Continuous shrinkage prior revisited: a collapsing behavior and remedy

Se Yoon Lee, Debdeep Pati, Bani K. Mallick
seyoonlee@stat.tamu.edu debdeep@stat.tamu.edu bmallick@stat.tamu.edu

Department of Statistics, Texas A & M University
High-dimensional Linear Regression
High-dimensional linear regression

Consider the high-dimensional setting: predict a vector $y \in \mathbb{R}^n$ from a set of features $X \in \mathbb{R}^{n \times p}$, with $p \gg n$

$$y = X\beta + \sigma \epsilon, \quad \epsilon \sim \mathcal{N}_n(0, I_n), \quad \text{and } \beta \text{ is sparse}.$$

- $\beta$ is said to be \textit{sparse} if substantially many coefficients in $\beta$ are assumed to be zeros or approximately zeros.

Main objective of high-dimensional linear regression

Main purpose is to estimate $\beta = (\beta_1, \cdots, \beta_j, \cdots, \beta_p)^\top \in \mathbb{R}^p$. 
Example: gene-expression data study

\[
\begin{bmatrix}
  y_1 \\
  \vdots \\
  y_i \\
  \vdots \\
  y_n
\end{bmatrix}
= 
\begin{bmatrix}
  x_{11} & \cdots & x_{1j} & \cdots & x_{1p} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{i1} & \cdots & x_{ij} & \cdots & x_{ip} \\
  \vdots & \ddots & \vdots & \ddots & \vdots \\
  x_{n1} & \cdots & x_{nj} & \cdots & x_{np}
\end{bmatrix}
\begin{bmatrix}
  \beta_1 \\
  \vdots \\
  \beta_j \\
  \vdots \\
  \beta_p
\end{bmatrix}
+ \sigma
\begin{bmatrix}
  \epsilon_1 \\
  \vdots \\
  \epsilon_i \\
  \vdots \\
  \epsilon_n
\end{bmatrix}
\]

- $y_i$: a health response from the $i$-th subject.
- $x_{ij}$: the gene-expression level of $j$-th gene from the $i$-th subject.
Sparsity level in general

- Let $q$ denote the number of signals in $\beta = (\beta_1, \cdots, \beta_p)^T \in \mathbb{R}^p$.
- We define the following ratio as the sparsity level:

**Definition (Sparsity level in general)**

$$s = \frac{q}{p} = \frac{\text{the number of relevant predictors}}{\text{total number of predictors}}$$
Sparsity level in cancer study

Definition (Sparsity level in cancer study)

\[ \frac{q}{p} = s \]

the number of interesting genes

the number of protein-coding genes

Remark:
- Geneticists now like to discover more and more interesting genes; growing \( q \) movement.

A recent request and movement in gene-expression data study

We need a sparse method which works nicely not only under ultra-sparse regime (\( s \) is very small) but also under moderately-sparse regime (\( s \) is not too small). Our motivation of research is that in Bayesian context, this issue is closely related with tail of a prior.
List of additionally discovered genes relevant with breast cancer are ATM, BARD1, BRIP1, CDH1, CHEK2, MRE11A, MSH6, NBN, PALB2, PMS2, RAD50, RAD51C, STK11, and TP53. 
(Visit https://www.breastcancer.org/risk/factors/genetics.)
“Genetic variations can have large or small effects on the likelihood of developing a particular disease. ... Current research is focused on identifying genetic changes that have a small effect on disease risk but are common in the general population. Although each of these variations only slightly increases a person’s risk, having changes in several different genes may combine to increase disease risk significantly. Changes in many genes, each with a small effect, may underlie susceptibility to many common diseases, including cancer, obesity, diabetes, heart disease, and mental illness.”
Bayesian approaches to high-dimensional linear regression
Bayesian approaches to high-dimensional linear regression

Consider the high-dimensional setting with $p \gg n$

$$y = X\beta + \sigma \epsilon, \quad \epsilon \sim \mathcal{N}_n(0, I_n), \quad \text{and } \beta \text{ is sparse.}$$

Bayesian approaches

Sample from the posterior distribution:

$$\pi(\beta | y) \propto \mathcal{N}_n(y | X\beta, \sigma^2 I_n) \cdot \pi(\beta),$$

for some sparsity favoring prior $\pi(\beta)$.

(Typically, in Bayesian analysis, prior for the error variance $\sigma^2$ is given by the Jeffreys prior: this is not our concern.)

There are two types of priors for $\pi(\beta)$:

- Spike-and-slab priors
- Continuous shrinkage priors (main talk)
Spike-and-slab type priors

Historically, spike-and-slab priors have been considered as natural solutions.

**Spike-and-slab type priors (George & McCulloch, 1997)**

Each component of the $\beta = (\beta_1, \cdots, \beta_p)^T$ is assumed to be drawn from

$$
\beta_j|\tau, \phi \sim (1 - \tau) \cdot \delta_0(\beta_j) + \tau \cdot f_\phi(\beta_j), \quad (j = 1, \cdots, p),
$$

where $\tau = \text{Pr}(\beta_j \neq 0)$ and $\phi \sim r(\phi)$. $\delta_0$ is the Dirac-delta function and $f_\phi$ is a density on $\mathbb{R}$ with parameter $\phi$.

**A drawback:** computation is prohibitive.

![Spike-and-slab density](image)
Continuous shrinkage priors

Basic idea of continuous shrinkage priors is to mimic the spike-and-slab prior by a single continuous density.

Global-local-shrinkage priors

\[ \beta_j | \lambda_j, \tau, \sigma^2 \sim \mathcal{N}_1(0, \lambda_j^2 \tau^2 \sigma^2), \quad \sigma^2 \sim h(\sigma^2), \quad (j = 1, \ldots, p), \]
\[ \lambda_j \sim f(\lambda_j), \quad \tau \sim g(\tau), \quad (j = 1, \ldots, p), \]

where \( h, f, \) and \( g \) are densities supported on \((0, \infty)\).

- \( \lambda_j \): local-scale parameter associated with \( \beta_j \). Detect signal or noise with magnitude of \( \lambda_j \). Provide the heterogeneity for signal detection.
- \( \tau \): global-scale parameter. Shrinkage \( \beta \) globally towards \( 0 \).
Hierarchical formulation of the Horseshoe

\[ \beta_j | \lambda_j, \tau, \sigma^2 \sim \mathcal{N}_1(0, \lambda_j^2 \tau^2 \sigma^2), \quad \sigma^2 \sim \pi(\sigma^2) \propto 1/\sigma^2, \quad (j = 1, \ldots, p), \]
\[ \lambda_j \sim \mathcal{C}^+(0, 1), \quad \tau \sim \mathcal{C}^+(0, 1), \quad (j = 1, \ldots, p), \]

where \( \mathcal{C}^+(0, 1) \) represents the half-Cauchy density:

\[ \pi(x) = \mathcal{C}^+(x|0, 1) = \frac{1}{1 + x^2}, \quad x > 0. \]

- The Horseshoe has no hyper-parameter.
- Posterior computation is very fast: use R package horseshoe.
How to visualize the Horseshoe?

Start by fixing $\tau > 0$, and consider the univariate setup:

$$
\beta | \lambda \sim \mathcal{N}_1(0, \tau^2 \lambda^2) \quad \text{and} \quad \lambda \sim \mathcal{C}^+(0, 1).
$$

And then, integrate out $\lambda$, to get marginal density

$$
\pi(\beta | \tau) = \int_0^\infty \mathcal{N}_1(\beta | 0, \tau^2 \lambda^2) \cdot \mathcal{C}^+(\lambda | 0, 1) d\lambda \\
= K \cdot e^{Z(\beta)} \cdot E_1\{Z(\beta)\},
$$

where $K = 1/(\tau^{1/2} \pi^{3/2})$ and $Z(\beta) = \beta^2 / (2\tau^2)$. $E_1(x) = \int_1^\infty e^{-xt} t^{-1} dt$ is the exponential integral function.

A key idea of Horseshoe

Many nice properties of the Horseshoe come from $E_1\{Z(\beta)\}$. 
Properties of the Horseshoe

The Horseshoe is characterized by two nice properties:

1. **Infinite spike at origin**, i.e., \( \lim_{\beta \to 0} \pi(\beta | \tau) = \infty, \tau > 0. \)
2. **Heavy-tail** of \( \pi(\beta | \tau) \) for any \( \tau > 0. \)
Spike-and-slab type prior versus the Horseshoe

The Horseshoe

Se Yoon Lee, Debdeep Pati, Bani K. Mallick

Continuous shrinkage prior revisited : a collapsing behavior and remedy
Other nice properties of the Horseshoe

1. Under the sparsity assumption

\[ s = \frac{q}{p} \to 0 \quad \text{and} \quad n, p \to \infty \]

The Horseshoe possesses nice theoretical properties. (Bai & Ghosh, 2018; Polson, Scott & Scott, 2010; van der Pas, Szabo, & van der Vaart, 2017)

2. For e.g., the Horseshoe estimator (posterior mean for \( \beta \)) is robust and attains the minimax-optimal rate up to a constant. (Van Der Pas et al. 2014, 2016)
How does the tail the Horseshoe behave?

First, let's define the tail-index typically used in extreme value theory.

**Definition (Tail-index = 1/shape parameter)**

Suppose random variable $X$ is distributed with cdf $F$. Then the tail-index of the density, $f = F'$, is the value $\alpha > 0$, satisfying

$$\bar{F}(x) = L(x) \cdot x^{-\alpha},$$

where $\bar{F} = 1 - F$ is survival function and $L$ is a slowly varying function. **The reciprocal of tail-index, $\xi = 1/\alpha$, is called the shape parameter.**

Key concept in understanding tail-behavior of a density $f$

**tail-heaviness of $f$ $\uparrow \iff$ tail-index $\alpha \downarrow 0 \iff$ shape parameter $\xi \uparrow \infty$**
Drawback: Restricted tail behavior of the Horseshoe

Theorem (Fixed tail-index of the Horseshoe)

Assume $\beta | \lambda, \tau \sim \mathcal{N}_1(0, \tau^2 \lambda^2)$, $\lambda \sim \mathcal{C}^+(0, 1)$, and $\tau > 0$. Then the tail-index of $\pi(\beta | \tau)$ is fixed with $\alpha = 1$ for any value of $\tau > 0$.

A drawback of the Horseshoe

1. The Horseshoe is only able to control the center part.
2. The Horsehoe does NOT have a tail-index controlling mechanism.
Global-local-\textit{tail} shrinkage priors
Idealistic tail behavior of continuous shrinkage prior

**KEY IDEA** of Global-local-tail shrinkage priors

Idealistically, as the sparsity-level $s = q/p$ increases, the shape parameter $\xi = 1/\alpha$ of the marginal density of $\beta$ should also accordingly increase to accommodate more signals.

---

**Figure**: Tail-part of the marginal density of $\beta$
Idealistic tail behavior of continuous shrinkage prior

**KEY IDEA** of Global-local-tail shrinkage priors

Idealistically, as the sparsity-level $s = q/p$ increases, the shape parameter $\xi = 1/\alpha$ of the marginal density of $\beta$ should also accordingly increase to accommodate more signals.

*Figure:* Tail-part of the marginal density of $\beta$. Tail is lifted to accommodate more signals.
Global-local-tail shrinkage prior

We introduce a new framework for continuous shrinkage priors:

**Definition (Hierarchy of global-local-tail shrinkage prior)**

\[
\beta_j | \lambda_j, \sigma^2 \sim \mathcal{N}_1(0, \lambda_j^2 \sigma^2), \quad \sigma^2 \sim h(\sigma^2), \quad (j = 1, \cdots, p),
\]

\[
\lambda_j | \tau, \xi \sim f(\lambda_j | \tau, \xi), \quad (j = 1, \cdots, p),
\]

\[(\tau, \xi) \sim g(\tau, \xi),\]

where \(f\) is a density supported on \((0, \infty)\) with the scale parameter \(\tau > 0\) and the shape parameter \(\xi > 0\), and \(g\) is a joint density supported on \((0, \infty) \times (0, \infty)\). \(h\) is a density supported on \((0, \infty)\).

- The local-scale density \(f\) plays a key role.
- Fully Bayesian inference about \(\xi\) is challenging because \(\xi\) exists as an exponent part of \(f\).
Global-local-tail shrinkage prior

Table: Unit scaled densities for $f$

| Distribution                        | $f(\lambda|\tau = 1, \xi)$                                      | Shape $\xi$ |
|-------------------------------------|-----------------------------------------------------------------|-------------|
| Half-$\alpha$-stable distribution   | non-closed form                                                 | $\xi$       |
| **Half-Cauchy distribution**        | $2\{\pi(1 + \lambda^2)\}^{-1}$                                | 1           |
| Half-Levy distribution              | $\lambda^{-3/2}\exp\{-1/(2\lambda)\}/\sqrt{2\pi}$            | 2           |
| Loggamma distribution               | $\{(1 + \lambda)^{-(1/\xi+1)}\}/\xi$                          | $\xi$       |
| Generalized extreme value distribution | $\exp\{- (1+\xi \lambda)^{-1/\xi}\}/(1+\xi \lambda)^{(1/\xi+1)}$ | $\xi$       |
| Generalized Pareto distribution     | $(1 + \xi \lambda)^{-(1/\xi+1)}$                              | $\xi$       |

- If choosing **half-Cauchy distribution** for $f$, one can get the Horseshoe: the Horseshoe $\in$ the global-local-tail shrinkage priors.
- Fully Bayesian inference about $\xi$ is challenging because $\xi$ exists as an exponent part of $f$. 
The GLT prior

**KEY IDEA of The GLT prior**

We use the Generalized Pareto Distribution (GPD) instead of half-Cauchy distribution.

**Definition (Hierarchy of the GLT prior \( \pi_{GLT}(\beta) \))**

\[
\begin{align*}
\beta_j | \lambda_j, \sigma^2 &\sim \mathcal{N}_1(0, \lambda_j^2 \sigma^2), \quad \sigma^2 \sim \pi(\sigma^2) \propto 1/\sigma^2, \quad (j = 1, \cdots, p), \\
\lambda_j | \tau, \xi &\sim \text{GPD}(\tau, \xi), \quad (j = 1, \cdots, p), \\
\tau | \xi &\sim \text{IG}(p/\xi + 1, 1), \\
\xi &\sim \log \mathcal{N}(\mu, \rho^2) \cdot \mathcal{I}(1/2, \infty), \quad \mu \in \mathbb{R}, \quad \rho^2 > 0.
\end{align*}
\]

(\text{GPD}: \text{GPD}; \text{IG}: \text{inverse-gamma}; \text{log}\mathcal{N}: \text{log-normal})
The GLT prior

Graphical model representation of

\[ y \mid \beta, \sigma^2 \sim \mathcal{N}_n(X\beta, \sigma^2 I_n) \quad \text{and} \quad \beta \sim \pi_{\text{GLT}}(\beta). \]

- \( \mu \) and \( \rho^2 \): hyper-parameters.
- We developed the \textit{elliptical slice sampler centered by the Hill estimator} which enables
  1. to learn the shape parameter \( \xi \) according to the sparsity-level,
  2. and render the posterior computation \textit{tuning-free}. 
Tail behavior of the GLT prior

For a fixed $\tau > 0$ and $\xi > 1/2$, consider a univariate setup:

$$\beta | \lambda \sim \mathcal{N}_1(0, \lambda^2) \quad \text{and} \quad \lambda \sim \mathcal{GPD}(\tau, \xi).$$

And then, integrate out $\lambda$, to get marginal density

$$\pi(\beta | \tau, \xi) = \int_{0}^{\infty} \mathcal{N}_1(\beta | 0, \lambda^2) \cdot \mathcal{GPD}(\lambda | \tau, \xi) d\lambda$$

$$= \sum_{k=0}^{\infty} a_k \{ \psi^S_k(\beta) + \psi^R_k(\beta) \},$$

where $a_k = (-1)^k \cdot K \cdot \binom{1/\xi + k}{k}$, $K = 1/(\tau ^{3/2} \pi^{1/2})$, $Z(\beta) = \beta^2 \xi^2 / (2 \tau^2)$,

$$\psi^S_k(\beta) = E_{k/2+1}\{Z(\beta)\}, \quad \text{and} \quad \psi^R_k(\beta) = Z(\beta)^{-\frac{1+1/\xi + k}{2}} \gamma\{(1 + 1/\xi + k)/2, Z(\beta)\}.$$

$E_s(x) = \int_{1}^{\infty} e^{-xt} t^{-s} dt$, is the generalized exponential-integral function.

Key massage of this slide

- Existence of $\{\psi^R_k(\beta)\}_{k=0}^{\infty}$ provides a great flexibility to the shape of the density $\pi(\beta | \tau, \xi)$. We call the series of functions tail lifters.

- The Horseshoe does NOT have tail lifters.
The Horseshoe versus The GLT prior

Let’s summarize properties of two priors.

### Characteristics of the Horseshoe

1. **Infinite spike at origin**, i.e., \( \lim_{\beta \to 0} \pi(\beta | \tau) = \infty. \)
2. **Heavy-tail** of \( \pi(\beta | \tau) \). Tail-index is \( \alpha = 1 \) (shape parameter is \( \xi = 1 \)).
3. **[Drawback]** Restricted tail behavior of \( \pi(\beta | \tau) \). Fixed tail-index.

### Characteristics of the GLT prior

1. **Infinite spike at origin**, i.e., \( \lim_{\beta \to 0} \pi(\beta | \tau, \xi) = \infty. \)
2. **Heavy-tail** of \( \pi(\beta | \tau, \xi) \).
3. **Controllable tail behavior** of \( \pi(\beta | \tau, \xi) \) through controlling \( \xi \).
The Horseshoe versus The GLT prior \((\tau = 1)\)

- Comparison between \(\pi_{\text{HS}}(\beta|\tau)\) and \(\pi_{\text{GLT}}(\beta|\tau, \xi)\) when \(\tau = 1\).
- \(\pi_{\text{HS}}(\beta|\tau)\): black curve.
- \(\pi_{\text{GLT}}(\beta|\tau, \xi)\): colored in red \((\xi = 1)\), green \((\xi = 1.5)\), blue \((\xi = 2)\), and violet \((\xi = 3)\), respectively.
The Horseshoe versus The GLT prior ($\tau = 0.001$)

- Comparison between $\pi_{HS}(\beta|\tau)$ and $\pi_{GLT}(\beta|\tau, \xi)$ when $\tau = 0.001$.
  - $\pi_{HS}(\beta|\tau)$: black curve.
  - $\pi_{GLT}(\beta|\tau, \xi)$: colored in red ($\xi = 1$), green ($\xi = 1.5$), blue ($\xi = 2$), and violet ($\xi = 3$), respectively.
- When $\tau$ is very small, $\pi_{HS}(\beta|\tau)$ numerically becomes a Dirac-delta function.
Framework changes in high-dimensional statistical modelling (Bayesian & Continuous shrinkage)

\[ y = X\beta + \sigma \epsilon, \quad \epsilon \sim \mathcal{N}_n(0, I_n), \text{ and } \beta \text{ is sparse} \]

Global Shrinkage Priors

\[ \beta_j | \tau \sim \mathcal{N}_1(0, \tau^2 \sigma^2) \]
\[ \tau \sim g(\tau) \]

Global-Local Shrinkage Priors

\[ \beta_j | \lambda_j, \tau \sim \mathcal{N}_1(0, \lambda_j^2 \tau^2 \sigma^2) \]
\[ \lambda_j \sim f(\lambda_j) \]
\[ \tau \sim g(\tau) \]

Global-Local-Tail Shrinkage Priors

\[ \beta_j | \lambda_j \sim \mathcal{N}_1(0, \lambda_j^2 \sigma^2) \]
\[ \lambda_j | \tau, \xi \sim f(\lambda_j | \tau, \xi) \]
\[ (\tau, \xi) \sim g(\tau, \xi) \]
Application I: Breast cancer data study
Application I: Breast cancer data study

Consider the high-dimensional linear regression:

\[ y = X\beta + \sigma \varepsilon, \quad \varepsilon \sim \mathcal{N}_n(0, I_n). \]

We collected data from The Cancer Genome Atlas (TCGA):

- \((y, X) \in \mathbb{R}^n \times \mathbb{R}^{n \times p}: n = 729\) breast cancer patients, \(p = 3250\) genes.
- \(y_i\): logarithm of overall-survival time from the \(i\)-th patient.
- \(x_{ij}\): the gene-expression level of \(j\)-th gene from the \(i\)-th patient.
Application I: Breast cancer data study

To see the behavior of the Horseshoe as the number of genes used increase, we constructed four datasets, $\mathcal{B}_1 \subset \mathcal{B}_2 \subset \mathcal{B}_3 \subset \mathcal{B}_4$;

$$\mathcal{B}_1 = (y, X[:, 1:500]) \text{ (500 genes)},$$

$$\mathcal{B}_2 = (y, X[:, 1:1000]) \text{ (1,000 genes)},$$

$$\mathcal{B}_3 = (y, X[:, 1:2000]) \text{ (2,000 genes)},$$

and $$\mathcal{B}_4 = (y, X[:, 1:3250]) \text{ (3,250 genes; full data)}.$$

Guess on the sparsity level of the four datasets

Sparsity level of the datasets may increase as the number of genes increases. (Of course, sparsity level is unknown.)
Top 50 gene ranking plot (the Horseshoe)

Figure: $\mathcal{B}_1$ (top-left), $\mathcal{B}_2$ (top-right), $\mathcal{B}_3$ (bottom-left), and $\mathcal{B}_4$ (bottom-right).
**Top 10 interesting genes (the Horseshoe)**

**Table:** Top 10 interesting genes selected by the Horseshoe,

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{B}_1$</td>
<td>NGEF(−)</td>
<td>PLN(−)</td>
<td>C3orf59(+)</td>
<td>C21orf63(+</td>
<td>LOC100130331(−)</td>
</tr>
<tr>
<td>$\mathcal{B}_2$</td>
<td>FAM138F(−)</td>
<td>SLC39A4(−)</td>
<td>PLN(−)</td>
<td>NGEF(−)</td>
<td>PCGF5(+)</td>
</tr>
<tr>
<td>$\mathcal{B}_3$</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>$\mathcal{B}_4$</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{B}_1$</td>
<td>FCGR2A(−)</td>
<td>HES4(+</td>
<td>BCAP31(−)</td>
<td>GSTM1(+)</td>
<td>TOB2(−)</td>
</tr>
<tr>
<td>$\mathcal{B}_2$</td>
<td>HES4(+)</td>
<td>FCGR2A(−)</td>
<td>FCGR2C(−)</td>
<td>TOB2(−)</td>
<td>BCAP31(−)</td>
</tr>
<tr>
<td>$\mathcal{B}_3$</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>$\mathcal{B}_4$</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

**NOTE:** Contents of table is (gene name, direction). Gene with (+) and Gene with (−) may enhance and undermine immune system of breast cancer patients, respectively. When the horseshoe prior is applied to the datasets $\mathcal{B}_3$ and $\mathcal{B}_4$, genes are unranked because the horseshoe estimator collapsed.
Figure: $\mathcal{B}_1$ (top-left), $\mathcal{B}_2$ (top-right), $\mathcal{B}_3$ (bottom-left), and $\mathcal{B}_4$ (bottom-right).

Posterior means of $\xi$ are

$$2.188(\mathcal{B}_1) < 2.230(\mathcal{B}_2) < 2.382(\mathcal{B}_3) < 2.922(\mathcal{B}_4).$$
## Top 10 interesting genes (the GLT prior)

### Table: Top 10 interesting genes selected by the GLT prior

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B_1$</td>
<td>NGEF($-$)</td>
<td>C21orf63($+$)</td>
<td>PLN($-$)</td>
<td>C3orf59($+$)</td>
<td>FCGR2A($-$)</td>
</tr>
<tr>
<td>$B_2$</td>
<td>FAM138F($-$)</td>
<td>SLC39A4($-$)</td>
<td>NGEF($-$)</td>
<td>PCGF5($+$)</td>
<td>PLN($-$)</td>
</tr>
<tr>
<td>$B_3$</td>
<td>FAM138F($-$)</td>
<td>SLC39A4($-$)</td>
<td>NGEF($-$)</td>
<td>PLN($-$)</td>
<td>COL7A1($-$)</td>
</tr>
<tr>
<td>$B_4$</td>
<td>FAM138F($-$)</td>
<td>NSUN4($-$)</td>
<td>COL7A1($-$)</td>
<td>LOC150776($+$)</td>
<td>NGEF($-$)</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>$B_1$</td>
<td>BCAP31($-$)</td>
<td>GSTM1($+$)</td>
<td>LOC100130331($-$)</td>
<td>TOB2($-$)</td>
<td>ABCA17P($+$)</td>
</tr>
<tr>
<td>$B_2$</td>
<td>FCGR2A($-$)</td>
<td>CRHR1($+$)</td>
<td>TOB2($-$)</td>
<td>GSTM1($+$)</td>
<td>LOC150776($+$)</td>
</tr>
<tr>
<td>$B_3$</td>
<td>CRHR1($+$)</td>
<td>FCGR2A($-$)</td>
<td>RPLP1($+$)</td>
<td>HES4($+$)</td>
<td>TOB2($-$)</td>
</tr>
<tr>
<td>$B_4$</td>
<td>SMCHD1($+$)</td>
<td>RPLP1($+$)</td>
<td>HES4($+$)</td>
<td>SLC37A2($-$)</td>
<td>SLC39A4($-$)</td>
</tr>
</tbody>
</table>
Literature coherency with oncology/genetics for the selected genes via the GLT prior for the dataset $B_4$

<table>
<thead>
<tr>
<th>Rank</th>
<th>Gene (direction)</th>
<th>Note</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FAM138F(−)</td>
<td>Increasing a risk of breast and ovarian cancer</td>
<td>[11, 27]</td>
</tr>
<tr>
<td>2</td>
<td>NSUN4(−)</td>
<td>Related with ovarian and prostate cancer</td>
<td>[15]</td>
</tr>
<tr>
<td>3</td>
<td>COL7A1(−)</td>
<td>Related with cell migration (metastasis)</td>
<td>[33]</td>
</tr>
<tr>
<td>4</td>
<td>LOC150776(+)</td>
<td>Less studied in oncology and genetics</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>NGEF(−)</td>
<td>Related with obesity-related diseases</td>
<td>[31]</td>
</tr>
<tr>
<td>6</td>
<td>SMCHD1(+)</td>
<td>Important in regulation</td>
<td>[14]</td>
</tr>
<tr>
<td>7</td>
<td>RPLP1(+)</td>
<td>Important in protein synthesis</td>
<td>[9]</td>
</tr>
<tr>
<td>8</td>
<td>HES4(+)</td>
<td>Gene knockdown increases a brain disease</td>
<td>[1]</td>
</tr>
<tr>
<td>9</td>
<td>SLC37A2(−)</td>
<td>Negatively related with survival probability</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>SLC39A4(−)</td>
<td>Negatively related with survival probability</td>
<td>[14]</td>
</tr>
<tr>
<td>11</td>
<td>MFRP(−)</td>
<td>Related with ovarian cancer</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>ARSA(+)</td>
<td>Positively related with survival probability</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>BHLHE41(+)</td>
<td>High recovery from fatigue or short sleep</td>
<td>[24]</td>
</tr>
</tbody>
</table>

See our Supplemental materials for more detail.
Application II: Prostate cancer data study (Bradley Efron, 2008, 2010)
Prostate cancer data study (Bradley Efron, 2008, 2010)

- Dataset is based on two-sample test-statistics, \( \{y_j\}_{j=1}^P \) \( (p = 6,033 \text{ genes}) \): Data is available in R package sda. See Efron (2008) for detail.
- Histogram of \( \{y_j\}_{j=1}^P \).

Prostate cancer data study (Bradley Efron, 2008, 2010)

Consider a sparse normal mean model:

\[
\begin{bmatrix}
y_1 \\
\vdots \\
y_j \\
\vdots \\
y_p
\end{bmatrix}
= 
\begin{bmatrix}
\beta_1 \\
\vdots \\
\beta_j \\
\vdots \\
\beta_p
\end{bmatrix}
+ \sigma
\begin{bmatrix}
\epsilon_1 \\
\vdots \\
\epsilon_j \\
\vdots \\
\epsilon_p
\end{bmatrix},
\epsilon_j \sim i.i.d. \mathcal{N}_1(0, 1),
\]

where \( \beta = (\beta_1, \cdots, \beta_j, \cdots, \beta_p)^\top \in \mathbb{R}^p \) is assumed to be sparse. We will use the Horseshoe and GLT priors for \( \beta \in \mathbb{R}^p \) to compare.
Theoretically idealistic shrinkage

1. If $y_j$ is a noise $\Rightarrow \hat{\beta}_j$ should shrink. ($\hat{\beta}_j \approx 0$)

2. If $y_j$ is a signal $\Rightarrow \hat{\beta}_j$ should NOT shrink. ($\hat{\beta}_j \approx y_j$)
The Horseshoe applied to the prostate cancer data

- To investigate the behavior of the Horseshoe as the number of genes used increases, we constructed four datasets,
  \[ \mathcal{P}_1 = \{ y_j \}_{j=1}^{p=50}, \quad \mathcal{P}_2 = \{ y_j \}_{j=1}^{p=100}, \quad \mathcal{P}_3 = \{ y_j \}_{j=1}^{p=200}, \quad \mathcal{P}_4 = \{ y_j \}_{j=1}^{p=6033} \]
The GLT prior applied to the prostate cancer data

- We constructed seven datasets,
  \[ \mathcal{P}_1 = \{y_j\}_{j=1}^{p=50}, \mathcal{P}_2 = \{y_j\}_{j=1}^{p=100}, \mathcal{P}_3 = \{y_j\}_{j=1}^{p=200}, \mathcal{P}_4 = \{y_j\}_{j=1}^{p=500}, \mathcal{P}_5 = \{y_j\}_{j=1}^{p=1000}, \mathcal{P}_6 = \{y_j\}_{j=1}^{p=3000}, \text{ and } \mathcal{P}_7 = \{y_j\}_{j=1}^{p=6033} \]
The GLT prior applied to the prostate cancer data

Posterior means of $\xi$ are

\[
1.620(\mathcal{P}_1) < 1.662(\mathcal{P}_2) < 1.789(\mathcal{P}_3) \\
< 1.905(\mathcal{P}_4) < 1.991(\mathcal{P}_5) < 2.760(\mathcal{P}_6) < 3.636(\mathcal{P}_7).
\]

- As the number of genes considered increases, the posterior mean of $\xi$ also accordingly increases to accommodate more signals.
Simulated data study
Simulation Setting

Consider a simulation environment \((n, p, q, \text{SNR})\):

\[
y = X\beta_0 + \sigma_0 \epsilon, \quad \epsilon \sim \mathcal{N}_n(0, I_n)
\]

\[
\beta_0 = \begin{pmatrix}
\beta_{01}, \ldots, \beta_{0q}, \\
\beta_{0q+1}, \ldots, \beta_{0p}
\end{pmatrix}^\top
\]

\[
= \begin{pmatrix}
1, \ldots, 1, 0, \ldots, 0
\end{pmatrix}^\top,
\]

where \(\beta_0\) consists of \(q\) signals and \((p-q)\) noises.

Generation of an artificial Data \((y, X) \in \mathbb{R}^n \times \mathbb{R}^{n \times p}\)

**Step 1.** Generate row vectors \(\{x_i\}_{i=1}^n\) of design matrix \(X \in \mathbb{R}^{n \times p}\) from

\[
x_i \sim \mathcal{N}_p(0, I_p),
\]

and then column-wisely standardize the matrix so that each vector \(X[\cdot, j]\) has zero mean with unit \(l_2\)-norm.

**Step 2.** Generate \(n\)-dimensional Gaussian errors \(\epsilon \sim \mathcal{N}_n(0, I_n)\).

**Step 3.** Get \(y = X\beta_0 + \sigma_0 \epsilon\) where \(\sigma_0^2 = \text{var}(X\beta_0)/\{\text{SNR} \cdot \text{var}(\epsilon)\}\).
Simulated data study

- Fix \( n = 100, \ p = 500, \) and SNR=5.
- Generated four artificial datasets;
  1. Generate \( \mathcal{A}_1 = (y, X) \) by choosing \( q = 2 \); then, sparsity level is \( s = \frac{2}{500} = 0.004 \)
  2. Generate \( \mathcal{A}_2 = (y, X) \) by choosing \( q = 5 \); then, sparsity level is \( s = \frac{5}{500} = 0.01 \)
  3. Generate \( \mathcal{A}_3 = (y, X) \) by choosing \( q = 8 \); then, sparsity level is \( s = \frac{8}{500} = 0.016 \)
  4. Generate \( \mathcal{A}_4 = (y, X) \) by choosing \( q = 13 \); then, sparsity level is \( s = \frac{13}{500} = 0.026 \)
Simulated data study (the Horseshoe)

\[ A_1, \quad A_2, \quad A_3, \quad A_4 \]

Figure: green: truth \( \beta_0 \) / blue: inference for signals / red: inference for noises

Posterior means of \( \tau \) are

\[ 1.41 \cdot 10^{-6}(A_1) < 0.05(A_2) < 0.13(A_3) \gg 6.53 \cdot 10^{-15}(A_4) \]
Simulated data study (the GLT prior)

\[ A_1, \quad A_2, \quad A_3, \quad A_4. \]

**Figure:** green: truth \( \beta_0 \) / blue: inference for signals / red: inference for noises

Posterior means of \( \xi \) are

\[ 2.010(A_1) < 2.134(A_2) < 2.235(A_3) < 2.347(A_4). \]
Replicated Numerical Studies
Simulation setup

- Generate 50 replicated datasets under

\[
y = X\beta_0 + \sigma_0 \epsilon, \quad \epsilon \sim \mathcal{N}_n(0, I_n)
\]

\[
\beta_0 = (\beta_{01}, \ldots, \beta_{0q}, \beta_{0q+1}, \ldots, \beta_{0p})^\top
\]

\[
= (1, \ldots, 1, 0, \ldots, 0)^\top.
\]

- Change the sparsity-level $q/p$ from 0.001 to 0.1. (SNR is fixed with 5)

Performance metrics

Report the medians of mean squared errors (MSE) corresponding to signal and noise coefficients measured across the 50 replicates:

\[
\text{MSE}_S = \frac{1}{q} \sum_{j=1}^{q} (\hat{\beta}_j - 1)^2 \quad \text{and} \quad \text{MSE}_N = \frac{1}{p-q} \sum_{j=q+1}^{p} (\hat{\beta}_j)^2.
\]
Replicated Numerical Studies: varied sparsity level

(\(n, p\)) = (100, 500) (top) and (\(n, p\)) = (200, 1000) (bottom)

• and ▲: GLT / ■: Horseshoe
Summary

- Verified an absence of the tail-controlling mechanism of the Horseshoe.
- Introduced a new framework for continuous shrinkage priors, called the \textit{global-local-tail shrinkage priors}.
- Proposed the GLT prior which is a member of the new framework.
- Demonstrated the tail-adaptability of the GLT prior through two gene expression datasets & numerical studies.
More detail about the GLT prior

More details are included in our manuscript, for e.g.,

- Technical detail of the posterior computation.
- Automatic Hyper-parameter Tuning Technique: *elliptical slice Sampler centered by the Hill estimator*.
- Discussion on random shrinkage coefficients.
- Theoretical detail of the GLT prior.
- More about numerical studies.
- Curve fitting study via sparse kernel Gaussian regression.
Future work ("Big-data Bayesian inference")

- Mean Field Variational Bayes for the Horseshoe and GLT priors in high-dimensional setting. Currently, we are making a R package, called the VBhorseGLT.
- Application of the GLT prior for large-scale gene-association network.
Thank you very much!