrow{D} \text{Bernoulli}(\Phi(\mu_\infty/\sigma_\infty)).$$

- The bagged posterior probability of model 1 converges to a r.v.:

$$\pi^*(\mathbf{m}_{1,N} \mid X_{1:N}) \xrightarrow{D} \Phi(c^{1/2}Z)$$

where  $Z \sim \mathcal{N}(\mu_\infty/\sigma_\infty, 1)$ .

- In particular, if  $\mu_\infty = 0$  and  $c > 0$ , then

$$\pi(\mathbf{m}_{1,N} \mid X_{1:N}) \xrightarrow{D} \text{Bernoulli}(1/2)$$

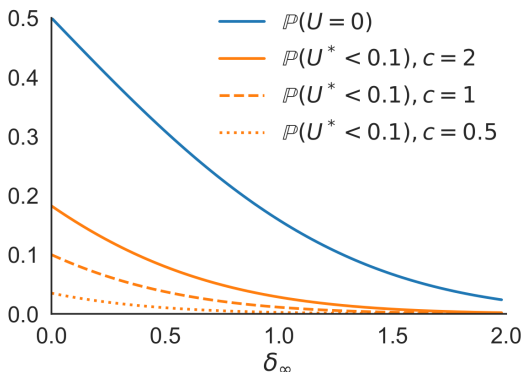
$$\pi^*(\mathbf{m}_{1,N} \mid X_{1:N}) \xrightarrow{D} \text{Uniform}(0, 1).$$

- Meanwhile, if  $c = 0$  then

$$\pi^*(\mathbf{m}_{1,N} \mid X_{1:N}) \xrightarrow{D} 1/2.$$

# Theoretical results: Model selection

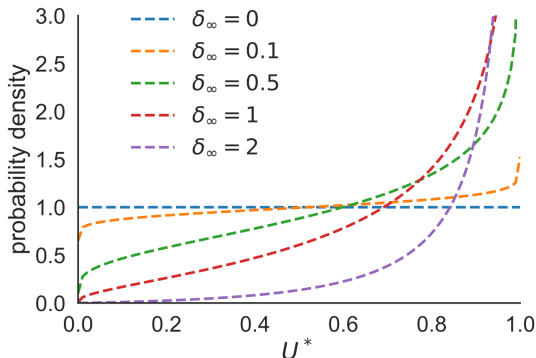
The standard posterior overwhelmingly favors the wrong model with non-negligible probability. The bagged posterior does much better.



- Standard posterior probability of model 1 converges to  $U$ .
- Bagged posterior probability of model 1 converges to  $U^*$ .
- $\delta_\infty := \mu_\infty / \sigma_\infty =$  mean effect size in favor of model 1.

# Theoretical results: Model selection

The bagged posterior converges to a continuous r.v.  $U^*$  on  $[0, 1]$ , avoiding misleading extreme probabilities close to 0 or 1. (Shown:  $c = 1$ .)

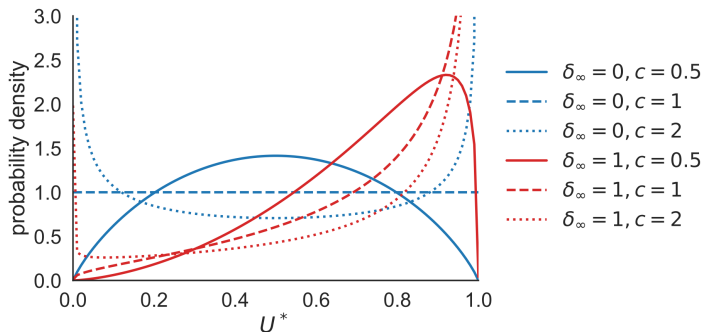


$$U^* = \Phi(c^{1/2}Z) \text{ where } Z \sim \mathcal{N}(\mu_\infty/\sigma_\infty, 1)$$

- $\delta_\infty := \mu_\infty/\sigma_\infty = \text{mean effect size in favor of model 1.}$

# Theoretical results: Model selection

Choosing  $M$  smaller makes the bagged posterior tend to be more uniform over the set of plausible models.



- $c = \lim_{N \rightarrow \infty} M/N$ , where  $M = M(N)$ .
  - ▶ For instance,  $c \in \{0.5, 1, 2\}$  when  $M \in \{0.5N, N, 2N\}$ , respectively.
- $\delta_\infty := \mu_\infty / \sigma_\infty = \text{mean effect size in favor of model 1}$ .

# Outline

- 1 Motivation
- 2 Background
- 3 Methodology (Bagged posteriors)
- 4 Theory
- 5 Applications
  - Variable selection
  - Phylogenetic tree inference
  - Hierarchical mixed effects logistic regression

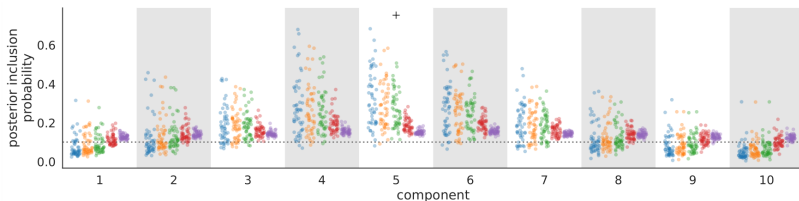


## Application: Variable selection

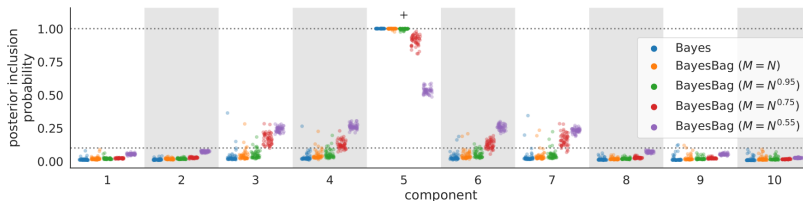
- We consider a standard Bayesian variable selection model for linear regression.
- Specifically, under the prior, each variable is included with probability  $q_0$ , independently, and we integrate out Normal and InverseGamma priors on the coefficients and variance, respectively.
- First, we simulate datasets from (1) the assumed model and (2) a model with nonlinearly transformed covariates.
- In both scenarios, the true coefficient vector is sparse.
- We consider using  $M = N^\alpha$  for  $\alpha \in \{1, 0.95, 0.75, 0.55\}$  to compute the bagged posterior.

## Application: Variable selection

When the model is correct, the bagged posterior with  $M = N^\alpha$  is similar to the standard posterior when  $\alpha = 1$  and more stable as  $\alpha$  decreases.



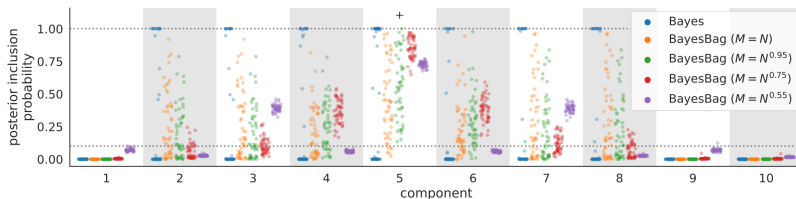
1-sparse-linear,  $N = 5 \times 10^1$



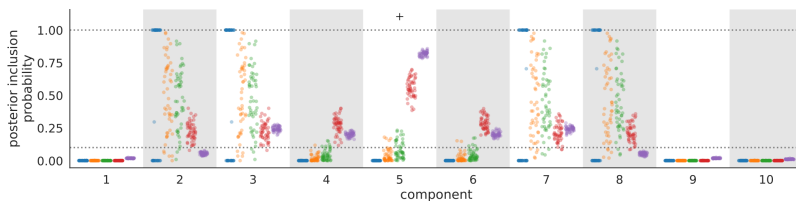
1-sparse-linear,  $N = 5 \times 10^3$

# Application: Variable selection

When the model is incorrect, the bagged posterior avoids the self-contradictory results produced by the standard posterior.



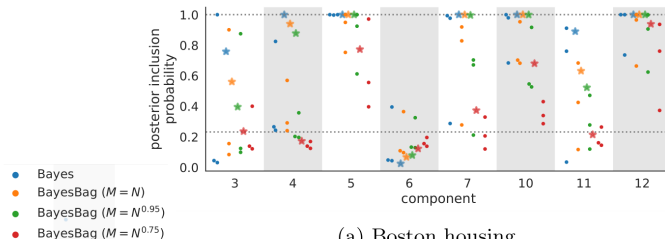
$$N = 5 \times 10^3$$



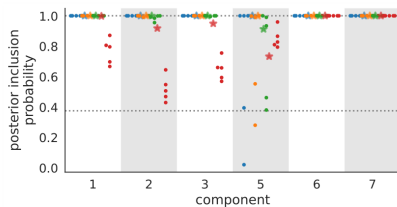
$$N = 5 \times 10^4$$

# Application: Variable selection

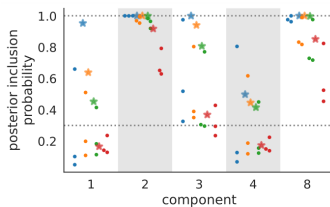
On real datasets, the bagged posterior yields greater reproducibility across subsets of the data.



(a) Boston housing



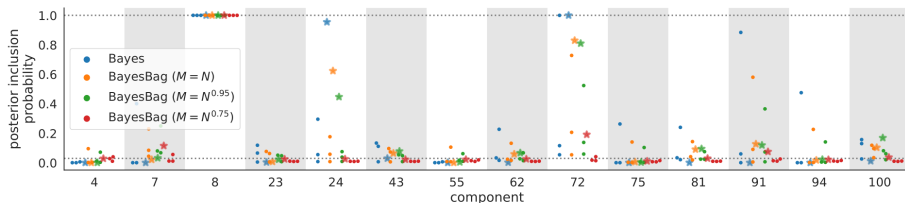
(b) California housing



(c) Diabetes

# Application: Variable selection

On real datasets, the bagged posterior yields greater reproducibility across subsets of the data.

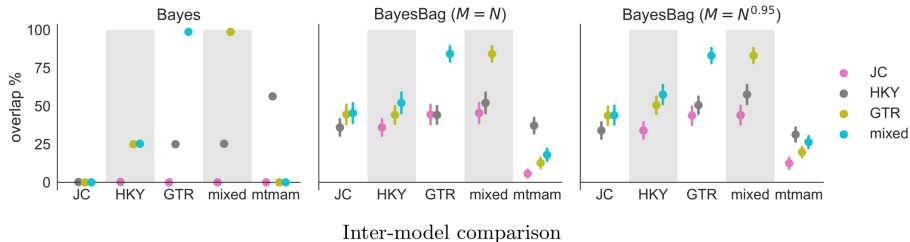


(d) Residential building

## Application: Phylogenetic tree inference

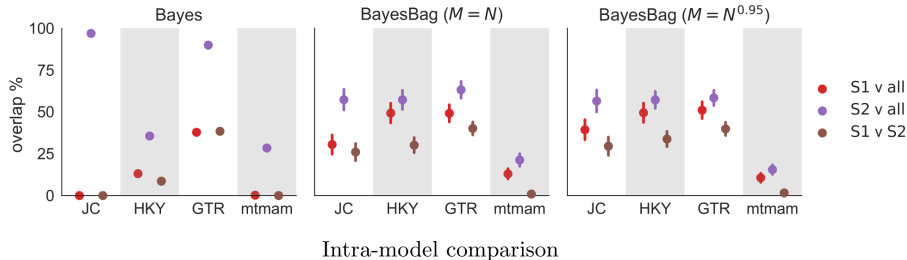
- We use a standard Bayesian package for phylogenetic inference (MrBayes 3.2, Ronquist et al., 2012).
- We used the whale dataset from Yang (2008), consisting of mitochondrial DNA from 13 whale species.
- To compute the posterior on trees, MrBayes was run using five different models for the evolutionary process (JC, HKY, GTR, mixed, and mtmam).
- For the bagged posterior, we used  $M \in \{N, N^{0.95}\}$  and  $B = 100$ .
- To assess reproducibility, we computed the overlap of 99% highest posterior density regions for selected pairs of posteriors.

# Application: Phylogenetic tree inference



- First, we consider the posterior overlap for each pair of evolutionary models.
- The standard posteriors sometimes have extremely low overlap, suggesting poor reproducibility.
- Meanwhile, the bagged posteriors exhibit more reasonable overlaps for each pair.

# Application: Phylogenetic tree inference



- Then, we split the genetic data into two parts, and compute the overlap for (1) the posteriors of the two splits, and (2) the posteriors for each split and the full data.
- Again, the standard posterior exhibits poor reproducibility, while the bagged posterior is more self-consistent.

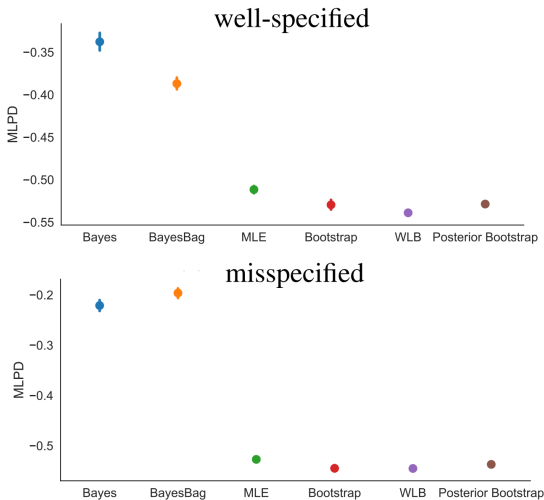


## Application: Hierarchical mixed effects logistic regression

- Finally, we consider a mixed effects model from Browne and Draper (2006), applied to prenatal care data from Guatemalan communities.
- We compare the predictive performance of the standard posterior, the bagged posterior, and four methods based on maximum likelihood estimation (with the random effects integrated out):
  - ▶ the standard MLE,
  - ▶ the bootstrapped MLE,
  - ▶ the weighted likelihood bootstrap (Newton and Raftery, 1994), and
  - ▶ the posterior bootstrap (Lyddon, Walker and Holmes, 2018).

# Application: Hierarchical mixed effects logistic regression

The bagged posterior performs favorably compared to the other methods in terms of mean log predictive density (MLPD).



# Conclusion

- Bagging the posterior is an easy-to-use and widely applicable method that improves upon standard Bayesian inference by making it more stable, accurate, and reproducible.
- Directions for future work or improvements:
  - ▶ Extensions to non-i.i.d. settings such as time-series and spatial data.
  - ▶ Improved computation of bagged posteriors.
  - ▶ Finite-sample theory for bagged posteriors.
  - ▶ Improved model assessment/criticism techniques and theory.

# Robust inference and model selection using bagged posteriors

Jeff Miller

Joint work with Jonathan Huggins

Harvard T.H. Chan School of Public Health  
Department of Biostatistics

University of Washington || Biostatistics Seminar || Jan 13, 2022

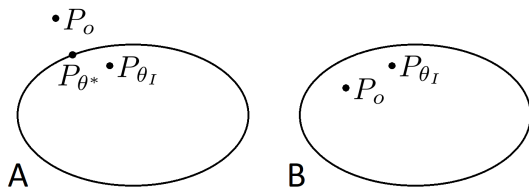
Slides: <http://jwmi.github.io/talks/uw2022.pdf>

Preprint 1: <https://arxiv.org/abs/1912.07104>

Preprint 2: <https://arxiv.org/abs/2007.14845>

## Perspective 2: Model is an idealization of a true process

- Model is interpretable, but not exactly right of course.
- Ideal parameter  $\theta_I$  is of interest.
- Data is from  $P_0$ , which we think of as a perturbation of  $P_{\theta_I}$ .
- The objective is to understand — not to fit.
- This perspective is common in science & medicine.



## Theoretical results: Parameter inference

- Now, consider the bagged posterior on a parameter  $\theta \in \mathbb{R}^D$ .
- Given dataset  $x = x_{1:N}$ , let  $X^*$  be a random bootstrap dataset.
- Let  $\mu(x)$  and  $\Sigma(x)$  denote the mean and covariance matrix of the standard posterior  $p(\theta|x)$ .
- By the law of total expectation, the mean of the bagged posterior is

$$\mathbb{E}(\mu(X^*) \mid x) = \frac{1}{NM} \sum_{x^*} \mu(x^*).$$

- By the law of total variance, the covariance of the bagged posterior is

$$\mathbb{E}(\Sigma(X^*) \mid x) + \text{Cov}(\mu(X^*) \mid x).$$

# Theoretical results: Parameter inference

- Thus, the covariance of the bagged posterior decomposes as the sum of two terms:

①  $E(\Sigma(X^*) \mid x)$

- ★  $\approx$  mean of the posterior covariance matrix under its sampling distribution.
- ★ Bayesian model-based uncertainty averaged with respect to frequentist sampling variability.

②  $\text{Cov}(\mu(X^*) \mid x)$

- ★  $\approx$  covariance of the posterior mean under its sampling distribution.
- ★ Frequentist sampling-based uncertainty of the Bayesian model-based point estimate.

## Theoretical results: Parameter inference

- Suppose  $X_1, \dots, X_N \sim P_0$  i.i.d., and let  $\theta_0$  minimize the KL divergence from  $P_0$ .
- For the standard posterior, by Bernstein–von Mises we know that

$$N^{1/2}(\theta - \hat{\theta}_N) | X_{1:N} \xrightarrow{D} \mathcal{N}(0, J_{\theta_0}^{-1})$$

where  $\theta \sim p(\theta | X_{1:N})$ ,  $\hat{\theta}_N$  is the MLE, and  $J_{\theta_0} = -\mathbb{E}(\nabla^2 \log p_{\theta}(X_i))$ .

- Meanwhile, we also know that the MLE is asymptotically normal:

$$N^{1/2}(\hat{\theta}_N - \theta_0) | X_{1:N} \xrightarrow{D} \mathcal{N}(0, J_{\theta_0}^{-1} I_{\theta_0} J_{\theta_0}^{-1}).$$

where  $I_{\theta_0} = \text{Cov}(\nabla \log p_{\theta}(X_i))$ .

- Hence, asymptotically, the standard posterior is correctly calibrated if these two covariance matrices coincide.



## Theoretical results: Parameter inference

- We prove a Bernstein–von Mises theorem for the bagged posterior, showing that the asymptotic covariance is

$$(J_{\theta_0}^{-1} + J_{\theta_0}^{-1} I_{\theta_0} J_{\theta_0}^{-1})/c$$

where  $c = \lim_{N \rightarrow \infty} M/N$ , and the asymptotic mean is the same as for the standard posterior.

- This is the asymptotic form of the total covariance decomposition.
- When the model is correct,  $c = 2$  recovers the standard posterior, asymptotically, since then  $J_{\theta_0}^{-1} = J_{\theta_0}^{-1} I_{\theta_0} J_{\theta_0}^{-1}$ .
- In general,  $c = 1$  is a safe choice, since it is guaranteed to prevent overconfident credible regions, asymptotically.

## Application: Linear regression

- To illustrate in the parameter inference setting, we consider a standard Bayesian linear regression model.
- As before, we use Normal and InverseGamma priors on the coefficients and variance.
- We simulate data from three scenarios:
  - 1 the assumed model (“default”),
  - 2 the coefficient vector has only one nonzero entry (“1-sparse”), and
  - 3 the covariates are nonlinearly transformed (“nonlinear”).
- For the bagged posterior, we selected  $M$  using an approach based on our asymptotic theory (see Huggins and M., 2019 for details).

# Application: Linear regression

The bagged posterior usually recovers the KL-optimal parameter better in terms of relative squared error (RSE) and log posterior density (LPD).

