Estimating Sparse High Dimensional Linear Models using Global-Local Shrinkage

Daniel F. Schmidt

Centre for Biostatistics and Epidemiology
The University of Melbourne

Monash University
May 11, 2017

Outline

1. Bayesian Global-Local Shrinkage Hierarchies
   - Linear Models
   - Bayesian Estimation of Linear Models
   - Global-Local Shrinkage Hierarchies

2. Bayesreg toolbox
   - The Bayesreg hierarchy
   - Gibbs sampling
   - The toolbox

Outline

1. Bayesian Global-Local Shrinkage Hierarchies
   - Linear Models
   - Bayesian Estimation of Linear Models
   - Global-Local Shrinkage Hierarchies
2. Bayesreg toolbox
   - The Bayesreg hierarchy
   - Gibbs sampling
   - The toolbox

Problem Description

- Consider the linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \beta_0 \mathbf{1}_n + \boldsymbol{\varepsilon},$$

  where
  - $\mathbf{y} \in \mathbb{R}^n$ is a vector of targets;
  - $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a matrix of features;
  - $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of regression coefficients;
  - $\beta_0 \in \mathbb{R}$ is the intercept;
  - $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a vector of random disturbances.
- Let $\boldsymbol{\beta}^*$ denote the *true* values of the coefficients
- Task: we observe $\mathbf{y}$, $\mathbf{X}$ and must estimate $\boldsymbol{\beta}^*$
  - We do not require that $p < n$

Sparse Linear Models (1)

- The value $\beta_j^* = 0$ is special
  $\Rightarrow$ means that feature j is not associated with targets
- Define the index of sparsity by

$$||\boldsymbol{\beta}^*||_0,$$

  where $||\mathbf{x}||_0$ is the ℓ_0 (counting) "norm"
- A linear model is sparse if $||\boldsymbol{\beta}^*||_0 \ll p$
- Sparsity is useful when $p \geq n$
  - Enables us to estimate less entries of $\boldsymbol{\beta}$
  - If trying to find which $\beta_j^* \neq 0$, conditions on $||\boldsymbol{\beta}^*||_0$ required

Sparse Linear Models (2)

- Why is sparsity useful?
- Loosely, an estimator sequence $\hat{\theta}_n$ is asymptotically *efficient* if

$$\limsup_{n\to\infty}\{\mathbb{E}[(\hat{\theta}_n - \theta^*)^2]\} = 1/J(\theta^*)$$

  where $J(\theta^*)$ is the Fisher information.
- Estimators exist for which the above bound can be beaten but only on a set of measure zero (Hodges 51, Le Cam 53)
- Sparse models have a special set of measure zero
  - The set $\beta_j^* = \{0\}$ has measure zero, but is extremely important
  - Good sparse estimators achieve superefficiency for $\beta_j^* = 0$

Maximum Likelihood Estimation of $\boldsymbol{\beta}$

- Assume we have a probabilistic model for the disturbances ε_i
  - Then, a standard way of estimating $\boldsymbol{\beta}$ is maximum likelihood

$$\{\hat{\boldsymbol{\beta}}, \hat{\beta}_0\} = \underset{\boldsymbol{\beta}, \beta_0}{\arg\max} \{p(\mathbf{y} \,|\, \boldsymbol{\beta}, \beta_0, \mathbf{X})\}$$

- If $\varepsilon_i \sim N(0, \sigma^2)$, then $\hat{\boldsymbol{\beta}}$ is the least squares estimator.

- Has several drawbacks:
  - Requires $p < n$ for uniqueness
  - Potentially high variance
  - Cannot produce sparse estimates

- Traditional “fixes” to maximum likelihood
  - Remove some covariates
  - Exploits sparsity

Penalized Regression (1)

Method	Type	Comments
Ridge	Convex	(+) Computationally efficient (−) Suffers from potentially high estimation bias
Lasso Elastic net	Convex	(+) Convex optimisation problem (+) Can produce sparse estimates (−) Suffers from potentially high estimation bias (−) Can have model selection consistency problems
Non-convex shrinkers (SCAD, MCP, etc.)	Non-convex	(+) Reduced estimation bias (+) Improved model selection consistency (+) Can produce sparse estimates (−) Non-convex optimisation; difficult, multi-modal
Subset selection	Non-convex	(+) Model selection consistency (−) Computationally intractable (−) High statistically unstable

Penalized Regression (2)

- All methods require an additional model selection step
    - Cross validation
    - Information criteria
    - Asymptotically optimal choices
- Quantifying statistical uncertainty is problematic
    - Uncertainty in λ difficult to incorporate
    - For sparse methods standard errors of β difficult
      $\Rightarrow$ Bootstrap requires special modifications
- Bayesian inference provides natural solutions to these problems

Bayesian Linear Regression (1)

- Assuming normal disturbances, the Bayesian regression

$$\begin{aligned} \mathbf{y} \,|\, \boldsymbol{\beta}, \beta_0 &\sim N(\mathbf{X}\boldsymbol{\beta} + \beta_0 \mathbf{1}_n, \sigma^2 \mathbf{I}_n), \\ \beta_0 &\sim d\beta_0, \\ \boldsymbol{\beta} \,|\, \sigma^2 &\sim \pi(\boldsymbol{\beta} \,|\, \sigma^2) d\boldsymbol{\beta}, \end{aligned}$$

  where
  - $\pi(\boldsymbol{\beta} \,|\, \sigma^2)$ is a prior distribution over $\boldsymbol{\beta}$;
  - σ^2 is the noise variance.

- Inferences about $\boldsymbol{\beta}$ formed using the posterior distribution

$$\pi(\boldsymbol{\beta}, \beta_0 \,|\, \mathbf{y}) \propto p(\mathbf{y} \,|\, \boldsymbol{\beta}, \beta_0, \sigma^2) \pi(\boldsymbol{\beta} \,|\, \sigma^2).$$

  Inference usually performed by MCMC sampling.

Bayesian Linear Regression (2)

- "Spike-and-slab" variable selection

$$\begin{aligned}
\mathbf{y} \,|\, \boldsymbol{\beta}, \beta_0 &\sim N(\mathbf{X}\boldsymbol{\beta} + \beta_0 \mathbf{1}_n, \sigma^2 \mathbf{I}_n), \\
\beta_0 &\sim d\beta_0, \\
\beta_j \,|\, I_j, \sigma^2 &\sim \left[I_j \pi(\beta_j \,|\, \sigma^2) + (1 - I_j)\delta_0(\beta_j) \right] d\beta_j, \\
I_j &\sim \mathrm{Be}(\alpha) \\
\alpha &\sim \pi(\alpha) d\alpha
\end{aligned}$$

  where

  - $I_j \in \{0, 1\}$ are indicators and $\mathrm{Be}(\cdot)$ is a Bernoulli distribution;
  - $\delta_z(x)$ denotes at a Dirac point-mass at $x = z$;
  - $\alpha \in (0, 1)$ is the *a priori* inclusion probability.

- Considered "gold standard"

  $\Rightarrow$ computationally intractable as involves exploring 2^p models

Bayesian Linear Regression (3)

- Variable selection with continuous shrinkage priors

- Treat the prior distribution for $\boldsymbol{\beta}$ as a Bayesian penalty
  - Taking
    $$\beta_j \,|\, \sigma^2, \lambda \sim \mathrm{N}(0, \lambda^2 \sigma^2)$$
    leads to Bayesian ridge regression;
  - or, taking
    $$\beta_j \,|\, \sigma^2, \lambda \sim \mathrm{La}(0, \lambda/\sigma)$$
    where $\mathrm{La}(a, b)$ is a Laplace distribution with location a and scale b leads to Bayesian lasso.

- More generally ...

Global-Local Shrinkage Hierarchies (1)

- The global-local shrinkage hierarchy
  ⇒ generalises many popular Bayesian regression priors

$$\begin{aligned}
\mathbf{y} \,|\, \boldsymbol{\beta}, \beta_0, \sigma^2 &\sim N(\mathbf{X}\boldsymbol{\beta} + \beta_0 \mathbf{1}_n, \sigma^2 \mathbf{I}_n), \\
\beta_0 &\sim d\beta_0, \\
\beta_j \,|\, \lambda_j^2, \tau^2, \sigma^2 &\sim N(0, \lambda_j^2 \tau^2 \sigma^2) \\
\lambda_j &\sim \pi(\lambda_j) d\lambda_j \\
\tau &\sim \pi(\tau) d\tau
\end{aligned}$$

- Models priors for β_j as scale-mixtures of normals
  ⇒ choice of $\pi(\lambda_j)$, $\pi(\tau)$ controls behaviour

Global-Local Shrinkage Hierarchies (2)

- The global-local shrinkage hierarchy
  ⇒ generalises many popular Bayesian regression priors

$$\begin{aligned}
\mathbf{y} \,|\, \boldsymbol{\beta}, \beta_0, \sigma^2 &\sim N(\mathbf{X}\boldsymbol{\beta} + \beta_0 \mathbf{1}_n, \sigma^2 \mathbf{I}_n), \\
\beta_0 &\sim d\beta_0, \\
\beta_j \,|\, \lambda_j^2, \tau^2, \sigma^2 &\sim N(0, \lambda_j^2 \tau^2 \sigma^2) \\
\lambda_j &\sim \pi(\lambda_j) d\lambda_j \\
\tau &\sim \pi(\tau) d\tau
\end{aligned}$$

- Local shrinkers λ_j control selection of important variables
  ⇒ play the role of indicators I_j in spike-and-slab

Global-Local Shrinkage Hierarchies (3)

- The global-local shrinkage hierarchy
  ⇒ generalises many popular Bayesian regression priors

$$\begin{array}{rcl} \mathbf{y} \,|\, \boldsymbol{\beta}, \beta_0, \sigma^2 & \sim & N(\mathbf{X}\boldsymbol{\beta} + \beta_0 \mathbf{1}_n, \sigma^2 \mathbf{I}_n), \\ \beta_0 & \sim & d\beta_0, \\ \beta_j \,|\, \lambda_j^2, \tau^2, \sigma^2 & \sim & N(0, \lambda_j^2 \tau^2 \sigma^2) \\ \lambda_j & \sim & \pi(\lambda_j) d\lambda_j \\ \tau & \sim & \pi(\tau) d\tau \end{array}$$

- Global shrinker τ controls for multiplicity
  ⇒ plays the role of inclusion probability α in spike-and-slab

Local Shrinkage Priors (1)

- What makes a good prior for local variance components?
- Denote the marginal prior of β_j by

$$\pi(\beta_j \,|\, \tau, \sigma) = \int_0^\infty \left(\frac{1}{\lambda_j^2 \tau^2 \sigma^2} \right)^{\frac{1}{2}} \exp\left(-\frac{\beta_j^2}{2\lambda_j^2 \tau^2 \sigma^2} \right) \pi(\lambda_j) d\lambda_j$$

- Carvalho, Polson and Scott (2010) proposed two desirable properties of $\pi(\beta_j \,|\, \tau, \sigma)$

Local Shrinkage Priors (2)

- Two desirable properties:
  1. Should concentrate sufficient mass near $\beta_j = 0$ such that

     $$\lim_{\beta_j \to 0} \pi(\beta_j \,|\, \tau, \sigma) \to \infty$$

     to guarantee fast rate of posterior contraction when $\beta_j^* = 0$
  2. Should have sufficiently heavy tails so that

     $$\mathbb{E}\left[\beta_j \,|\, \mathbf{y}\right] = \hat{\beta}_j + o_{\hat{\beta}_j}(1)$$

     to guarantee asymptotic (in effect-size) unbiasedness

Local Shrinkage Priors (2)

- Two desirable properties:
  1. Should concentrate sufficient mass near $\beta_j = 0$ such that

$$\lim_{\beta_j \to 0} \pi(\beta_j \,|\, \tau, \sigma) \to \infty$$

     to guarantee fast rate of posterior contraction when $\beta_j^* = 0$

  2. Should have sufficiently heavy tails so that

$$\mathbb{E}\left[\beta_j \,|\, \mathbf{y}\right] = \hat{\beta}_j + o_{\hat{\beta}_j}(1)$$

     to guarantee asymptotic (in effect-size) unbiasedness

Local Shrinkage Priors (3)

- Classic shrinkage priors do not satisfy either property
  - Bayesian ridge takes $\lambda_j \sim \delta_1(\lambda_j) d\lambda_j$, leading to

    $$\beta_j \,|\, \tau, \sigma^2 \sim N(0, \tau^2 \sigma^2).$$

    which expects β_js to be same squared magnitude

    1. Does not model sparsity
    2. Large bias if β^* mix of weak and strong signals

  - Bayesian lasso takes $\lambda_j \sim \mathrm{Exp}(1)$, lead to

    $$\beta_j \,|\, \tau, \sigma \sim \mathrm{La}(0, 2^{-3/2} \sigma \tau)$$

    which expects β_js to be same absolute magnitude

    1. Super-efficient at $\beta_j^* = 0$ but not fast enough contraction,
    2. Large bias if β^* sparse with few strong signals

Local Shrinkage Priors (3)

- Classic shrinkage priors do not satisfy either property
  - Bayesian ridge takes $\lambda_j \sim \delta_1(\lambda_j) d\lambda_j$, leading to

    $$\beta_j \,|\, \tau, \sigma^2 \sim N(0, \tau^2 \sigma^2).$$

    which expects β_js to be same squared magnitude

    1. Does not model sparsity
    2. Large bias if $\boldsymbol{\beta}^*$ mix of weak and strong signals

  - Bayesian lasso takes $\lambda_j \sim \mathrm{Exp}(1)$, lead to

    $$\beta_j \,|\, \tau, \sigma \sim \mathrm{La}(0, 2^{-3/2} \sigma \tau)$$

    which expects β_js to be same absolute magnitude

    1. Super-efficient at $\beta_j^* = 0$ but not fast enough contraction,
    2. Large bias if $\boldsymbol{\beta}^*$ sparse with few strong signals

The horseshoe prior (1)

- The “horseshoe” prior satisfies both properties
  - The horseshoe prior takes

    $$\lambda_j \sim \mathrm{C}^+(0,1),$$

    with $C^+(0,A)$ a half-Cauchy distribution with scale A.
  - Does not admit closed-form for marginal prior, but has bounds

    $$\frac{K}{2}\log\left(1+\frac{4}{b^2}\right) < \pi(\beta_j|\tau,\sigma) < \frac{K}{2}\log\left(1+\frac{2}{b^2}\right),$$

    where $b = \beta_j\tau\sigma$ and $K = (2\pi^3)^{-1/2}$.
    - ✓ Has a pole at $\beta_j = 0$;
    - ✓ Has polynomial tails in β_j

The horseshoe prior (1)

- The "horseshoe" prior satisfies both properties
  - The horseshoe prior takes

    $$\lambda_j \sim \mathrm{C}^+(0,1),$$

    with $C^+(0, A)$ a half-Cauchy distribution with scale A.
  - Does not admit closed-form for marginal prior, but has bounds

    $$\frac{K}{2}\log\left(1+\frac{4}{b^2}\right) < \pi(\beta_j|\tau,\sigma) < \frac{K}{2}\log\left(1+\frac{2}{b^2}\right),$$

    where $b = \beta_j\tau\sigma$ and $K = (2\pi^3)^{-1/2}$.
    - ✓ Has a pole at $\beta_j = 0$;
    - ✓ Has polynomial tails in β_j

The horseshoe prior (1)

- The "horseshoe" prior satisfies both properties
  - The horseshoe prior takes

    $$\lambda_j \sim \mathrm{C}^{+}(0, 1),$$

    with $C^{+}(0, A)$ a half-Cauchy distribution with scale A.
  - Does not admit closed-form for marginal prior, but has bounds

    $$\frac{K}{2} \log\left(1 + \frac{4}{b^2}\right) < \pi(\beta_j | \tau, \sigma) < \frac{K}{2} \log\left(1 + \frac{2}{b^2}\right),$$

    where $b = \beta_j \tau \sigma$ and $K = (2\pi^3)^{-1/2}$.
    - ✓ Has a pole at $\beta_j = 0$;
    - ✓ Has polynomial tails in β_j

The horseshoe prior (1)

- The "horseshoe" prior satisfies both properties
  - The horseshoe prior takes

    $$\lambda_j \sim \mathrm{C}^+(0,1),$$

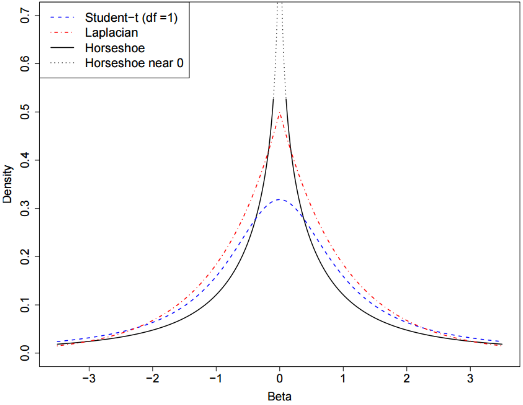
    with $C^+(0,A)$ a half-Cauchy distribution with scale A.
  - Does not admit closed-form for marginal prior, but has bounds

    $$\frac{K}{2}\log\left(1+\frac{4}{b^2}\right) < \pi(\beta_j|\tau,\sigma) < \frac{K}{2}\log\left(1+\frac{2}{b^2}\right),$$

    where $b = \beta_j \tau \sigma$ and $K = (2\pi^3)^{-1/2}$.
    - ✓ Has a pole at $\beta_j = 0$;
    - ✓ Has polynomial tails in β_j

The horseshoe prior (2)

- Flat, Cauchy-like tails and infinitely tall spike at the origin



The horseshoe prior and two close cousins: Laplacian and Student-t.

Higher order horseshoe priors

- More generally, we can model λ_j as a product of k half-Cauchy variables

- The HS$_k$ (our notation) prior is

$$\lambda_j \sim C_1 C_2 \ldots C_k$$

where $C_i \sim C^+(0,1)$, $i = 1, \ldots, k$.

- Generalises several existing priors
  - HS$_0$ is ridge regression;
  - HS$_1$ is the usual horseshoe;
  - HS$_2$ is the horseshoe+ prior (Bhadra et al, 2015).

- Tail weight and mass at $\beta_j = 0$ increase as k grows $\Rightarrow$ models β as increasingly sparse

Horseshoe estimator

- The horseshoe estimator also takes

  $$\tau \sim C^{+}(0,1)$$

  though most heavy tailed priors will perform similarly
- How does the horseshoe prior work in practice?
  - The horseshoe prior works well high dimensional, sparse regressions
  - Experiments show it performs similarly to “spike-and-slab” at variable selection
  - Continuous nature of prior means mixing is much better for large p
  - Posterior mean has strong prediction properties

Outline

1. Bayesian Global-Local Shrinkage Hierarchies
   - Linear Models
   - Bayesian Estimation of Linear Models
   - Global-Local Shrinkage Hierarchies

2. Bayesreg toolbox
   - The Bayesreg hierarchy
   - Gibbs sampling
   - The toolbox

A Bayesian regression toolbox (1)

Motivation

- We have lots of genomic/epigenomic data
  - Large numbers of genomic markers measured along genome
  - Associated disease outcomes (breast and prostate cancer, etc.)
  - Dimensionality is large (total $p > 5,000,000$ in some cases).
  - Number of true associations expected to be small
  - Both association discovery and prediction of importance
- We wished to apply horseshoe-type methods
- But no flexible, easy-to-use and efficient toolbox existed
- So we (myself, Enes Makalic) wrote the Bayesreg toolbox for MATLAB and R

A Bayesian regression toolbox (2)

- Requirements:
    - Handle large p (at least 10,000$+$)
    - Implement both normal and logistic linear regression
    - Implement the horseshoe priors
- Additionally:
    - Handle group-structures within variables, for example, genes
    - Perform (grouped) variable selection even when $p > n$

The Bayesreg hierarchy (1)

- Bayesreg uses the following hierachy

$$\begin{aligned} z_i \,|\, \mathbf{x}_i, \boldsymbol{\beta}, \beta_0, \omega_i^2, \sigma^2 &\sim N(\mathbf{x}_i' \boldsymbol{\beta} + \beta_0, \sigma^2 \omega_i^2), \\ \sigma^2 &\sim \sigma^{-2} \, d\sigma^2, \\ \omega_i^2 &\sim \pi(\omega_i^2) \, d\omega_i^2, \\ \beta_0 &\sim d\beta_0, \\ \beta_j \,|\, \lambda_j^2, \tau^2, \sigma^2 &\sim N(0, \lambda_j^2 \tau^2 \sigma^2), \\ \lambda_j^2 &\sim \pi(\lambda_j^2) \, d\lambda_j^2, \\ \tau^2 &\sim \pi(\tau^2) \, d\tau^2, \end{aligned}$$

- We use scale-mixture representation of likelihood
- Continuous data $z_i = y_i$; binary data $z_i = (y_i - 1/2)/\omega_i^2$

The Bayesreg hierarchy (2)

- Bayesreg uses the following hierachy

$$\boxed{\begin{array}{rcl} z_i \,|\, \mathbf{x}_i, \boldsymbol{\beta}, \beta_0, \omega_i^2, \sigma^2 & \sim & N(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0, \sigma^2\omega_i^2), \\ \sigma^2 & \sim & \sigma^{-2}\, d\sigma^2, \\ \omega_i^2 & \sim & \pi(\omega_i^2)\, d\omega_i^2, \end{array}}$$

$$\begin{array}{rcl} \beta_0 & \sim & d\beta_0, \\ \beta_j \,|\, \lambda_j^2, \tau^2, \sigma^2 & \sim & N(0, \lambda_j^2\tau^2\sigma^2), \\ \lambda_j^2 & \sim & \pi(\lambda_j^2)\, d\lambda_j^2, \\ \tau^2 & \sim & \pi(\tau^2)\, d\tau^2, \end{array}$$

- The z_i follow a (potentially) heteroskedastic Gaussian $\Rightarrow \pi(\omega_i)$ determines data model (normal, logistic, etc.)

The Bayesreg hierarchy (3)

- Bayesreg uses the following hierachy

$$\begin{aligned} z_i \,|\, \mathbf{x}_i, \boldsymbol{\beta}, \beta_0, \omega_i^2, \sigma^2 &\sim N(\mathbf{x}_i' \boldsymbol{\beta} + \beta_0, \sigma^2 \omega_i^2), \\ \sigma^2 &\sim \sigma^{-2} \, d\sigma^2, \\ \omega_i^2 &\sim \pi(\omega_i^2) \, d\omega_i^2, \end{aligned}$$

$$\boxed{\begin{aligned} \beta_0 &\sim d\beta_0, \\ \beta_j \,|\, \lambda_j^2, \tau^2, \sigma^2 &\sim N(0, \lambda_j^2 \tau^2 \sigma^2), \\ \lambda_j^2 &\sim \pi(\lambda_j^2) \, d\lambda_j^2, \\ \tau^2 &\sim \pi(\tau^2) \, d\tau^2, \end{aligned}}$$

- The priors for $\boldsymbol{\beta}$ follow a global-local shrinkage hierarchy $\Rightarrow \pi(\lambda_j^2)$ determines the estimator (horseshoe, lasso, etc.)

Gibbs sampling (1)

- Sampler for $\beta \mid \cdots$
  - Is a multivariate normal of the form

$$\beta \mid \cdots \sim N(\mathbf{A}^{-1}\mathbf{e}, \mathbf{A}^{-1})$$

    where $\mathbf{A} = (\mathbf{B} + \mathbf{D})$ and $\mathbf{D}$ is diagonal.
  - This form allows for specialised sampling algorithms
    - If $p/n < 2$ we use Rue's algorithm $O(p^3)$
    - Otherwise we use Bhattarchaya's algorithm, $O(n^2 p)$

- Sampler for $\sigma^2 \mid \cdots$
  - We integrate out the βs to improve mixing
  - Conditional distribution is an inverse-gamma
    $\Rightarrow$ Uses quantities computed when sampling β

Gibbs sampling (2)

Sampler for $\lambda_j \,|\, \cdots$

- Recall that $\lambda_j \sim C^+(0,1)$
  - Conditional distribution for λ_j is

$$\pi(\lambda_j \,|\, \beta_j, \tau, \sigma) \propto \left(\frac{1}{\lambda_j^2 \tau^2 \sigma^2} \right)^{\frac{1}{2}} \exp\left(-\frac{\beta_j^2}{2\lambda_j^2 \tau^2 \sigma^2} \right) (1 + \lambda_j^2)^{-1}$$

  which is not a standard distribution

- When we started this work in 2015, only one slow slice sampler (monomvn) existed for horseshoe
- Since our implementation there have been several competing samplers

Alternative Horseshoe Samplers

- Slice sampling
  - Heavy tails can cause mixing issues
  - Requires CDF inversions
  - Does not easily extend to higher-order horseshoe priors
- NUTS sampler (Stan implementation)
  - Very slow
  - Numerically unstable for true horseshoe
  - Unable to handle heavier tailed priors
- Elliptical slice sampler
  - Computationally efficient
  - Cannot be applied if $p > n$
  - Cannot handle grouped variable structures

Our approach (1)

- Based on auxiliary variables
- Let x and a be random variables such that

$$x^2 \,|\, a \sim IG(1/2, 1/a), \quad \text{and} \quad a \sim IG(1/2, 1/A^2)$$

then $x \sim C^+(0, A)$, where $IG(\cdot, \cdot)$ denotes the inverse-gamma distribution with pdf

$$p(z \,|\, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} \exp\left(-\frac{\beta}{z}\right)$$

- Inverse-gamma conjugate with normal for scale parameters
    - Also conjugate with itself

Our approach (2)

- Rewrite prior for β_j as

$$\begin{aligned} \beta_j \,|\, \lambda_j^2, \tau^2, \sigma^2 &\sim N(0, \lambda_j^2 \tau^2 \sigma^2) \\ \lambda_j^2 | \nu_j &\sim IG(1/2, 1/\nu_j) \\ \nu_j &\sim IG(1/2, 1) \end{aligned}$$

- Leads to simple Gibbs sampler for λ_j and ν_j

$$\begin{aligned} \lambda_j^2 \,|\, \cdot &\sim IG\left(1, \frac{1}{\nu_j} + \frac{\beta_j^2}{2\tau^2\sigma^2}\right), \\ \nu_j \,|\, \cdot &\sim IG\left(1, 1 + \frac{1}{\tau^2}\right) \end{aligned}$$

- Both are simply inverted exponential random variables
  $\Rightarrow$ extremely quick and stable sampling

Higher order horseshoe priors

- The HS$_k$ prior is $\lambda_j \sim C_1 C_2 \ldots C_k$, where $C_i \sim C^+(0,1)$.
- Prior for λ_j has very complex form, but
  - Can rewrite prior as the hierarchy

$$\begin{aligned}
\lambda_j &\sim C^+(0, \phi_j^{(1)}) \\
\phi_j^{(1)} &\sim C^+(0, \phi_j^{(2)}) \\
&\vdots \\
\phi_j^{(k-1)} &\sim C^+(0, 1)
\end{aligned}$$

  - We can apply our expansion to easily sample λ_j and the $\phi_j^{(\cdot)}$s $\Rightarrow$ currently only sampler than can efficiently handle $k > 1$

Group structures (1)

Group structures exist naturally in predictor variables

- A multi-level categorical predictor - a group of dummy variables
- A continuous predictor - composition of basis functions (additive models)
- Prior knowledge such as genes grouped in the same biological pathway - a natural group
- We wanted our toolbox to take exploit such structures

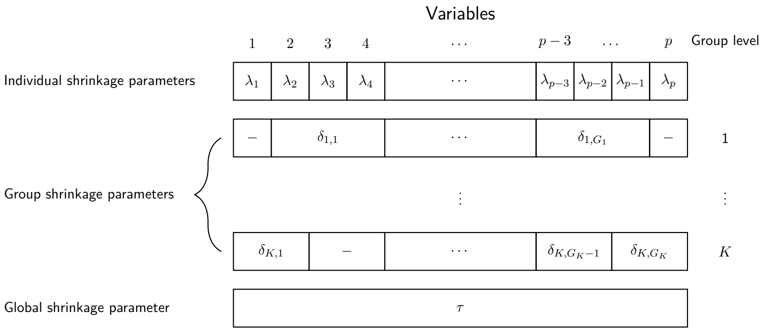
Group structures (2)

- We add additional group-specific shrinkage parameters
  - For convenience we assume K levels of disjoint groupings
  - Assume β_j belongs to group g_k at level k

$$\beta_j \mid \cdots \sim N(0, \lambda_j^2 \delta_{1,g_1}^2 \cdots \delta_{K,g_K}^2 \tau^2 \sigma^2)$$

- Group shrinkers are given appropriate prior distributions
  - Our horseshoe sampler trivially adapted to group shrinkers $\Rightarrow$ conditional distribution of δ_{k,g_k} is inverse-gamma
  - In contrast, slice-sampler requires inversions of gamma CDFs
- Paper detailing this work about to be submitted

Group structures (3)



Variables

	1	2	3	4	$\cdots$	$p-3$	$\cdots$		p	Group level
Individual shrinkage parameters	λ_1	λ_2	λ_3	λ_4	$\cdots$	λ_{p-3}	λ_{p-2}	λ_{p-1}	λ_p	
Group shrinkage parameters	$-$	$\delta_{1,1}$			$\cdots$	δ_{1,G_1}			$-$	1
					$\vdots$					$\vdots$
	$\delta_{K,1}$		$-$		$\cdots$	δ_{K,G_K-1}		δ_{K,G_K}		K
Global shrinkage parameter	τ									

An illustration of possible group structures of total p number of variables with 1 level of individual variables, K levels of grouped variables and 1 level of all variables.

The Bayesreg toolbox (1)

The Bayesreg toolbox currently has:

- Data models
    - Gaussian, logistic, Laplace, Robust student-t regression
- Priors
    - Bayesian ridge and g-prior regression
    - Bayesian lasso
    - Horseshoe
    - Higher order horseshoe (horseshoe$+$, etc.)
- Other features
    - Variable ranking
    - Some basic variable selection criteria
    - Easy to use

The Bayesreg toolbox (2)

- In comparison to Stan:
  - On simple problem with $p = 10$ and $n = 442$
  - Stan took 50 seconds to produce $1,000$ samples
  - Bayesreg took $< 0.1s$
- In comparison to other slice-sampling implementations:
  - Speed and mixing for small p considerable better
  - For large p performance is similar
  - Scope of options much smaller (no horseshoe$+$, no grouping)
- Currently being used by group at University College London to fit logistic regressions for brain lesion work involving $p = 50,000$ predictors

The Bayesreg toolbox (3)

- In current development version:
  - Negative binomial regression for count data
  - Multi-level variable grouping
  - Higher order horseshoe priors (beyond HS_2)
- To be added in the near future:
  - Posterior sparsification tools
  - Autoregressive (time-series) residuals
  - Additional diagnostics

The Bayesreg toolbox (3)

- In current development version:
  - Negative binomial regression for count data
  - Multi-level variable grouping
  - Higher order horseshoe priors (beyond HS_2)
- To be added in the near future:
  - Posterior sparsification tools
  - Autoregressive (time-series) residuals
  - Additional diagnostics

Sparse Bayesian Point Estimates

- We obtain m samples from the posterior
  - What if we want a single point estimate?
- An attractive choice is the Bayes estimator with squared-prediction loss
  - (Potentially) admissable, invariant to reparameterisation
  - Reduces to posterior mean $\bar{\beta}$ of β in standard parameterisation
- Ironically, even if the prior promotes sparsity (i.e., horseshoe), the posterior mean will not be sparse
  - The *exact* posterior mode may be sparse, but is impossible to find from samples
- A number of simple sparsification rules exist
  - Most do not work when $p > n$
  - Largely consider only marginal effects

Decoupled Shrinkage and Selection (DSS) estimator (1)

- Polson et al. (2016) recently introduced the DSS procedure
  1. Obtain samples for β from posterior distribution
  2. Form a new data vector incorporating the effects of shrinkage
     $$\bar{\mathbf{y}} = \mathbf{X}\bar{\beta}.$$
  3. Find "sparsified" approximations of $\bar{\beta}$ by solving
     $$\beta_\lambda = \underset{\beta}{\arg\min} \left\{ ||\bar{\mathbf{y}} - \mathbf{X}\beta||_2^2 + \lambda ||\beta||_0 \right\}$$
     or a similar penalised estimator for different values of λ
  4. Select one of the sparsified models β_λ
- The authors use adaptive lasso in place of intractable ℓ_0 penalisation

Decoupled Shrinkage and Selection (DSS) estimator (1)

- Polson et al. (2016) recently introduced the DSS procedure
  1. Obtain samples for β from posterior distribution
  2. Form a new data vector incorporating the effects of shrinkage
     $$\bar{\mathbf{y}} = \mathbf{X}\bar{\beta}.$$
  3. Find "sparsified" approximations of $\bar{\beta}$ by solving
     $$\beta_\lambda = \operatorname*{arg\,min}_{\beta} \left\{ ||\bar{\mathbf{y}} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_0 \right\}$$
     or a similar penalised estimator for different values of λ
  4. Select one of the sparsified models β_λ
- The authors use adaptive lasso in place of intractable ℓ_0 penalisation

Decoupled Shrinkage and Selection (DSS) estimator (1)

- Polson et al. (2016) recently introduced the DSS procedure
  1. Obtain samples for β from posterior distribution
  2. Form a new data vector incorporating the effects of shrinkage

     $$\bar{\mathbf{y}} = \mathbf{X}\bar{\beta}.$$

  3. Find "sparsified" approximations of $\bar{\beta}$ by solving

     $$\boldsymbol{\beta}_\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ ||\bar{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_0 \right\}$$

     or a similar penalised estimator for different values of λ
  4. Select one of the sparsified models β_λ
- The authors use adaptive lasso in place of intractable ℓ_0 penalisation

Decoupled Shrinkage and Selection (DSS) estimator (1)

- Polson et al. (2016) recently introduced the DSS procedure
  1. Obtain samples for $\boldsymbol{\beta}$ from posterior distribution
  2. Form a new data vector incorporating the effects of shrinkage
     $$\bar{\mathbf{y}} = \mathbf{X}\bar{\boldsymbol{\beta}}.$$
  3. Find "sparsified" approximations of $\bar{\boldsymbol{\beta}}$ by solving
     $$\boldsymbol{\beta}_\lambda = \underset{\boldsymbol{\beta}}{\arg\min} \left\{ ||\bar{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_0 \right\}$$
     or a similar penalised estimator for different values of λ
  4. Select one of the sparsified models $\boldsymbol{\beta}_\lambda$
- The authors use adaptive lasso in place of intractable ℓ_0 penalisation

Decoupled Shrinkage and Selection (DSS) estimator (2)

- While clever, the initial DSS proposal has several weaknesses:
  1. It does not apply to non-continuous data
  2. Selection of degree of sparsification is done by an ad-hoc rule
  3. It cannot be applied to selection of groups of variables
- Current work being done with PhD student Zemei Xu addresses all three problems

Generalised DSS estimator (1)

- We first generalise the procedure to arbitrary data types
  - Let $p(\mathbf{y} \,|\, \boldsymbol{\theta}, \mathbf{X})$ be the data model in our Bayesian hierarchy
  - Partition the parameter vector as $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$
  - $\boldsymbol{\gamma}$ are additional parameters, such as σ^2 if the model is normal

- Given $\mathbf{X}$, the posterior predictive density defines a probability density over possible values of "$\mathbf{y}$", say $\tilde{\mathbf{y}}$, that could arise

$$p(\tilde{\mathbf{y}} \,|\, \mathbf{y}, \mathbf{X}) = \int p(\tilde{\mathbf{y}} \,|\, \boldsymbol{\theta}, \mathbf{X}) \pi(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{X}) d\boldsymbol{\theta}$$

  - Incorporates all posterior beliefs about $\boldsymbol{\theta}$
  - Defines a complete distribution over $\tilde{\mathbf{y}}$

Generalised DSS estimator (2)

- Recall the “ideal” sparsification scheme from DSS:

$$\boldsymbol{\beta}_\lambda = \arg\min_{\boldsymbol{\beta}} \left\{ ||\bar{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_0 \right\}$$

- We can now replace the sum-of-squares goodness of fit term by an expected likelihood goodness-of-fit term

$$L_{\tilde{\mathbf{y}}}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = -\mathbb{E}_{\tilde{\mathbf{y}}} \left[\log p(\tilde{\mathbf{y}} \,|\, \boldsymbol{\beta}, \boldsymbol{\gamma}) \right]$$

  and use a sparsifying penalized likelihood estimator
- Simple for binary data as mixture of Bernoullis is a Bernoulli

Generalised DSS estimator (3)

- Second problem solved by selecting using information criteria
  - Formed as sum of likelihood plus dimensionality penalty
- We adapt information criterion to the DSS problem by using

$$\mathrm{GIC}(\lambda) = \min_{\boldsymbol{\gamma}} \left\{ L_{\tilde{\mathbf{y}}}(\boldsymbol{\beta}_\lambda, \boldsymbol{\gamma}) + \alpha(n, k_\lambda, \boldsymbol{\gamma}) \right\}$$

where $k_\lambda = ||\boldsymbol{\beta}||_0$ is the degrees-of-freedom of $\boldsymbol{\beta}_\lambda$;

- Some common choices for $\alpha(\cdot)$
  - $\alpha(n, k_\lambda, \boldsymbol{\gamma}) = (k_\lambda/2)\log n$ for the BIC;
  - $\alpha(n, k_\lambda, \boldsymbol{\gamma}) = nk_\lambda/(n - k_\lambda - 2)$ for the corrected AIC
  - We also considered an MML criterion
- Select the $\boldsymbol{\beta}_\lambda$ that minimises the information criterion score

Generalised DSS estimator (4)

- We have extended this further to selecting groups of variables
  - Very relevant for testing genes and pathways in genomic data
- We compared our generalised DSS to grouped spike-and-slab
  - Info criterion approach (using MML) outperformed original ad-hoc proposal of Polson et al.
  - Performed as well as spike-and-slab in overall selection error
  - However, over $20,000$ times faster!
- Analysis of iCOGs data is about to begin
  - Very large dataset, $n = 120,000$ and $p = 2,000,000$.

Conclusion

- MATLAB Bayesreg toolbox
  - http://au.mathworks.com/matlabcentral/fileexchange/60335-bayesian-regularized-linear-and-logistic-regression
- R package available as package "bayesreg" from CRAN
- A pre-print describing the toolbox in detail:
  - "High-Dimensional Bayesian Regularised Regression with the BayesReg Package", E. Makalic and D. F. Schmidt, arXiv preprint: https://arxiv.org/pdf/1611.06649v1/
- Thank you – questions?