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

Harvard Statistics Colloquium || Cambridge, MA || Feb 1, 2021

Slides: http://jwmi.github.io/talks/Harvard2021.pdf
Preprint 1: https://arxiv.org/abs/1912.07104
Preprint 2: https://arxiv.org/abs/2007.14845

Outline

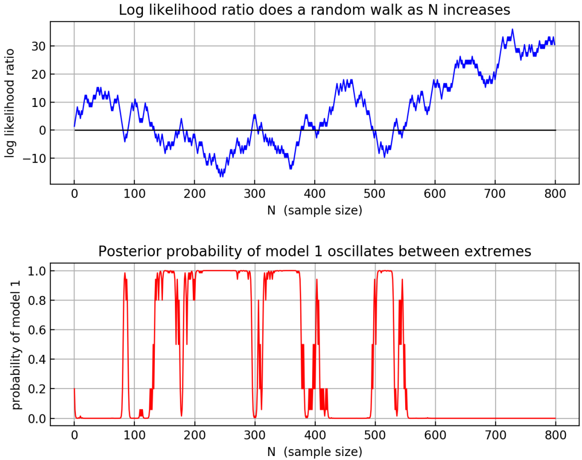
Outline

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

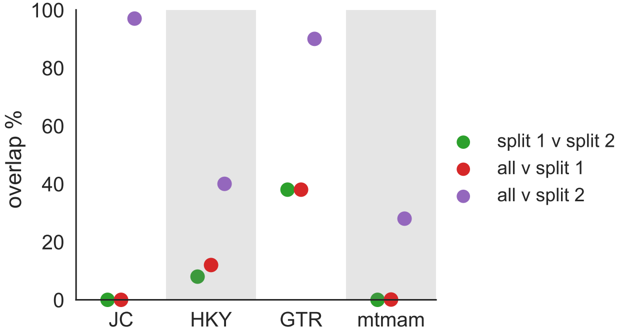
- Suppose $X_1, \ldots, X_N \sim \text{Bernoulli}(p)$ i.i.d.
- Consider the (yes, contrived!) situation in which we only consider two models: (1) $p = 0.2$ and (2) $p = 0.8$, but the true value is $p = 0.501$.



Example: Phylogenetic tree inference for whale species

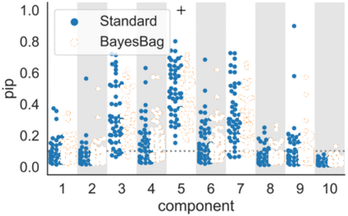
- This is not just a contrived issue – it frequently occurs in practice in phylogenetic inference.
  - Alfaro et al. (2003), Douady et al. (2003), Wilcox et al. (2002).
- Bayesian phylogenetic inference is very widely used, however, it often yields self-contradictory results due to misspecification.

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

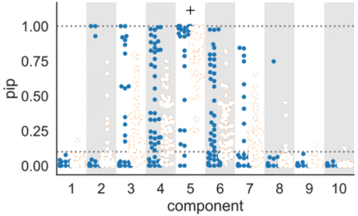


Example: Variable selection in linear regression

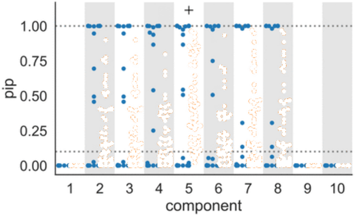
- Similarly, variable selection is unstable when there is misspecification.
- Posterior inclusion probabilities (pips) often flip-flop as N grows.



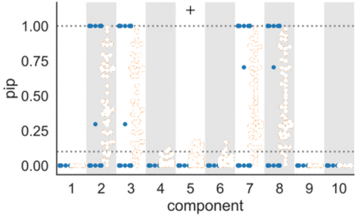
(a) $N = 5 \times 10^1$



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



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

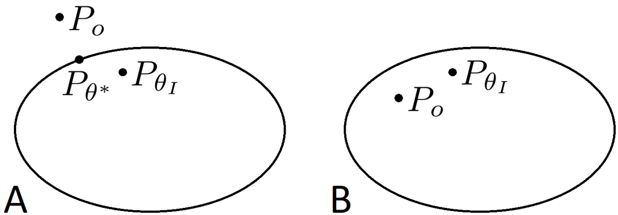


(d) $N = 5 \times 10^4$

Outline

What do we mean by misspecification? Two scenarios

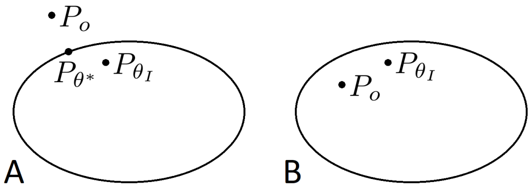
- Notation:
  - P_o = distribution of the observed data
  - θ^* = pseudo-true parameter (KL-nearest point in model to P_o)
  - θ_I = ideal parameter (the truth before perturbation)
  - We think of P_o as a perturbation of P_{θ_I}.
- Scenario A: P_o is not in the model class.
- Scenario B: P_o is in the model class, but $P_o \neq P_{\theta_I}$.



- If there is no perturbation, then $P_o = P_{\theta^*} = P_{\theta_I}$.

What is the quantity of interest?

- The choice of method depends on the quantity of interest.
- Two main perspectives:
  1. *Fitting/prediction:* Model is a tool for approximating P_o.
     - ★ Want to predict future observations.
     - ★ Pseudo-true parameter θ^* is of interest.
  2. *Inference:* Model is an idealization of a true process.
     - ★ Want to recover unknown true parameters.
     - ★ Ideal parameter θ_I is of interest.



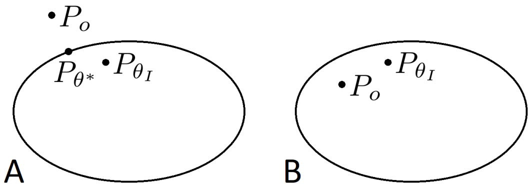
Perspective 1: Model is a tool for approximating P_o

- Pseudo-true parameter θ^* is of interest.
- Common when doing prediction using classification or regression.

- The posterior concentrates at θ^* (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 $\zeta \in (0, 1)$

Perspective 2: Model is an idealization of a true process

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



Some previous methods, categorized

- Perspective 1: Fitting/prediction, focus on pseudo-true parameter θ^*.
  - 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.
- Perspective 2: Inference, focus on idealized parameter θ_I.
  - Coarsened posterior (M. & Dunson, 2019)
  - Nonparametric perturbation models (M., forthcoming)
- In this talk, we focus on perspective 1.

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: $x^* = (x_1^*, \ldots, x_M^*)$.
  - 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:

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

    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 $x^*_{(1)}, \ldots, x^*_{(B)}$ 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?
  - 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.

- Interpretation of M:
  - 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.
  - As M/N increases, the bagged posterior becomes more concentrated.

- 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.
- 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

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

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

$$q(\theta) \approx \frac{1}{N^N} \sum_{x^*_{1:N}} p(\theta \mid x^*_{1:N}),$$

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

Outline

Overview of theoretical results

- We consider the setting of i.i.d. data $X_1, \ldots, X_N \sim P_0$.
- **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 $\mathfrak{m}_{1,N}$ 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
  $$\lim_{N\to\infty} \mathrm{E}(\Lambda_N/\sqrt{N}) = \mu_\infty \in \mathbb{R},$$
  2. $\mathrm{Var}(\Lambda_N/\sqrt{N}) = \sigma_\infty^2 \in (0,\infty)$ 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.:

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

- When $c > 0$, the bagged posterior probability of model 1 converges to a continuous r.v. with pdf

$$f(u) = \Phi'\big(c^{-1/2}\Phi^{-1}(u) - \mu_\infty/\sigma_\infty\big)c^{-1/2}/\Phi'(\Phi^{-1}(u)).$$

- When $c = 0$, the bagged posterior prob. of model 1 converges to 1/2:

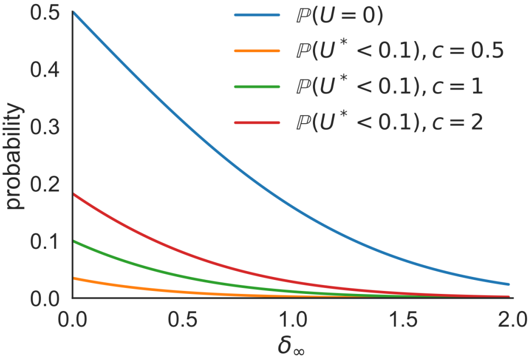
$$\pi^*(\mathfrak{m}_{1,N} \mid X_{1:N}) \xrightarrow{P} 1/2.$$

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

$$\pi(\mathfrak{m}_{1,N} \mid X_{1:N}) \xrightarrow{D} \mathrm{Bernoulli}(1/2)$$
$$\pi^*(\mathfrak{m}_{1,N} \mid X_{1:N}) \xrightarrow{D} \mathrm{Uniform}(0,1).$$

Theoretical results: Model selection

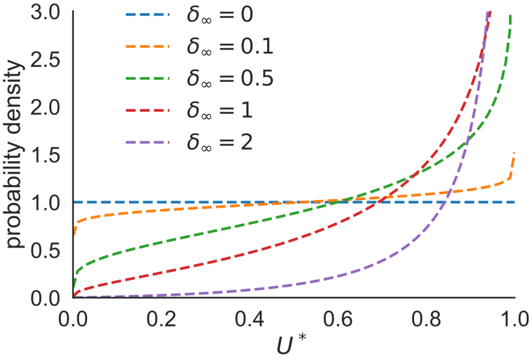
The standard posterior overwhelmingly favors the wrong model with non-negligible probability. The bagged posterior does much better.



- 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: $c = 1$.)

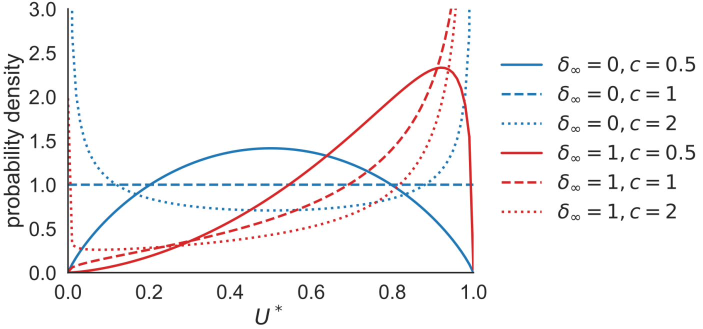


$$f_{U^*}(u) = \Phi'\big(c^{-1/2}\Phi^{-1}(u) - \mu_\infty/\sigma_\infty\big)c^{-1/2}/\Phi'(\Phi^{-1}(u))$$

- $\delta_\infty := \mu_\infty/\sigma_\infty$ = 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.



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

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

$$\mathrm{E}(\mu(X^*) \mid x) = \frac{1}{N^M} \sum_{x^*} \mu(x^*).$$

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

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

Theoretical results: Parameter inference

- Thus, the covariance of the bagged posterior decomposes as the sum of two terms:
  1. $\mathrm{E}(\Sigma(X^*) \mid x)$
     - ★ $\approx$ mean of the posterior covariance matrix under its sampling distribution.
     - ★ Bayesian model-based uncertainty averaged with respect to frequentist sampling variability.
  2. $\mathrm{Cov}(\mu(X^*) \mid x)$
     - ★ $\approx$ 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 $X_1, \ldots, X_N \sim P_0$ i.i.d., and let θ_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 $J_{\theta_0} = -\mathrm{E}(\nabla^2 \log p_\theta(X_i))$.

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

- 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 $c = \lim_{N\to\infty} M/N$, 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.

Outline

Application: Variable selection

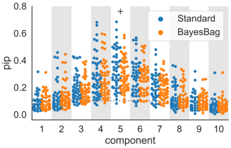
- We consider a standard Bayesian variable selection model for linear regression.
- Specifically, under the prior, each variable is included with probability q_0, independently, and we integrate out Normal and InverseGamma priors on the coefficients and variance, respectively.

- 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.

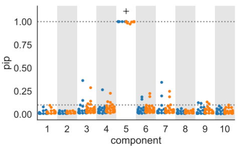
- We compute the bagged posterior using $M = N$.

Application: Variable selection

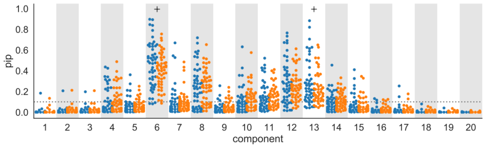
When the model is correct, the bagged and standard posteriors are similar.



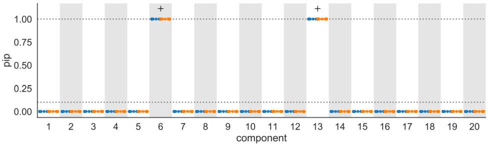
(a) 1-sparse-linear, $N = 5 \times 10^1$



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



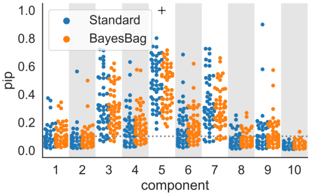
(c) 2-sparse-linear, $N = 10^2$



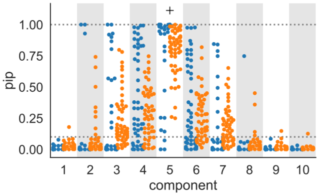
(d) 2-sparse-linear, $N = 10^4$

Application: Variable selection

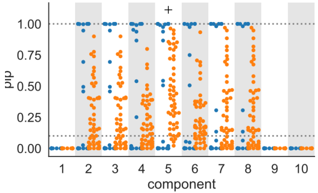
When the model is incorrect, the bagged posterior avoids the self-contradictory results produced by the standard posterior.



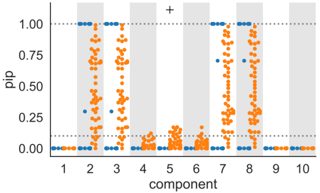
(a) $N = 5 \times 10^1$



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



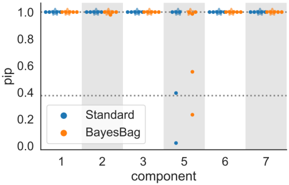
(c) $N = 5 \times 10^3$



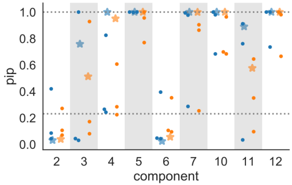
(d) $N = 5 \times 10^4$

Application: Variable selection

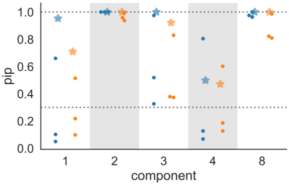
On real datasets, the difference is not dramatic, but the bagged posterior does yield greater reproducibility across subsets of the data.



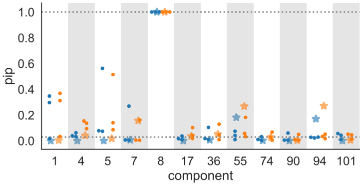
(a) California housing ($\mathcal{I} = 1.00$)



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



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

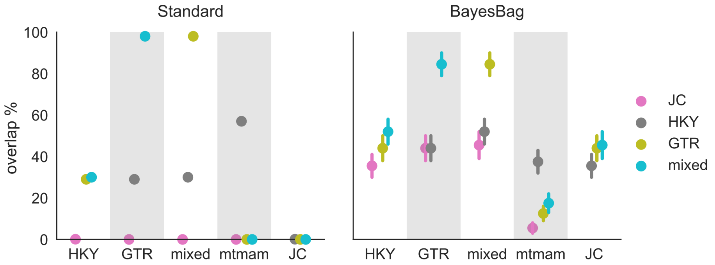


(d) Residential building ($\lambda = 16, \mathcal{I} =$ NA)

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 $M = N$ and $B = 100$.
- To assess reproducibility, we computed the overlap of 99% highest posterior density regions for selected pairs of posteriors.

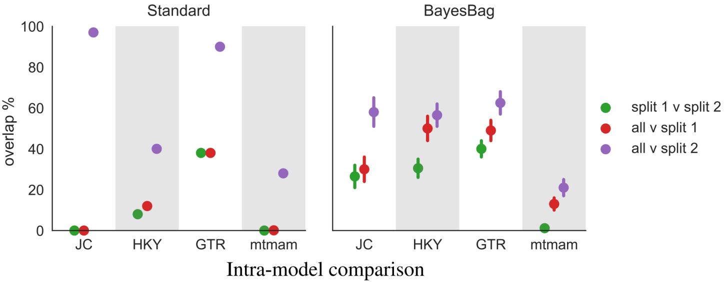
Application: Phylogenetic tree inference



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



- 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: Linear regression

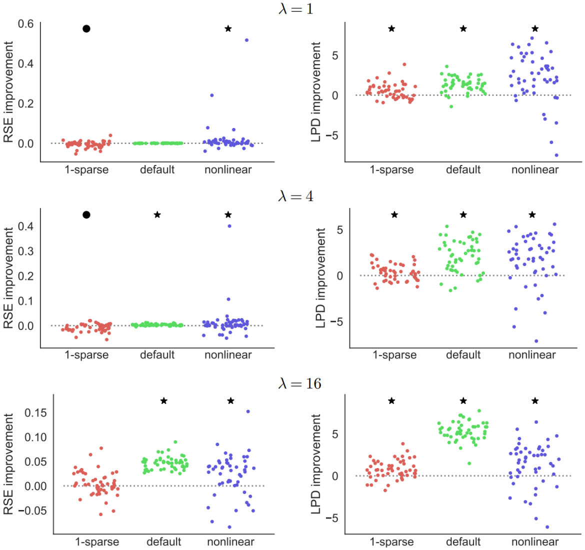
- To illustrate in the parameter inference setting, we consider a standard Bayesian linear regression model.
- As before, we use Normal and InverseGamma priors on the coefficients and variance.

- 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).

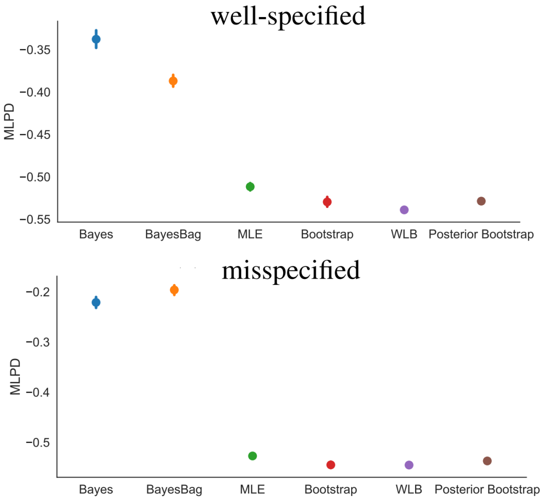


Application: Hierarchical mixed effects logistic regression

- 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 (e.g., Pierre Jacob proposed an unbiased MCMC approach).
  - 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

Harvard Statistics Colloquium || Cambridge, MA || Feb 1, 2021

Slides: http://jwmi.github.io/talks/Harvard2021.pdf
Preprint 1: https://arxiv.org/abs/1912.07104
Preprint 2: https://arxiv.org/abs/2007.14845