Scalable Spike-and-Slab

Niloy Biswas
Harvard University

Joint work with
Lester Mackey (Microsoft Research)
Xiao-Li Meng (Harvard)

International Conference on Machine Learning 2022

Variable selection with continuous spike-and-slab priors

High-dimensional data: $y \in \mathbb{R}^n$, design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $n \ll p$.

Linear regression: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}$ where $\epsilon \sim \mathcal{N}(0, I_n)$

Continuous Spike-and-Slab Prior [George and McCulloch, 1993]

$$
\begin{aligned}
&\sigma^2 \sim \mathrm{InvGamma}\Big(\frac{a_0}{2}, \frac{b_0}{2}\Big) \\
&z_j \underset{j=1,\dots,p}{\overset{i.i.d.}{\sim}} \mathrm{Bernoulli}(q), \\
&\beta_j | z_j, \sigma^2 \underset{j=1,\dots,p}{\overset{ind}{\sim}} (1 - z_j)\underbrace{\mathcal{N}(0, \sigma^2\tau_0^2)}_{Spike} + z_j \underbrace{\mathcal{N}(0, \sigma^2\tau_1^2)}_{Slab}
\end{aligned}
$$

Hyperparameters: $q \in (0, 1)$, $\tau_1^2 \gg \tau_0^2 > 0$, and $a_0, b_0 > 0$.
Choose small q to incorporate sparsity.

Variable selection with spike-and-slab priors

High-dimensional data: $y \in \mathbb{R}^n$, design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $n \ll p$.

Linear regression: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}$ where $\epsilon \sim \mathcal{N}(0, I_n)$

Continuous Spike-and-Slab Prior [George and McCulloch, 1993]

$$\sigma^2 \sim \mathrm{InvGamma}\Big(\frac{a_0}{2}, \frac{b_0}{2}\Big)$$

$$z_j \underset{j=1,\dots,p}{\overset{i.i.d.}{\sim}} \mathrm{Bernoulli}(q),$$

$$\beta_j | z_j, \sigma^2 \underset{j=1,\dots,p}{\overset{ind}{\sim}} (1 - z_j)\underbrace{\mathcal{N}(0, \sigma^2\tau_0^2)}_{Spike} + z_j \underbrace{\mathcal{N}(0, \sigma^2\tau_1^2)}_{Slab}$$

Inference using $\mathbb{P}(z_j = 1|\boldsymbol{y})$. Guan and Stephens, 2011, Zhou et al., 2013, ...

How to sample from the posterior?

Bayesian computation for spike-and-slab priors

High-dimensional data: $y \in \mathbb{R}^n$, design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $n \ll p$.

Markov chain Monte Carlo methods:

- Naïve MCMC: $\mathcal{O}(p^3)$ cost per iteration
- State-of-the-art (SOTA) MCMC: $\mathcal{O}(n^2 p)$ cost per iteration
  Bhattacharya, 2016
- *For large datasets, $\mathcal{O}(n^2 p)$ cost can become prohibitive.*
  e.g. GWAS with $n \approx 10^3, p \approx 10^5$, SOTA takes 1 minute per iteration

Approximate inference methods:

- Approx. MCMC: $\mathcal{O}(\max\{n\|z_t\|_1^2, np\})$ cost at iteration t
  Narisetty et al., 2019
- Variational inference Ray et al., 2020, Ray and Szabó, 2021
- *Does not converge to the spike-and-slab posterior*

Faster Bayesian computation which converges to spike-and-slab posterior?

Bayesian computation for spike-and-slab priors

High-dimensional data: $y \in \mathbb{R}^n$, design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $n \ll p$.

Markov chain Monte Carlo methods:

- Naïve MCMC: $\mathcal{O}(p^3)$ cost per iteration
- State-of-the-art (SOTA) MCMC: $\mathcal{O}(n^2 p)$ cost per iteration
- *For large datasets, $\mathcal{O}(n^2 p)$ cost become prohibitive.*
  e.g. GWAS with $n \approx 10^3, p \approx 10^5$, SOTA takes 1 minute per iteration

Faster Bayesian computation which converges to spike-and-slab posterior?
Scalable Spike-and-Slab (S^3) MCMC: $\mathcal{O}(\max\{n^2 p_t, np\})$ cost at iteration t

p_t never larger than $\|z_t - z_{t-1}\|_1$: the number of covariates switching spike-and-slab states between iterations t and $t-1$.
e.g. for GWAS with $n \approx 10^3, p \approx 10^5$, 50$\times$ faster than SOTA