Reproducible model selection using bagged posteriors

Jeff Miller

Joint work with Jonathan Huggins

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

EcoSta2022 ‖ June 4, 2022 ‖
Session on "Inference of hierarchical and nonparametric structures"

Slides: http://jwmi.github.io/talks/ecosta2022.pdf
Preprint: https://arxiv.org/abs/2007.14845

Outline

Outline

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 $X_1, \ldots, X_N \sim \text{Bernoulli}(p)$ 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 $p = 0.501$.
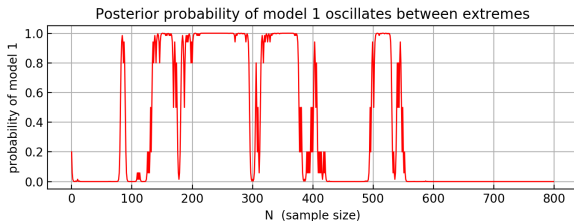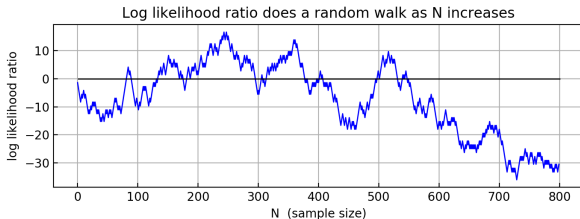
Example: Phylogenetic tree inference for whale species

- 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
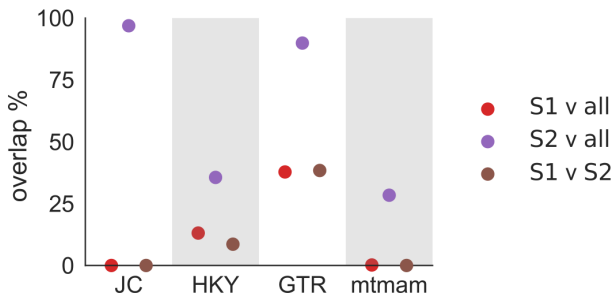
Outline

Bagged posterior (BayesBag)

- Basic idea: Use bagging on the posterior, that is, average the posterior over many bootstrapped datasets.

- More precisely:
  - Original data set: $x = (x_1, \ldots, x_N)$.
  - Bootstrapped copy of original data set: $x^* = (x_1^*, \ldots, x_M^*)$.
  - Set of models under consideration: $\mathfrak{M} = \{\mathfrak{m}_1, \mathfrak{m}_2, \ldots\}$.
  - Posterior obtained by treating x^* as the original data set:

  $$\pi(\mathfrak{m} \mid x^*) \propto \pi(\mathfrak{m}) p(x_{1:M}^* \mid \mathfrak{m}).$$

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

  $$\pi^*(\mathfrak{m} \mid x) := \frac{1}{N^M} \sum_{x^*} \pi(\mathfrak{m} \mid x^*),$$

  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 $\pi^*(\mathfrak{m} \mid x)$ by generating B bootstrap datasets $x^*_{(1)}, \ldots, x^*_{(B)}$ and forming the simple Monte Carlo approximation

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

- Any posterior computation technique for the standard posterior can be used to compute each term $\pi(\mathfrak{m} \mid x^*_{(b)})$.
  - 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?
  - Unlike B, bigger M is not always better.
  - The choice of M affects the concentration of the bagged posterior.
  - Thus, M is connected to calibration of uncertainty.

- Recommended choice of M:
  - 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.
  - Even smaller M may work well under significant misspecification or when there are many models relative to the amount of data.

- The role of M is subtly different in the model selection setting compared to the parameter inference setting.

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.

- In concurrent work, we (Huggins & Miller, 2019) we have investigated bagged posteriors in the parameter inference setting.

- 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.

Outline

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 $\mathfrak{m}_{1,N}$ and $\mathfrak{m}_{2,N}$.

- Letting $\Lambda_N = \log \dfrac{p(X_{1:N}|\mathfrak{m}_{1,N})}{p(X_{1:N}|\mathfrak{m}_{2,N})}$ (the log-likelihood ratio), suppose:

  ① $\mathfrak{m}_{1,N}$ and $\mathfrak{m}_{2,N}$ are asymptotically comparable in the sense that

$$\lim_{N \to \infty} \mathrm{E}_{P_0}(\Lambda_N/\sqrt{N}) = \mu_\infty \in \mathbb{R},$$

  ② $\mathrm{Var}_{P_0}(\Lambda_N/\sqrt{N}) = \sigma_\infty^2 \in (0, \infty)$ for all N, and

  ③ $M/N \to c \in [0, \infty)$ as $N \to \infty$, where $M = M(N) \to \infty$.

- The effect size μ_∞/σ_∞ is the evidence in favor of model 1.

Theoretical results: Model selection

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

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

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

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

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

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

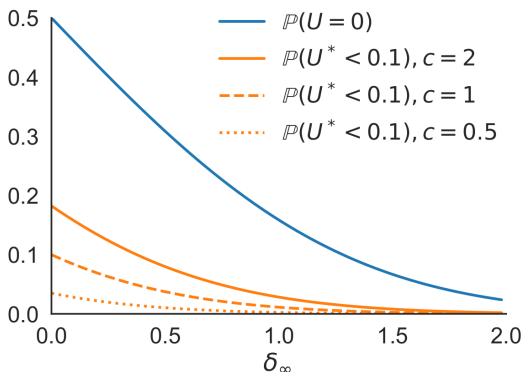$$\pi(\mathfrak{m}_{1,N} \mid X_{1:N}) \xrightarrow{D} \mathrm{Bernoulli}(1/2)$$

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

- Meanwhile, if $c = 0$ then

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

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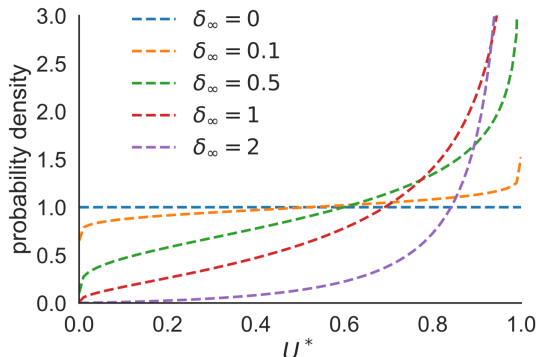
$$U^* = \Phi(c^{1/2}Z) \text{ where } Z \sim \mathcal{N}(\mu_\infty/\sigma_\infty, 1)$$

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

Outline

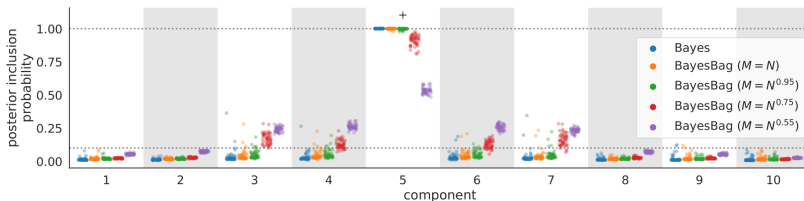# Application: Variable selection

- 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 α decreases.
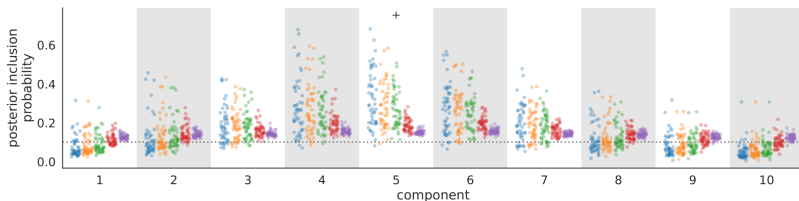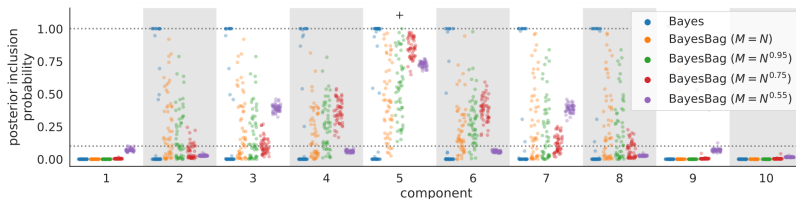
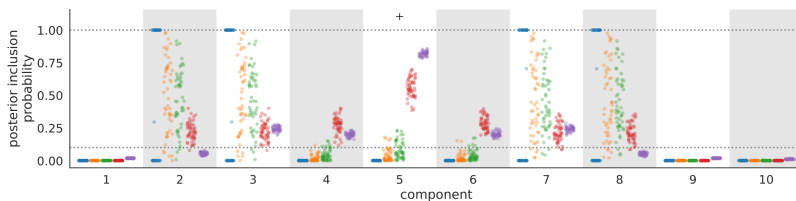

1-sparse-linear, $N = 5 \times 10^1$

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

Application: Variable selection

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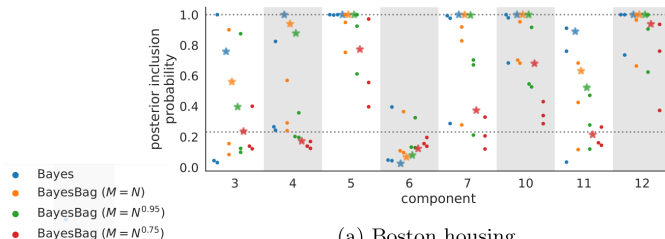self-contradictory results produced by the standard posterior.
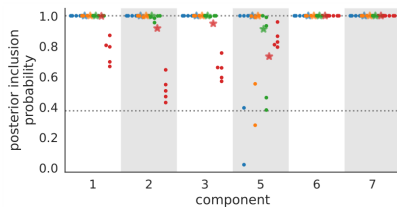

$$N = 5 \times 10^3$$

$$N = 5 \times 10^4$$
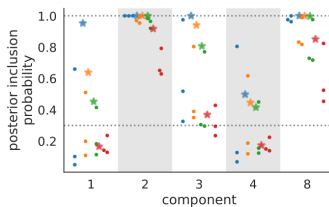
Application: Variable selection

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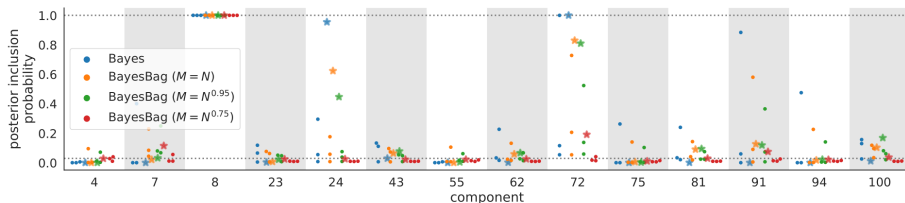

(a) Boston housing

(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

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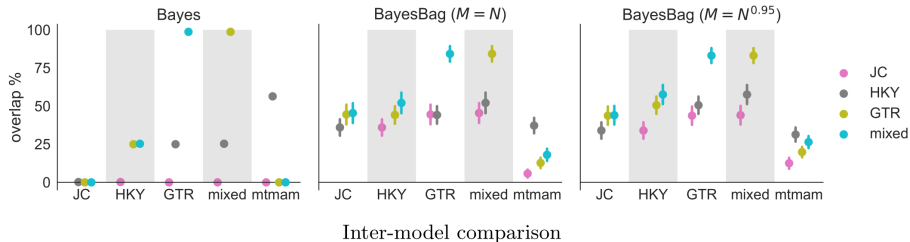
- First, we consider the posterior overlap for each pair of evolutionary models.
- The standard posteriors sometimes have extremely low overlap, suggesting poor reproducibility.
- Meanwhile, the bagged posteriors exhibit more reasonable overlaps for each pair.

Application: Phylogenetic tree inference

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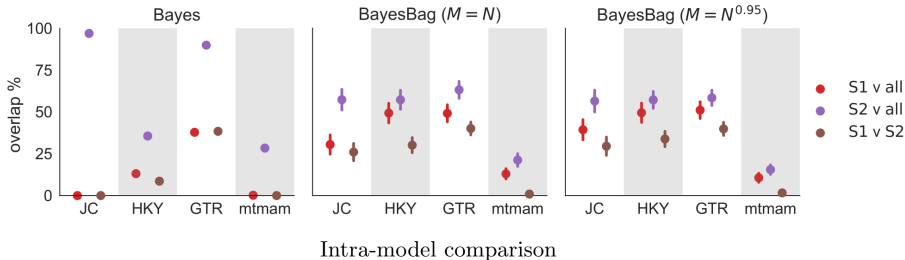
- Then, we split the genetic data into two parts, and compute the overlap for (1) the posteriors of the two splits, and (2) the posteriors for each split and the full data.
- Again, the standard posterior exhibits poor reproducibility, while the bagged posterior is more self-consistent.

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.

Reproducible model selection using bagged posteriors

Jeff Miller

Joint work with Jonathan Huggins

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

EcoSta2022 ‖ June 4, 2022 ‖
Session on "Inference of hierarchical and nonparametric structures"

Slides: http://jwmi.github.io/talks/ecosta2022.pdf
Preprint: https://arxiv.org/abs/2007.14845