A Method for Bayesian Inference in Massive Datasets

Greg Ridgeway
Department of Statistics
University of Washington

with
David Madigan and Thomas Richardson

Outline

- Bayesian Inference
  - Example: Logistic regression
- Monte Carlo methods
  - Importance sampling
  - Effective sample size
- Massive datasets
  - Posterior simulation
  - Predictive weight trimming

Bayesian Inference

- Model for the data, $y_i \sim f(y \mid \theta)$
- Prior knowledge, $f(\theta)$
- Bayes' Theorem

$$f(\theta \mid y) = \frac{f(y \mid \theta) f(\theta)}{\int f(y \mid \theta) f(\theta) d\theta} \propto f(y \mid \theta) f(\theta)$$

Example 1: Beta-Bernoulli

- We observe $y=\{0,1,1,1,1,0,0,1,0,1\}$
- A model

$$y_i = \begin{cases} 1 & \text{with probability } \theta \\ 0 & \text{with probability } 1-\theta \end{cases}$$

$$f(y_i \mid \theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

- A prior

$$f(\theta) = I(0 < \theta < 1)$$

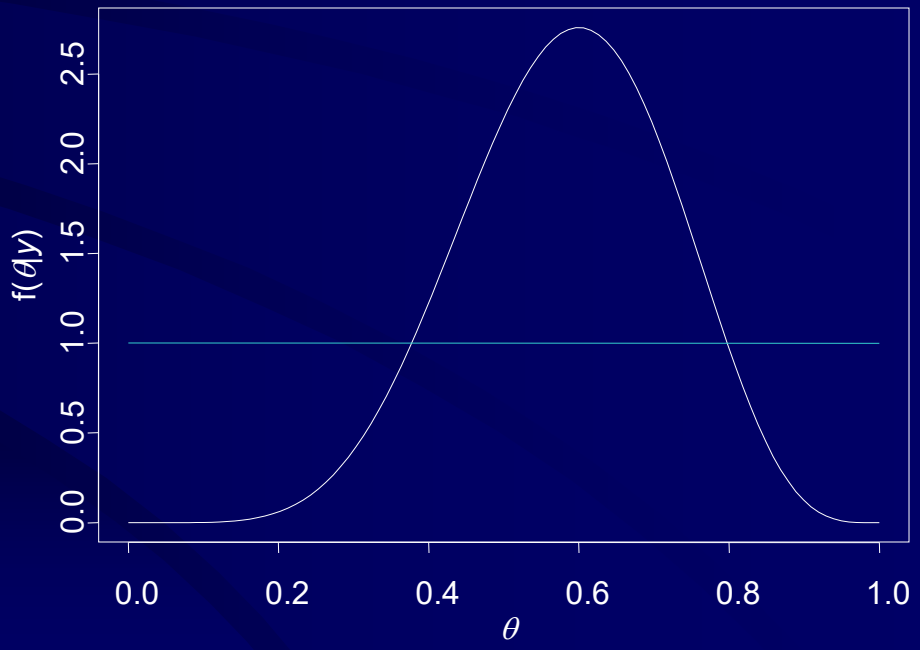
Beta-Bernoulli Posterior

- The posterior

$$f(\theta \mid y) = \frac{\theta^{\sum y_i}(1-\theta)^{n-\sum y_i}}{\int \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}\, d\theta} \propto \theta^{6}(1-\theta)^{4}$$

$$\equiv \text{Beta}(7,5)$$

- Summarize with
  - density
  - mean
  - variance



Example 2: Logistic regression

- We might have covariates with each y_i.

$$\{\underline{x}_i, y_i\}, y_i \in \{0,1\}$$

- Let the probability depend on $\underline{x}$.

$$f(y_i \mid p(x_i)) = p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

- Linear logistic regression

$$\log \frac{p(x_i)}{1 - p(x_i)} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots$$

Logistic regression posterior

The posterior distribution for the parameters of a logistic regression model is complex.

$$f(\underline{\beta} \mid \underline{x}, \underline{y}) \propto f(\underline{\beta}) \prod_{i=1}^{N} \left(1 + e^{-(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij})} \right)^{-y_i} \left(1 + e^{\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij}} \right)^{y_i - 1}$$

Mean and variance of $f(\beta | \underline{x}, \underline{y})$?

Monte Carlo methods

- Summaries from Bayesian data analyses

$$E(h(\theta) \mid x_1, x_2, \ldots, x_n) = \int h(\theta) f(\theta \mid x_1, x_2, \ldots, x_n) d\theta$$

- Monte Carlo integration

$$f(\theta \mid x_1, x_2, \ldots, x_n)$$

$$\hat{E}(h(\theta) \mid x) = \frac{1}{M} \sum_{i=1}^{M} h(\theta_i)$$

Sampling

- Many distributions are easy to sample from; beta, multinomial, multivariate normal, …

- Adaptive rejection sampling
- Markov Chain Monte Carlo
- Importance sampling

Importance sampling

- $f(\theta)$ - target density
- $g(\theta)$ - sampling density

$$\int \theta \cdot f(\theta) d\theta = \int \theta \frac{f(\theta)}{g(\theta)} g(\theta) d\theta$$

$$\approx \frac{1}{m} \sum_{i=1}^{m} \theta_i w_i$$

where $w_i = \frac{f(\theta_i)}{g(\theta_i)}$ and $\theta_i \sim g(\theta)$

Accuracy of importance sampling

- Accuracy depends on the variance of the weights

$$\mathrm{Var}_g\left(\frac{1}{m}\sum_{i=1}^{m}\theta_i w_i\right) \approx (1+\mathrm{Var}_g(w)) \times \mathrm{Var}_f\left(\frac{1}{m}\sum_{i=1}^{m}\theta_i\right)$$

$$\text{where } w(\theta) = \frac{f(\theta)}{g(\theta)}$$

Effective sample size

- A rule-of-thumb for *ESS*

$$ESS = m \frac{1}{1 + \mathrm{Var}_g (w / \mathrm{E}_g\, w)}$$

- Drawing m times from $g(\theta)$ is like drawing *ESS* times from $f(\theta)$

Massive Datasets

$$\int h(\theta) f(\theta \mid x_1, x_2, \ldots, x_n) d\theta$$

- The dimension of θ does not affect the convergence rate, but it does slow down each iteration.
- The sample size, n, can also substantially slow down the iterations, especially if it exceeds the computer's main memory.

Example: Logistic regression

The posterior distribution for the parameters of a logistic regression model is complex.

$$f(\underline{\beta} \mid \underline{x}, \underline{y}) \propto f(\underline{\beta}) \prod_{i=1}^{N} \left(1 + e^{-(\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij})}\right)^{-y_i} \left(1 + e^{\beta_0 + \sum_{j=1}^{d} \beta_j x_{ij}}\right)^{y_i - 1}$$

An evaluation for any β requires a scan of the dataset.

Patching the posterior

G. Ridgeway, D. Madigan, T. Richardson [in progress]

$$\int h(\theta) f(\theta \mid x_1, x_2, \ldots, x_n) d\theta = \int h(\theta) f(\theta \mid g_1, g_2) d\theta$$

$$= \int h(\theta) \frac{f(g_1 \mid \theta) f(g_2 \mid \theta) f(\theta)}{f(g_1, g_2)} d\theta$$

$$= \frac{\int h(\theta) f(g_2 \mid \theta) f(\theta \mid g_1) d\theta}{\int f(g_2 \mid \theta) f(\theta \mid g_1) d\theta}$$

$$= \frac{E_{\theta \mid g_1} h(\theta) f(g_2 \mid \theta)}{E_{\theta \mid g_1} f(g_2 \mid \theta)}$$

Patching the posterior

$$\int h(\theta) f(\theta \mid x_1, x_2, \ldots, x_n) d\theta = \frac{E_{\theta|g_1} h(\theta) f(g_2 \mid \theta)}{E_{\theta|g_1} f(g_2 \mid \theta)}$$

$$= \lim_{m \to \infty} \frac{\frac{1}{m} \sum h(\theta_i) w_i}{\frac{1}{m} \sum w_i}$$

where $w_i = f(g_2 \mid \theta_i)$ and $\theta_i \sim f(\theta \mid g_1)$

Efficiency

Efficient if $f(\theta \mid g_1)$ is close to $f(\theta \mid g_1, g_2)$

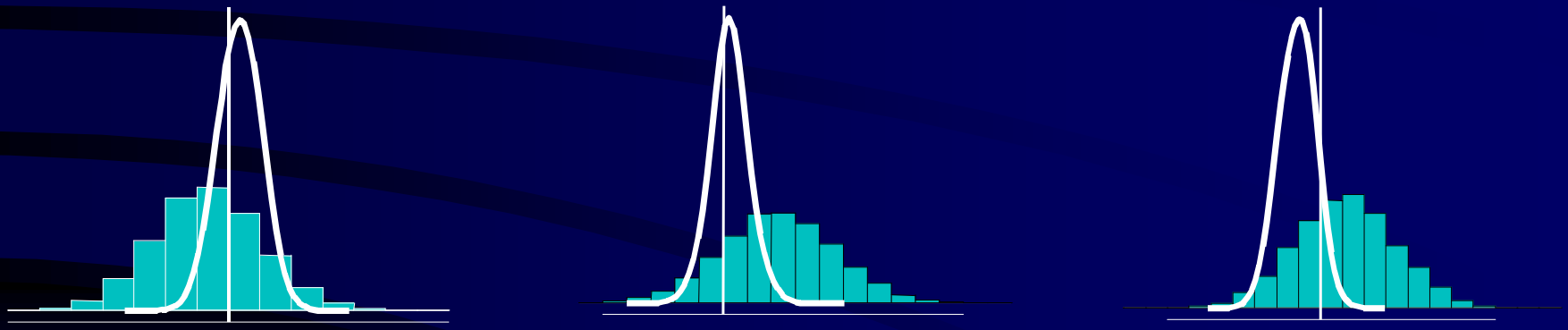
- Reasonably large g_1 implies that the locations will be close.
- Variance of $f(\theta \mid g_1, g_2)$ decreases as $|g_2|$ increases (on average).

$$\mathrm{Var}(\theta \mid g_1) = \mathrm{E}(\mathrm{Var}(\theta \mid g_1, g_2)) + \mathrm{Var}(\mathrm{E}(\theta \mid g_1, g_2))$$

$$\mathrm{Var}(\theta \mid g_1) \geq \mathrm{E}(\mathrm{Var}(\theta \mid g_1, g_2))$$

Example: Logistic regression

$|g_1| = 10{,}000$ and $|g_2| = 10{,}000$



Weight variance

Variance of the importance sampling weights dictate the rate of convergence.

Proposition Assume that

$$x_i \sim N(\mu, \sigma^2),\ \mu \sim N(\mu_0, \tau_0^2)$$

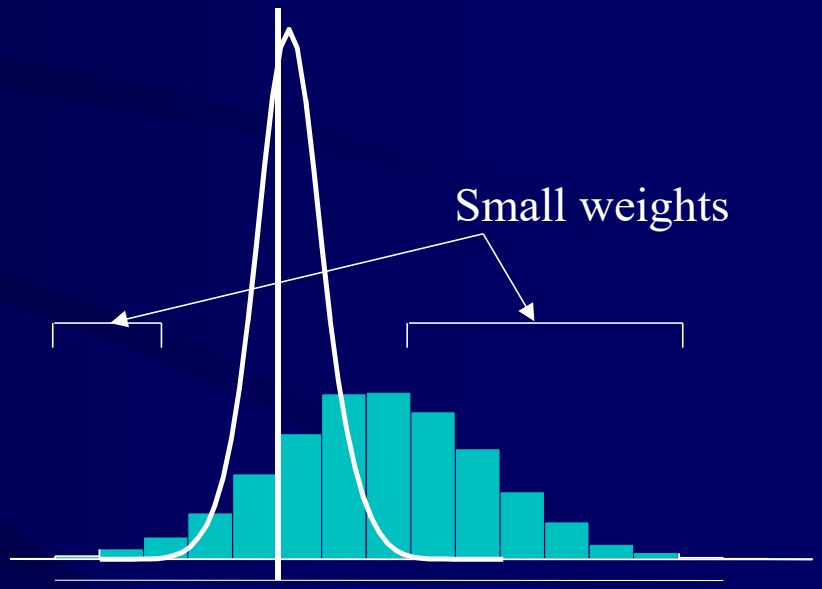
$$g_1 = \{x_1, \ldots, x_{n_1}\},\ g_2 = \{x_{n_1+1}, \ldots, x_{n_1+n_2}\}$$

then

$$\lim_{\tau_0 \to \infty} E_{g_1} E_{g_2} \operatorname{Var}\left(\frac{f(\mu \mid g_1, g_2)}{f(\mu \mid g_1)} \right) = \frac{n_2}{n_1}$$

Adaptive importance sampling

- Small *ESS* is due to wasted draws.
- As n_2 grows the size of these regions will increase.
- We can learn the weightless regions.



Adaptive importance sampling

- θ_i and w_i are deterministically linked.

$$w(\theta) = f(g_2|\theta)$$

1. Computing $w(\theta)$ is expensive.
2. θ's for which $w(\theta)$ is virtually 0 are useless.
3. Avoid computing $w(\theta)$ for values for "weightless" θ's.

Predictive weight trimming

- Draw $\theta_1,\ldots,\theta_m$ from $f(\theta \mid g_1)$.
- Compute $w_i = f(g_2|\theta_i)$ where $n_2 = n_1$.
- Construct a weight threshold predictor, $\mathrm{I}(\theta)$

$$1(w > \varepsilon) \approx \mathrm{I}(\theta)$$

- Repeat sampling from a truncated form

$$f(\theta \mid g_1) = \frac{f(\theta \mid g_1)\,\mathrm{I}(\theta)}{\int f(\theta \mid g_1)\,\mathrm{I}(\theta)d\theta}, \quad w(\theta) = f(g_2 \mid \theta)$$

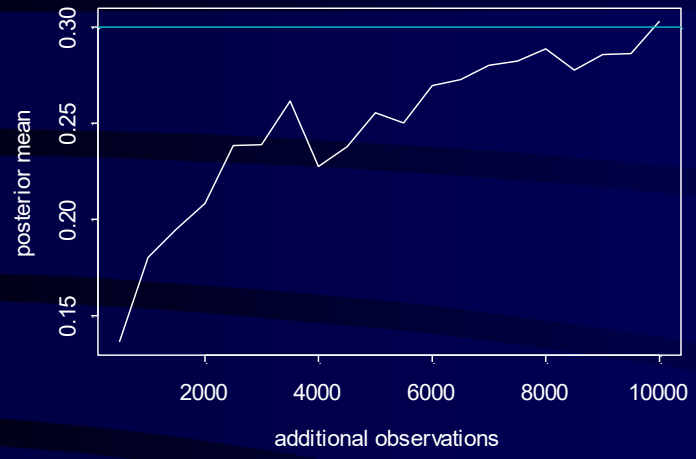
Preliminary results

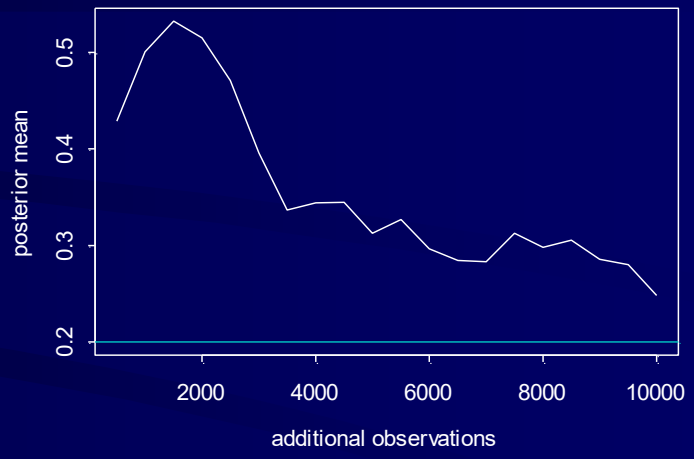
- Simulate 20,000 observations from a logistic regression model, half for g_1 and half for g_2.
- Use BUGS to draw from $f(\theta \mid g_1)$, m=10,000.
- Iterate through the $(\underline{x}_j, y_j)$ in g_2 and compute w_i.

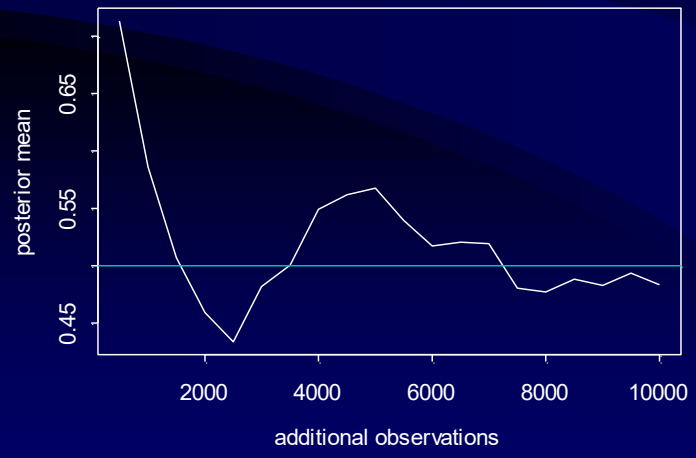
for i in 1,..., m do

$$\log(w_i) \leftarrow \log(w_i) + \log f(x_j \mid \theta_i)$$

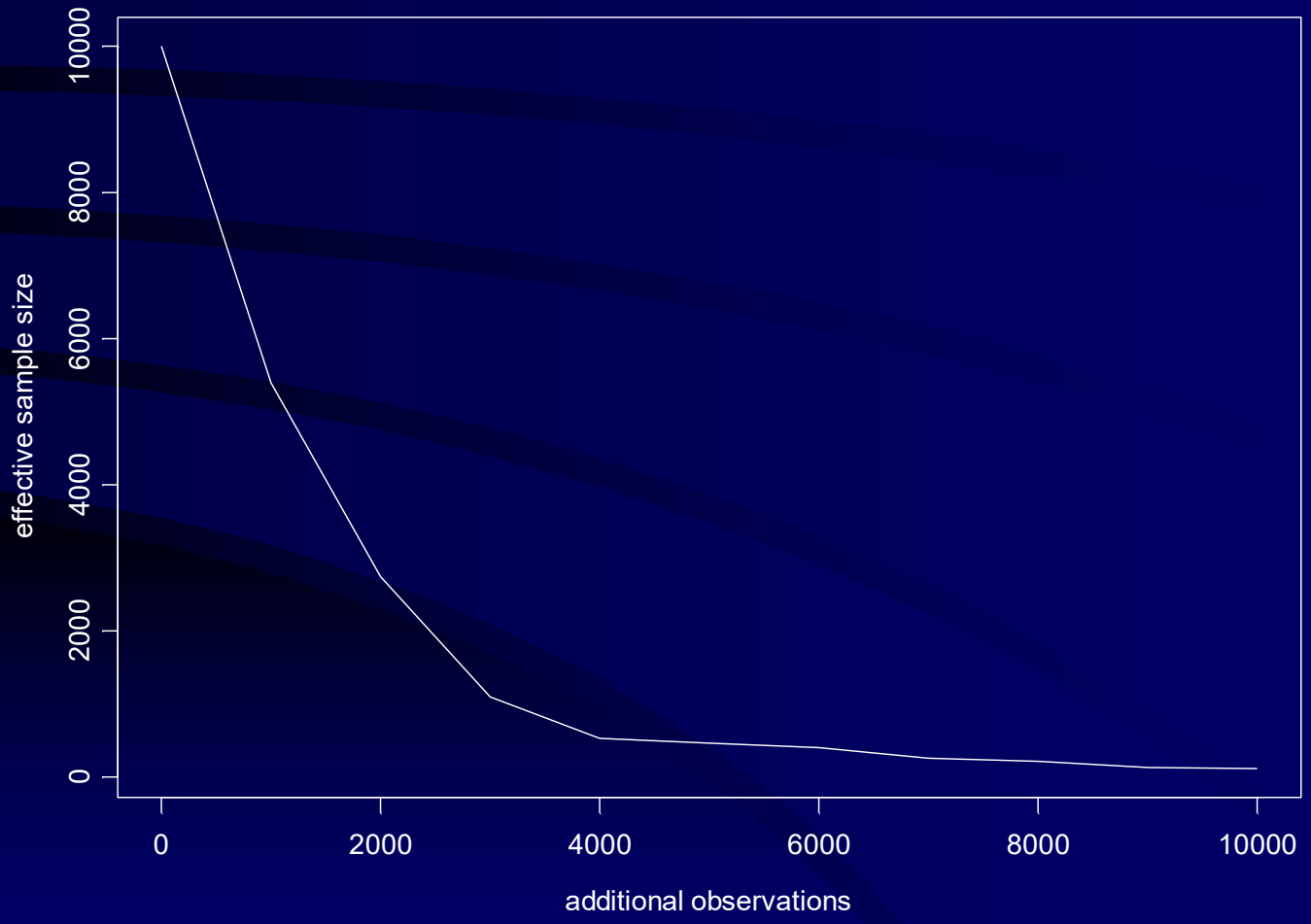
Posterior mean convergence



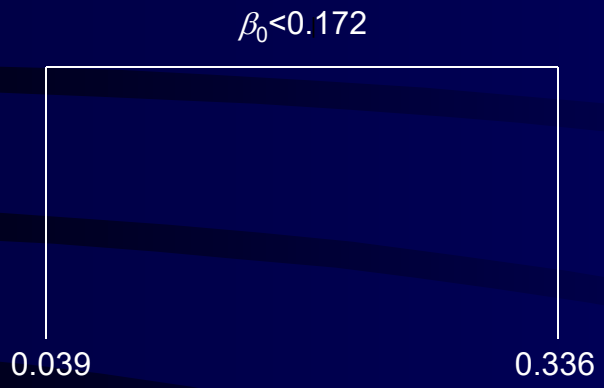


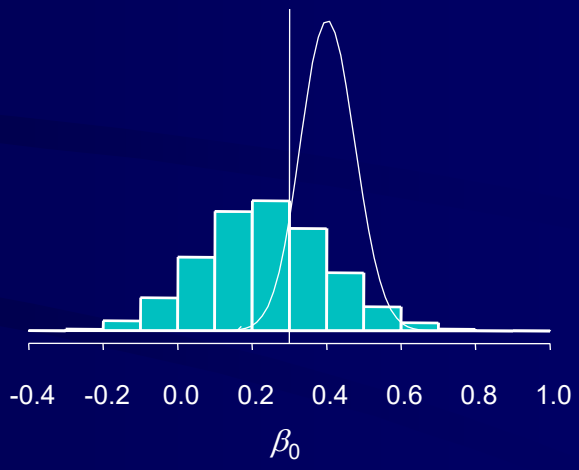


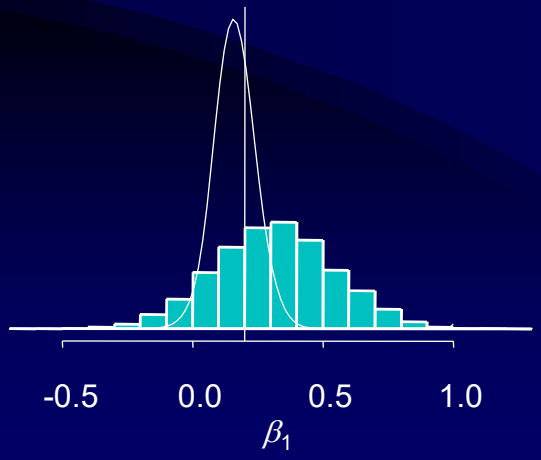
Effective sample size

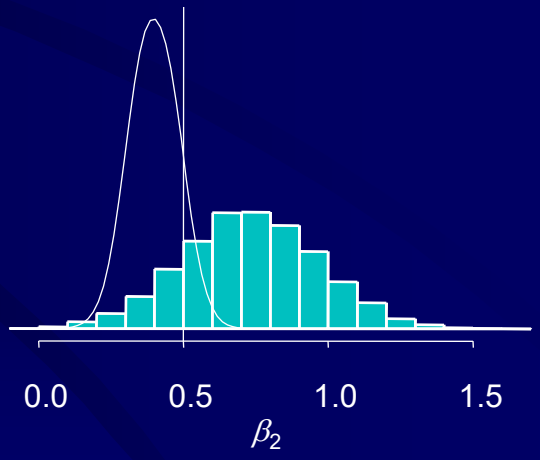


Predictive weight trimming

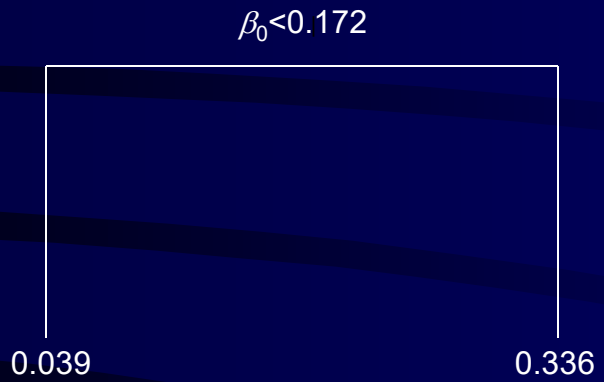


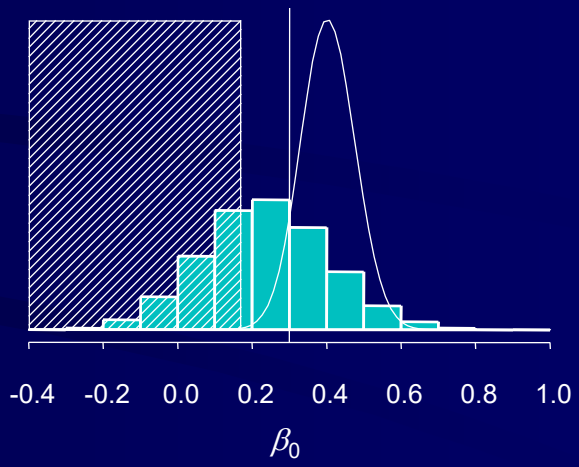


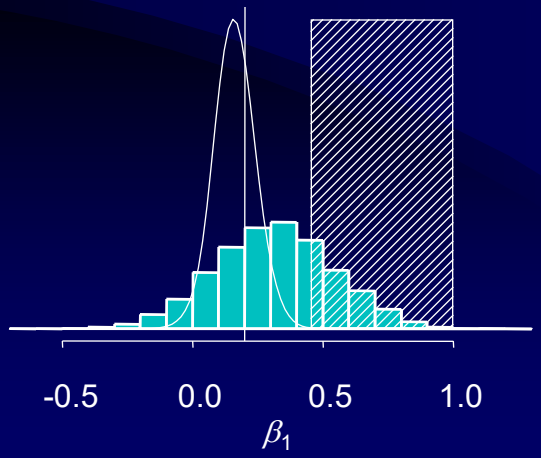


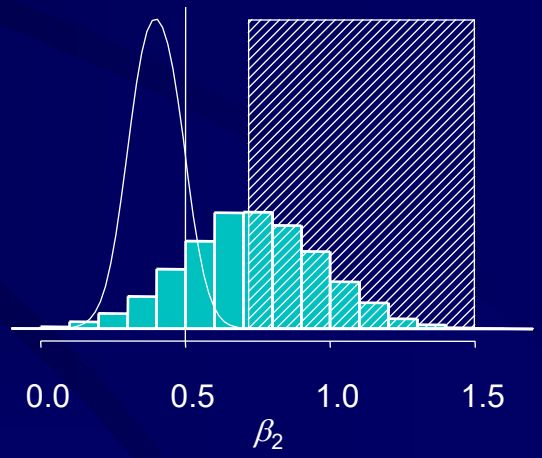


Predictive weight trimming









Features

- Using trees is much faster than using g_2 to indicate 0 weight.
- The trees are easy to construct.
- We can trim 70% of the draws which combined only have the weight of 1/5 of a draw.

Future Work

- Evaluate predictive weight trimming
- Investigate the convergence properties of the adaptive importance sampling
- Apply fast vector quantization algorithms to
  - select g_1
  - generate smaller pseudo-datasets with nearly equivalent information

Predictive weight trimming

- Draw $\theta_1,\ldots,\theta_m$ from $f(\theta \mid g_1)$.
- Compute $w_i = f(g_2|\theta_i)$ where $|g_2| = n_1$.
- Construct a weight threshold predictor, I(θ)

$$1(w > \varepsilon) \approx \mathrm{I}(\theta)$$

- Now sample from the mixture

$$f_\alpha(\theta \mid g_1) = \alpha \frac{f(\theta \mid g_1) I(\theta)}{\int f(\theta \mid g_1) I(\theta) d\theta} + (1-\alpha) f(\theta \mid g_1)$$