Scalable Nonparametric Sampling from Multimodal Posteriors with the Posterior Bootstrap

Edwin Fong[1,2] Simon Lyddon[1] Chris Holmes[1,2]

[1]University of Oxford, [2]The Alan Turing Institute

Thirty-sixth International Conference on Machine Learning

Table of Contents

Challenges in Bayesian Inference

Suppose we observe $y_{1:n} \overset{\text{iid}}{\sim} F_0$. We are interested in a parameter $\theta \in \Theta \subseteq \mathbb{R}^p$, which indexes a family of probability densities $\mathcal{F}_\Theta = \{f_\theta(y); \theta \in \Theta\}$.

Challenges in Bayesian Inference

Suppose we observe $y_{1:n} \stackrel{\text{iid}}{\sim} F_0$. We are interested in a parameter $\theta \in \Theta \subseteq \mathbb{R}^p$, which indexes a family of probability densities $\mathcal{F}_\Theta = \{f_\theta(y); \theta \in \Theta\}$.

Model misspecification

- Bayesian inference assumes that $f_0 \in \mathcal{F}_\Theta$
- Unlikely in large and complex datasets

Challenges in Bayesian Inference

Suppose we observe $y_{1:n} \overset{\text{iid}}{\sim} F_0$. We are interested in a parameter $\theta \in \Theta \subseteq \mathbb{R}^p$, which indexes a family of probability densities $\mathcal{F}_\Theta = \{f_\theta(y); \theta \in \Theta\}$.

Model misspecification

- Bayesian inference assumes that $f_0 \in \mathcal{F}_\Theta$
- Unlikely in large and complex datasets

Computation

- Markov chain Monte Carlo is inherently serial, computationally expensive, and struggles with multimodal posteriors
- Difficult to quantify the approximation of Variational Bayes, and poor uncertainty estimates

Bayesian Nonparametric Learning

We present a scalable Bayesian nonparametric learning (NPL) routine with the following properties:

Bayesian Nonparametric Learning

We present a scalable Bayesian nonparametric learning (NPL) routine with the following properties:

- A Dirichlet process (DP) prior on the unknown data distribution accounts for **model misspecification**.

Bayesian Nonparametric Learning

We present a scalable Bayesian nonparametric learning (NPL) routine with the following properties:

- A Dirichlet process (DP) prior on the unknown data distribution accounts for **model misspecification**.
- We sample from the NPL posterior through **parallel optimizations** of randomized objective functions.

Bayesian Nonparametric Learning

We present a scalable Bayesian nonparametric learning (NPL) routine with the following properties:

- A Dirichlet process (DP) prior on the unknown data distribution accounts for **model misspecification**.
- We sample from the NPL posterior through **parallel optimizations** of randomized objective functions.
- Our method is adept at sampling from **multimodal** posterior distributions via a random restart mechanism.

Table of Contents

Bayesian Nonparametric Learning [Lyddon et al., 2018]

Suppose we observe $y_{1:n} \stackrel{\text{iid}}{\sim} F_0$.

Our parameter of interest is defined:

$$\theta_0(F_0) = \arg\min_{\theta} \int \ell(y, \theta) dF_0(y) \tag{1}$$

Bayesian Nonparametric Learning [Lyddon et al., 2018]

Suppose we observe $y_{1:n} \overset{\text{iid}}{\sim} F_0$.

Our parameter of interest is defined:

$$\theta_0(F_0) = \arg\min_{\theta} \int \ell(y, \theta) dF_0(y) \tag{1}$$

- For example, $\ell(y, \theta) = |y - \theta|$ gives the median and $(y - \theta)^2$ gives the mean.

Bayesian Nonparametric Learning [Lyddon et al., 2018]

Suppose we observe $y_{1:n} \stackrel{\text{iid}}{\sim} F_0$.

Our parameter of interest is defined:

$$\theta_0(F_0) = \arg\min_{\theta} \int \ell(y, \theta) dF_0(y) \tag{1}$$

- For example, $\ell(y, \theta) = |y - \theta|$ gives the median and $(y - \theta)^2$ gives the mean.
- For model fitting, let $\ell(y, \theta) = -\log f_\theta(y)$, where f_θ is the density of some parametric model.

Our NPL Posterior

We elicit a Dirichlet process prior on the unknown sampling distribution:

$$F \sim \mathrm{DP}\left(\alpha, F_\pi\right) \tag{2}$$

Our NPL Posterior

We elicit a Dirichlet process prior on the unknown sampling distribution:

$$F \sim \mathrm{DP}\left(\alpha, F_\pi\right) \tag{2}$$

Calculate the posterior over F from the conjugacy of the DP:

$$\begin{aligned} [F|y_{1:n}] &\sim \mathrm{DP}\left(\alpha + n, G_n\right) \\ G_n &= \frac{\alpha}{\alpha + n} F_\pi + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{y_i} \end{aligned} \tag{3}$$

Our NPL Posterior

We elicit a Dirichlet process prior on the unknown sampling distribution:

$$F \sim \mathrm{DP}\left(\alpha, F_\pi\right) \tag{2}$$

Calculate the posterior over F from the conjugacy of the DP:

$$\begin{aligned} [F|y_{1:n}] &\sim \mathrm{DP}\left(\alpha + n, G_n\right) \\ G_n &= \frac{\alpha}{\alpha + n} F_\pi + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{y_i} \end{aligned} \tag{3}$$

Then the NPL posterior over θ is defined:

$$\tilde{\pi}(\theta|y_{1:n}) = \int \pi(\theta|F) d\pi(F|y_{1:n}) \tag{4}$$

where $\pi(\theta|F) = \delta_{\theta_0(F)}(\theta)$; the delta arises as θ is a deterministic functional of F as in (1).

Sampling from the NPL Posterior

Algorithm 1 NPL Posterior Sampling

for $i = 1$ **to** B **do**
  Draw $F^{(i)} \sim \pi(F|y_{1:n})$
  $\theta^{(i)} = \arg\min_\theta \int \ell(y,\theta) dF^{(i)}(y)$
end for

Here $\theta^{(i)} \sim \tilde{\pi}(\theta|y_{1:n})$ and B is the number of posterior samples.

Sampling from the NPL Posterior

Algorithm 1 NPL Posterior Sampling

for $i = 1$ **to** B **do**
  Draw $F^{(i)} \sim \pi(F|y_{1:n})$
  $\theta^{(i)} = \arg\min_\theta \int \ell(y, \theta) dF^{(i)}(y)$
end for

Here $\theta^{(i)} \sim \tilde{\pi}(\theta|y_{1:n})$ and B is the number of posterior samples.

- NPL posterior is usually intractable

Sampling from the NPL Posterior

Algorithm 1 NPL Posterior Sampling

for $i = 1$ **to** B **do**
  Draw $F^{(i)} \sim \pi(F|y_{1:n})$
  $\theta^{(i)} = \arg\min_\theta \int \ell(y, \theta) dF^{(i)}(y)$
end for

Here $\theta^{(i)} \sim \tilde{\pi}(\theta|y_{1:n})$ and B is the number of posterior samples.

- NPL posterior is usually intractable
- Embarrassingly parallel sampling scheme

Properties of NPL Posterior

Theoretical properties follow from properties of the DP

Properties of NPL Posterior

Theoretical properties follow from properties of the DP

Consistency at θ_0, from the properties of the DP. This is true irrespective of the choice of F_π.

Properties of NPL Posterior

Theoretical properties follow from properties of the DP

Consistency at θ_0, from the properties of the DP. This is true irrespective of the choice of F_π.

Asymptotic dominance of $\tilde{\pi}(\cdot|y_{1:n})$ over $\pi(\cdot|y_{1:n})$ for $\alpha = 0$:

$$\begin{aligned}\mathbb{E}_{y_{1:n}\sim q}\left[\mathrm{KL}(q(\cdot)||\pi(\cdot|y_{1:n})) - \mathrm{KL}(q(\cdot)||\tilde{\pi}(\cdot|y_{1:n}))\right] \\ = K(q(\cdot)) + o(n^{-1})\end{aligned}$$

for all distributions q, where K is a non-negative and possibly positive real-valued functional.

The Posterior Bootstrap

Draws of F from the posterior DP are almost surely discrete:

$$\begin{aligned}\theta(F) &= \arg\min_{\theta} \int \ell(y,\theta)dF(y) \\ &= \arg\min_{\theta} \sum_{k=1}^{\infty} w_k \ell(\tilde{y}_k,\theta)\end{aligned} \tag{5}$$

where $w_{1:\infty} \sim \text{GEM}(\alpha + n)$ and $\tilde{y}_{1:\infty} \overset{\text{iid}}{\sim} G_n$ from the stick-breaking construction.

The Posterior Bootstrap

Draws of F from the posterior DP are almost surely discrete:

$$\theta(F) = \arg\min_{\theta} \int \ell(y, \theta) dF(y) = \arg\min_{\theta} \sum_{k=1}^{\infty} w_k \ell(\tilde{y}_k, \theta) \tag{5}$$

where $w_{1:\infty} \sim \text{GEM}(\alpha + n)$ and $\tilde{y}_{1:\infty} \overset{\text{iid}}{\sim} G_n$ from the stick-breaking construction.

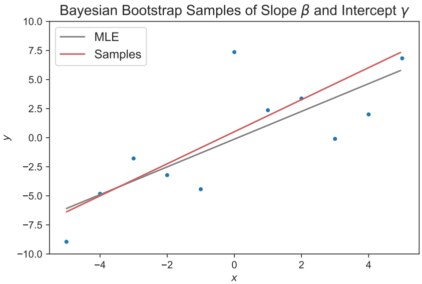
As an approximation, we can truncate the sum to obtain the **posterior bootstrap**.

The Posterior Bootstrap

Algorithm 2 Posterior Bootstrap Sampling

Define T as truncation limit
Observed samples are $y_{1:n}$
for $i = 1$ **to** B **do**
  Draw prior pseudo-samples $\tilde{y}_{1:T}^{(i)} \overset{\text{iid}}{\sim} F_\pi$
  Draw $(w_{1:n}^{(i)}, \tilde{w}_{1:T}^{(i)}) \sim \text{Dir}\,(1, \ldots, 1, \alpha/T, \ldots, \alpha/T)$
  $$\theta^{(i)} = \arg\min_\theta \left\{ \sum_{j=1}^{n} w_j^{(i)} \ell(y_j, \theta) + \sum_{k=1}^{T} \tilde{w}_k^{(i)} \ell(\tilde{y}_k^{(i)}, \theta) \right\}$$
end for

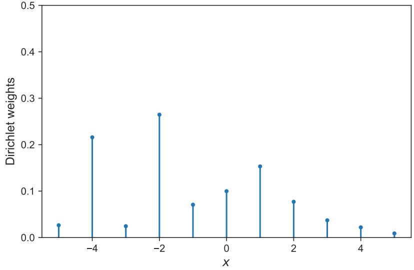
The Posterior Bootstrap for a Linear Model

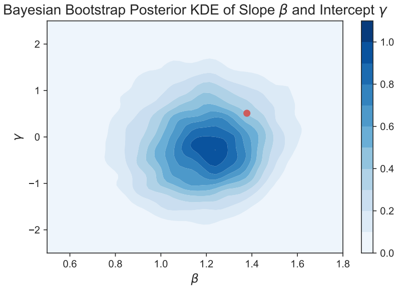


For a simple linear model

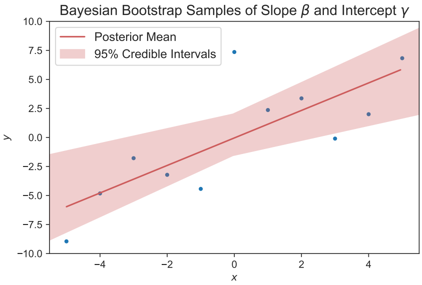
$$f_\beta(y|x) = \mathcal{N}(y; \beta x + \gamma, 1)$$

sample $\left(\beta^{(i)}, \gamma^{(i)}\right) \sim \tilde{\pi}(\beta, \gamma | y)$ with $\alpha = 0$. Here $n = 11$ and $B = 10000$.





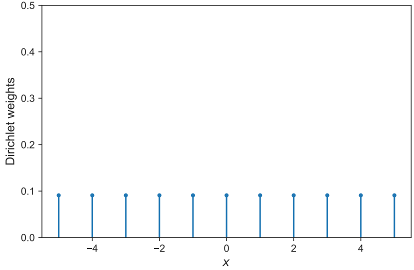
The Posterior Bootstrap for a Linear Model

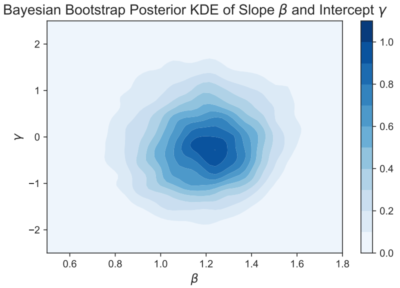


For a simple linear model

$$f_\beta(y|x) = \mathcal{N}(y; \beta x + \gamma, 1)$$

sample $\left(\beta^{(i)}, \gamma^{(i)}\right) \sim \tilde{\pi}(\beta, \gamma | y)$ with $\alpha = 0$. Here $n = 11$ and $B = 10000$.





Multimodality

If $\ell(y, \theta)$ is non-convex, then the NPL posterior may be multimodal.

Multimodality

If $\ell(y, \theta)$ is non-convex, then the NPL posterior may be multimodal.

We approximate the global minimization with random restart with R local minimizations.

Multimodality

If $\ell(y, \theta)$ is non-convex, then the NPL posterior may be multimodal.

We approximate the global minimization with random restart with R local minimizations.

Algorithm 3 Random Restart NPL Posterior Sampling

for $i = 1$ **to** B **do**
  Draw $F^{(i)} \sim \text{DP}(\alpha + n, G_n)$
  for $r = 1$ **to** R **do**
    Draw $\theta_r^{\text{init}} \sim \pi_0$
    $\theta_r^{(i)} = \text{local}\arg\min_\theta \left(\int \ell(y, \theta) dF^{(i)}(y), \theta_r^{\text{init}}\right)$
  end for
  $\theta^{(i)} = \arg\min_r \int \ell(y, \theta_r^{(i)}) dF^{(i)}(y)$
end for

Related Approaches

In [Lyddon et al., 2018], they let $\pi(F)$ be a mixture of Dirichlet processes:

$$F|\theta \sim \mathrm{DP}\left(\alpha, F_\theta\right); \quad \theta \sim \pi(\theta) \tag{6}$$

where $(f_\theta, \pi(\theta))$ is the conventional Bayesian likelihood and prior.

Related Approaches

In [Lyddon et al., 2018], they let $\pi(F)$ be a mixture of Dirichlet processes:

$$F|\theta \sim \mathrm{DP}\left(\alpha, F_\theta\right); \quad \theta \sim \pi(\theta) \tag{6}$$

where $(f_\theta, \pi(\theta))$ is the conventional Bayesian likelihood and prior.

- They recover conventional Bayesian inference for $\alpha \to \infty$

Related Approaches

In [Lyddon et al., 2018], they let $\pi(F)$ be a mixture of Dirichlet processes:

$$F|\theta \sim \mathrm{DP}\left(\alpha, F_\theta\right); \quad \theta \sim \pi(\theta) \tag{6}$$

where $(f_\theta, \pi(\theta))$ is the conventional Bayesian likelihood and prior.

- They recover conventional Bayesian inference for $\alpha \to \infty$
- Posterior $\pi(F|y_{1:n})$ requires sampling from Bayesian posterior $\pi(\theta|y_{1:n})$, which is the computationally difficult step

Related Approaches

- Bayesian bootstrap [Rubin, 1981] for $\alpha = 0$

Related Approaches

- Bayesian bootstrap [Rubin, 1981] for $\alpha = 0$
- Weighted likelihood bootstrap [Newton and Raftery, 1994] if we further set $\ell(y, \theta) = -\log f_\theta(y)$

Related Approaches

- Bayesian bootstrap [Rubin, 1981] for $\alpha = 0$
- Weighted likelihood bootstrap [Newton and Raftery, 1994] if we further set $\ell(y, \theta) = -\log f_\theta(y)$
- General Bayesian updating [Bissiri et al., 2016] also uses the expected loss to define a posterior

Table of Contents

Gaussian Mixture Model

Our Bayesian model for K-component diagonal GMM with non-conjugate prior is:

$$\begin{aligned}
\mathbf{y}_i | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma} &\sim \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{\mu}_k, \mathrm{diag}(\boldsymbol{\sigma}_k^2)\right) \\
\boldsymbol{\pi} | a_0 &\sim \mathrm{Dir}(a_0, \ldots, a_0) \\
\mu_{kj} &\sim \mathcal{N}(0, 1) \\
\sigma_{kj} &\sim \mathrm{logNormal}(0, 1)
\end{aligned} \tag{7}$$

Gaussian Mixture Model

Our Bayesian model for K-component diagonal GMM with non-conjugate prior is:

$$\begin{aligned} \mathbf{y}_i | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma} &\sim \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{\mu}_k, \text{diag}(\boldsymbol{\sigma}_k^2)\right) \\ \boldsymbol{\pi} | a_0 &\sim \text{Dir}(a_0, \ldots, a_0) \\ \mu_{kj} &\sim \mathcal{N}(0, 1) \\ \sigma_{kj} &\sim \text{logNormal}(0, 1) \end{aligned} \quad (7)$$

For NPL, we are interested in model fitting, so our loss function is simply the negative log-likelihood:

$$\ell(\mathbf{y}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}) = -\log \sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{y}; \boldsymbol{\mu}_k, \text{diag}(\boldsymbol{\sigma}_k^2)\right) \quad (8)$$

Gaussian Mixture Model: Toy Data

Toy data from a GMM with $K = 3$, $d = 1$ and the parameters:

$$\boldsymbol{\pi}_0 = \{0.1, 0.3, 0.6\}, \;\; \boldsymbol{\mu}_0 = \{0, 2, 4\}, \;\; \boldsymbol{\sigma}_0^2 = \{1, 1, 1\} \tag{9}$$

$$n_{train} = 1000, n_{test} = 250$$

Gaussian Mixture Model: Toy Data

Toy data from a GMM with $K = 3$, $d = 1$ and the parameters:

$$\boldsymbol{\pi}_0 = \{0.1, 0.3, 0.6\}, \ \ \boldsymbol{\mu}_0 = \{0, 2, 4\}, \ \ \boldsymbol{\sigma}_0^2 = \{1, 1, 1\} \tag{9}$$

$$n_{train} = 1000, n_{test} = 250$$

As $n >> d$, we elicit a noninformative NPL prior with $\alpha = 0$

Gaussian Mixture Model: Toy Data

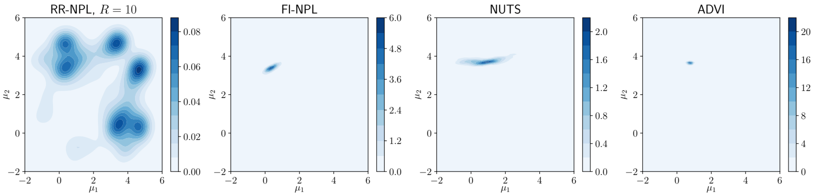


Figure 1: Posterior KDE of (μ_1, μ_2) in K=3 toy GMM problem

Gaussian Mixture Model: Toy Data

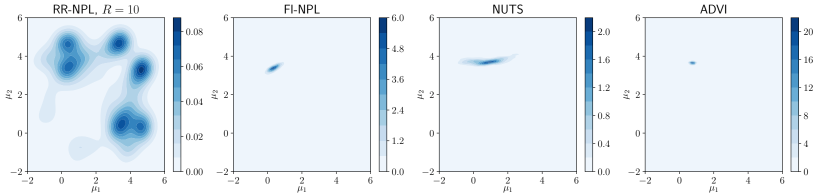


Figure 1: Posterior KDE of (μ_1, μ_2) in K=3 toy GMM problem

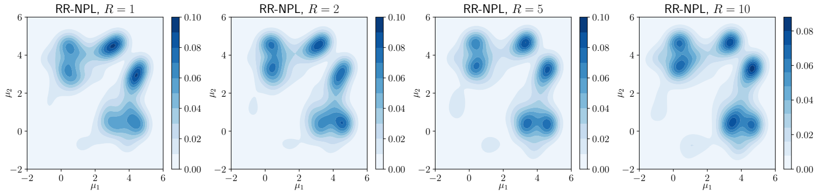


Figure 2: Posterior KDE of (μ_1, μ_2) in K=3 toy GMM problem for RR-NPL with increasing R

Sparse Logistic Regression

Our Bayesian model for sparse logistic is:

$$\begin{aligned} y_i|\mathbf{x}_i, \boldsymbol{\beta}, \beta_0 &\sim \text{Bernoulli}(\eta_i) \\ \eta_i &= \sigma(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) \\ \beta_j &\sim \text{Student-t}\left(2a, 0, \frac{b}{a}\right) \end{aligned} \tag{10}$$

Sparse Logistic Regression

Our Bayesian model for sparse logistic is:

$$\begin{aligned} y_i|\mathbf{x}_i, \boldsymbol{\beta}, \beta_0 &\sim \text{Bernoulli}(\eta_i) \\ \eta_i &= \sigma(\boldsymbol{\beta}^T \mathbf{x}_i + \beta_0) \\ \beta_j &\sim \text{Student-t}\left(2a, 0, \frac{b}{a}\right) \end{aligned} \tag{10}$$

For NPL, we use the loss:

$$\begin{aligned} \ell(y, \mathbf{x}, \boldsymbol{\beta}, \beta_0) = &- (y \log \eta + (1-y)\log(1-\eta)) \\ &+ \gamma\left(\frac{2a+1}{2}\right)\sum_{j=1}^{d} \log\left(1 + \frac{\beta_j^2}{2b}\right) \end{aligned} \tag{11}$$

Sparse Logistic Regression: UCI Datasets

We use 3 binary classification datasets from UCI ML repo: ‘Adult’ ($n = 36177, d = 96$), ‘Polish companies bankruptcy’ ($n = 8402, d = 64$), and ‘Arcene’ ($n = 100, d = 10000$)

Sparse Logistic Regression: UCI Datasets

We use 3 binary classification datasets from UCI ML repo: 'Adult' ($n = 36177, d = 96$), 'Polish companies bankruptcy' ($n = 8402, d = 64$), and 'Arcene' ($n = 100, d = 10000$)

Table 1: Mean log pointwise predictive density on held-out test data for LogReg

Data Set	Loss-NPL	NUTS	ADVI
Adult	**-0.326**	**-0.326**	-0.327
Polish	**-0.229**	-3.336	-0.247
Arcene	**-0.449**	-0.464	**-0.445**

Sparse Logistic Regression: UCI Datasets

We use 3 binary classification datasets from UCI ML repo: 'Adult' ($n = 36177, d = 96$), 'Polish companies bankruptcy' ($n = 8402, d = 64$), and 'Arcene' ($n = 100, d = 10000$)

Table 1: Mean log pointwise predictive density on held-out test data for LogReg

Data Set	Loss-NPL	NUTS	ADVI
Adult	**-0.326**	**-0.326**	-0.327
Polish	**-0.229**	-3.336	-0.247
Arcene	**-0.449**	-0.464	**-0.445**

Table 2: Run-time for 2000 samples for LogReg on 4 72-core Azure VMs

Data Set	Loss-NPL	NUTS	ADVI
Adult	2m24s	2h36m	26.9s
Polish	19.0s	1h20m	3.3s
Arcene	2m20s	4h31m	54.2s

Bayesian Sparsity-path-analysis: Genetics Dataset

Single-neucleotide polymorphisms from a genome-wide data set [Lee et al., 2012] with $n = 500$, $d = 50$

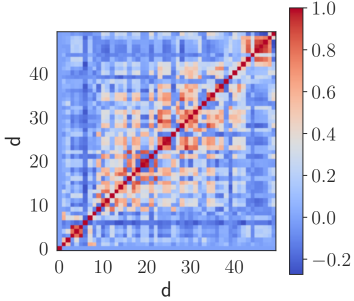


Figure 3: Block-like correlations of covariates $\boldsymbol{x}$

Bayesian Sparsity-path-analysis: Genetics Dataset

Single-neucleotide polymorphisms from a genome-wide data set [Lee et al., 2012] with $n = 500$, $d = 50$

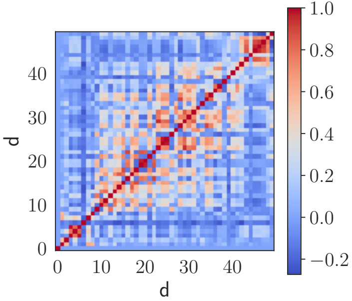


Figure 3: Block-like correlations of covariates $\boldsymbol{x}$

We simulated phenotype data from $y \sim \text{Bernoulli}(\sigma(\beta_0^T \mathbf{x}))$; β_0 has 5 non-zero components with the rest set to 0.

Bayesian Sparsity-path-analysis: Genetics Dataset

We vary the scale of the Student-t prior $c = b/a$ (same ℓ as before) to visualize how the responsibility of each covariate changes with sparsity

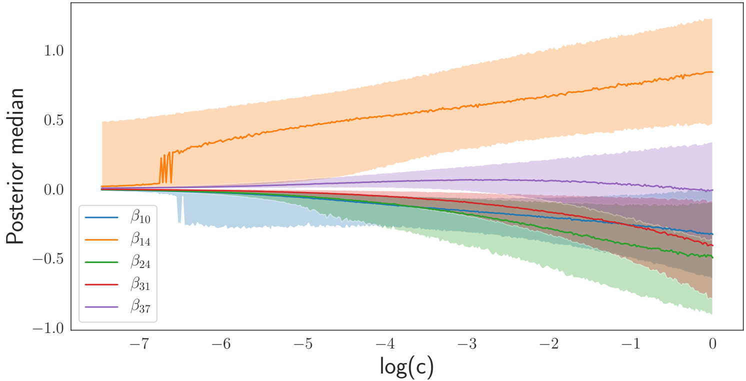


Figure 4: Lasso-type plot for posterior medians of non-zero β with 80% credible intervals against $\log(c)$ from genetic dataset. NPL required 5m 24s to generate 450×4000 posterior samples.

Bayesian Sparsity-path-analysis: Genetics Dataset

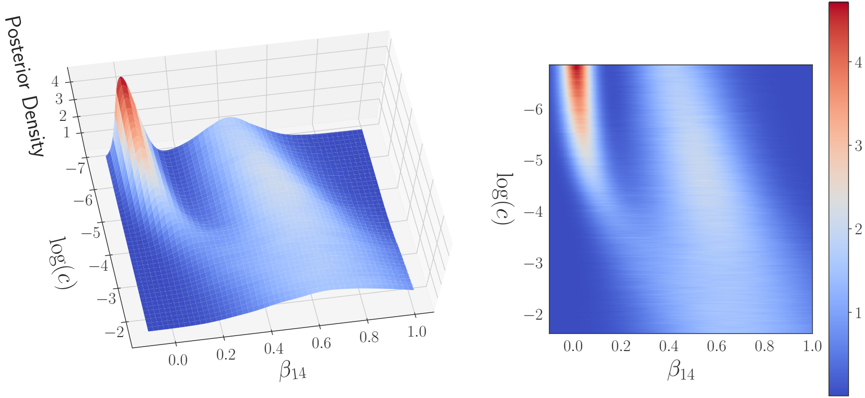


Figure 5: Posterior marginal KDE of β_{14} against $\log(c)$ from genetic dataset

Summary

We have introduced a scalable Bayesian nonparametric learning scheme which:

Summary

We have introduced a scalable Bayesian nonparametric learning scheme which:

- Takes into account **model misspecification**

Summary

We have introduced a scalable Bayesian nonparametric learning scheme which:

- Takes into account **model misspecification**
- Allows for an **embarrassingly parallel** sampling scheme through optimizations

Summary

We have introduced a scalable Bayesian nonparametric learning scheme which:

- Takes into account **model misspecification**
- Allows for an **embarrassingly parallel** sampling scheme through optimizations
- Can tackle **multimodal** posteriors

Summary

We have introduced a scalable Bayesian nonparametric learning scheme which:

- Takes into account **model misspecification**
- Allows for an **embarrassingly parallel** sampling scheme through optimizations
- Can tackle **multimodal** posteriors

Thank you! Any questions?

Come check out poster #235.

References

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(5):1103–1130.

Lee, A., Caron, F., Doucet, A., and Holmes, C. (2012).
Bayesian sparsity-path-analysis of genetic association signal using generalized t priors.
Statistical applications in genetics and molecular biology, 11 2.

Lyddon, S., Walker, S., and Holmes, C. C. (2018).
Nonparametric learning from Bayesian models with randomized objective functions.
In *Advances in Neural Information Processing Systems 31*, pages 2075–2085. Curran Associates, Inc.

Newton, M. and Raftery, A. (1994).
Approximate bayesian inference by the weighted likelihood bootstrap.
Journal of the Royal Statistical Society Series B-Methodological, 56:3 – 48.

Rubin, D. B. (1981).
The Bayesian bootstrap.
The Annals of Statistics, 9(1):130–134.