Tail-adaptive Bayesian shrinkage

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, 3143 TAMU, College Station, TX, U.S.A.

January 14, 2021

Abstract

Modern genomic studies are increasingly focused on discovering more and more interesting genes associated with a health response. Traditional shrinkage priors are primarily designed to detect a handful of signals from tens and thousands of predictors. Under diverse sparsity regimes, the nature of signal detection is associated with a tail behaviour of a prior. A desirable tail behaviour is called tail-adaptive shrinkage property where tail-heaviness of a prior gets adaptively larger (or smaller) as a sparsity level increases (or decreases) to accommodate more (or less) signals. We propose a global-local-tail (GLT) Gaussian mixture distribution to ensure this property and provide accurate inference under diverse sparsity regimes. Incorporating a peaks-over-threshold method in extreme value theory, we develop an automated tail learning algorithm for the GLT prior. We compare the performance of the GLT prior to the Horseshoe in two gene expression datasets and numerical examples. Results suggest that varying tail rule is advantageous over fixed tail rule under diverse sparsity domains.

Keywords: Tail-adaptive shrinkage; The GLT prior; The Horseshoe; Extreme value theory.
1. INTRODUCTION

Development of sophisticated data acquisition techniques in gene expression microarray among many other fields triggered the development of innovative statistical methods [12, 26, 34] to identify relevant predictors associated with a response out of a large number of predictors, but only with a smaller number of samples. This ‘large $p$, small $n$’ paradigm is arguably the most researched topic in the last decade.

Consider a high-dimensional linear regression: we have data $(y_i, x_i)$, $i = 1, \ldots, n$, where $y_i$ and $x_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})^\top$ are response and $p$ covariates for the $i$-th subject, respectively, and assume $p \gg n$. Let $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ represent $n$-dimensional response vector and $n$-by-$p$ dimensional design matrix with $x_i$ as the rows, respectively, and perturb with a Gaussian error $\epsilon$ to formulate

$$y = X\beta + \sigma \epsilon, \quad \epsilon \sim \mathcal{N}_n(0, I_n). \tag{1}$$

A natural assumption accompanied with the formulation (1) is a sparsity on the $p$-dimensional coefficient vector $\beta = (\beta_1, \cdots, \beta_j, \cdots, \beta_p)^\top \in \mathbb{R}^p$, that is, substantially many coefficients in $\beta$ are assumed to be zeros or approximately zeros [64]. This is known to be a reasonable assumption in gene expression data studies where only a fraction of genes may affect a health response [22]. The true non-zero coefficients in the $\beta$ are referred to as signal coefficients, while the remaining are called noise coefficients.

Throughout the paper, we make use of the following definition of a sparsity level. Let $q$ denotes the number of true signals among the $p$ coefficients in the $\beta$. Sparsity level is defined as a ratio representing a proportion of the signals among the total coefficients,

$$s = \frac{q}{p} = \frac{\text{the number of relevant predictors}}{\text{total number of predictors}}. \tag{2}$$

In most real data applications, the sparsity level $s$ (2) is an unknown quantity because the truth of the $\beta$ is unknown.
Statisticians have devised a number of penalized regression techniques for estimating $\beta$ under the sparsity assumption [34]. From a Bayesian point view, sparsity favoring mixture priors with separate control on the signal and noise coefficients have been proposed [28, 38, 50, 75]. Although they often lead to attractive theoretical properties [16, 17], computational issues and considerations that many of the $\beta_j$’s may be small but not exactly zero has led to a rich variety of continuous shrinkage priors [13, 14, 32, 53, 65], which can be unified through a global-local scale mixture representation [57]. Among the continuous shrinkage priors, the Horseshoe [13, 14] is possibly the most visible and acclaimed method. It is important to point out that the posterior obtained using Horseshoe has remarkable finite sample performance and enjoys several optimal theoretical properties [68, 69, 70, 71] when the underlying sparsity level $s$ (2) is very small, so-called an ultra sparse regime.

In practical application of the regression problem (1), it is often necessary to consider and allow a moderately sparse regime where the sparsity level $s$ (2) is not too small, deviating from an ultra sparse regime, as exemplified in Section 2. In this article, we empirically showed that under a moderately sparse regime, the results of the posterior inference via the Horseshoe (and some of its variants) can be nullified; the posterior mean of the $\beta$, the Horseshoe estimator, may end up being approximately the null-vector, $\hat{\beta} \approx 0$. We refer to this undesirable phenomenon as a “collapsing behaviour”, typically caused by an underestimation of a global-scale parameter; a similar phenomenon is also observed by [4].

The objective of this article is to develop a new sparse-favoring prior that works reasonably well across diverse sparse domains. A key idea to achieve the goal is to make the posterior tail behaviour of a prior for the $\beta$ adaptive to the sparsity level $s$ (2). We refer to this as the *tail-adaptive shrinkage property*, characterized by the following: (i) under an ultra sparse regime (sparsity level $s$ (2) is very small), tail-heaviness of a sparse-favoring prior gets adaptively smaller to accommodate a small number of signals, a posteriori; and (ii) under a moderately sparse regime (sparsity level $s$ (2) is not too small), tail-heaviness of a sparse-favoring prior gets adaptively larger to accommodate a larger number of signals, a posteriori.

As there is no scientific threshold that bisects a sparse domain into ultra-sparse or moderate-
sparse domains [46], to achieve the tail-adaptive shrinkage property, a suggested prior should be endowed with an automated algorithm that produces inference results where tail-heaviness of the prior is adaptive to an unknown sparsity level, a posteriori, yet free from expert-tuning if having hyper-parameters. In this article, we suggest a new prior from, a global-local-tail mixture of Gaussian distribution, and propose its member, “the GLT prior” which enjoys the tail-adaptive shrinkage property. An automated algorithm is enabled by a modern Markov Chain Monte Carlo (MCMC) sampler combined with a peaks-over-threshold method from extreme value theory [24, 41].

The rest of the paper is organized as follows. In Section 2, we revisit a recent trend in gene expression data studies which reinforces the need to estimate signals under moderate sparse situation. In Section 3, we define the GLT prior as a member of global-local-tail shrinkage priors. In Section 4, we introduce an existing continuous shrinkage formulation that is a special case of the global-local-tail shrinkage formulation, and particularly look at the Horseshoe. In Section 6, we apply the GLT prior and the Horseshoe to an actual gene expression dataset, and in Section 7, the two priors are investigated through a simulation study. Replicated numerical studies are conducted in Section 8. Section 9 contains a summary and some discussion.

2. A RECENT TREND IN HIGH-THROUGHPUT GENE EXPRESSION ANALYSIS
In cancer genomic studies, the denominator $p$ in the sparsity level $s$ (2) typically represents the number of protein-coding genes. Protein-coding genes are fundamental to oncology because a cause for cancer is tied to mutated protein-coding genes [62, 63]. Although the number of protein-coding genes that can be analyzed in genomic research was very small decades ago, Human Genome Project enabled a construction of a massive human genome database, which allowed a large value for the $p$. One of the pressing questions in cancer-biology is how many protein-coding genes are encoded in the human genome [19], and a recent estimate of the count is $p = 21,306$ [74]. Nowadays, it seems that the acquisition of the gene expression data poses little issues. For a decade, The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov/) collected clinicopathologic annotation data along with multi-platform molecular profiles of more than 11,000 human tumors across 33 different cancer types [43, 72].
There is a recent trend in the nominator $q$ in (2), the number of interesting genes. (Here, the meaning of ‘interesting’ depends on the context.) Practitioners now like to detect interesting genes whose effects on cancer are diverse. For example, geneticists discovered that BRCA1 and BRCA2 are linked to a breast cancer risk over 20 years ago: however, beyond BRCA1 & 2 movement has motivated to discover more interesting genes because bio-industry began to realize that the two BRCA genes do not tell the whole tale in increasing the risk of breast cancer [52, 60]. Recently, genes named, ATM, BARD1, BRIP1, CDH1, CHEK2, MRE11A, MSH6, NBN, PALB2, PMS2, RAD50, RAD51C, STK11, and TP53, were discovered as additional risk factors for the breast cancer [25, 62]. Furthermore, it is known that these genes have their unique physiological functionalities, possibly related to other cancers, and considering that the understanding of cancer at the molecular level has a relatively short history, more interesting genes will be discovered in the future [7]. This growing $q$ movement is also aligned with a research outline provided by the National Institutes of Health of the United States; refer to the page 58 from 2019-year guideline (https://ghr.nlm.nih.gov/primer).

To summarize, it is pertinent to develop a high-throughput gene expression techniques reflecting the recent trend of growing $q$ movement. Development of a sparse estimation method for the high-dimensional regression (1) that can detect signals under various sparsity regimes is crucial in cancer genomic studies to better understand diseases.

3. VARYING TAIL RULE – GLOBAL-LOCAL-TAIL SHRINKAGE PRIORS

3.1 Definition of tail-heaviness of a density

Throughout the paper, the notion of the tail-heaviness of a Lebesgue measurable function (hence, distribution function, prior, or posterior distributions, etc) is adopted from the extreme value theory and regular variation [24, 30, 40, 41, 44, 48].

Definition 1. A positive, Lebesgue measurable function $\rho$ on $(0, \infty)$ is regularly varying of index $\alpha \in \mathbb{R}$ if there exists $\alpha$ such that $\lim_{x \to \infty} \rho(cx)/\rho(x) = c^{-\alpha}$, for any $c > 0$. If $\alpha = 0$, then the function $\rho$ is said to be slowly varying.
The Karamata’s characterization theorem states that every regularly varying function $\rho$ of index $\alpha$ has a representation $\rho(x) = L(x) \cdot x^{-\alpha}$ where $L$ is a slowly varying function [40].

Consider a positive random variable $X \sim F$ where $F$ is the distribution function of $X$. To adapt the regular variation theory to the extreme value theory, the measurable function $\rho$ in the Definition 1 is replaced by the tail (survival) function of the $X$, that is, $\bar{F} = 1 - F$, leading to an equation $\lim_{x \to \infty} \bar{F}(cx)/\bar{F}(x) = c^{-\alpha}$, for any $c > 0$. By the Karamata’s characterization theorem, this also leads to $\bar{F}(x) = L(x) \cdot x^{-\alpha}$ where $L$ is a slowly varying function: see that $\alpha$ represents the rate of decay at infinity. In extreme value theory, the value $\alpha$ satisfying the equation is an essential quantity representing a tail-heaviness of the random variable $X$. The $\alpha$ is referred to as the tail-index of the random variable $X$ or the tail-index of the density $f = F'$, and its reciprocal $\xi = 1/\alpha$ is called the shape parameter [20, 24]. A distribution $F$ with positive $\xi > 0$ is called a heavy-tailed distribution (see page 268 of [47]). As value of $\xi$ increases, the tail-heaviness of the density $f$ accordingly increases. Although we illustrated the notion of tail-index via a positive random variable, this notion can be generalized to a real random variable in a similar fashion by measuring the tail-index at either $\infty$ or $-\infty$.

3.2 Global-local-tail shrinkage priors

We propose a new hierarchical formulation of continuous shrinkage priors, called the “global-local-tail shrinkage priors”, under the high-dimensional regression (1). As the name suggests, they can be represented as a global-local-tail Gaussian mixture distribution

$$\beta_{j}\mid \lambda_{j}, \sigma^2 \sim \mathcal{N}_{1}(0, \lambda_{j}^2 \sigma^2), \quad \sigma^2 \sim h(\sigma^2), \quad (j = 1, \cdots, p),$$

(3)

$$\lambda_{j}\mid \tau, \xi \sim f(\lambda_{j}\mid \tau, \xi), \quad (j = 1, \cdots, p),$$

(4)

$$(\tau, \xi) \sim g(\tau, \xi),$$

(5)

where $f$ is a density supported on $(0, \infty)$ with the scale parameter $\tau > 0$ and the shape parameter $\xi > 0$. (Then, by Karamata’s characterization theorem, the distribution function of $\lambda$, that is, $F(\lambda\mid \tau, \xi) = \int_{0}^{\lambda} f(t\mid \tau, \xi)dt$ for $\lambda > 0$, can be characterized by the equation $\bar{F}(\lambda\mid \tau, \xi) = \int_{0}^{\lambda} f(t\mid \tau, \xi)dt$ for $\lambda > 0$, can be characterized by the equation $\bar{F}(\lambda\mid \tau, \xi) =$
\[ 1 - F(\lambda|\tau, \xi) = L(\lambda) \cdot \lambda^{-1/\xi} \] where \( L \) is a slowly varying function.) The \( h \) is a density supported on \((0, \infty)\), while \( g \) is a joint density supported on \((0, \infty) \times (0, \infty)\). Following the literature of continuous shrinkage prior [9] and extreme value theory [24], the scale parameters \( \lambda_j \) (\( j = 1, \cdots, p \)) and \( \tau \) are referred to as the local-scale parameters and global-scale parameter, respectively, and \( \xi \) is called the shape parameter. Throughout of paper, we shall use the Jeffreys prior [36] for the measurement error \( \sigma^2 \), that is, \( h(\sigma^2) \propto \frac{1}{\sigma^2} \).

Table 1 lists some candidates of \( f \) with the unit scale \( \tau = 1 \). All distributions in the table are supported on \((0, \infty)\) with a positive shape parameter \( \xi > 0 \). The half-Cauchy and half-Levy distributions are derived from the half-\( \alpha \)-stable distribution with the tail-index \( \alpha \) by fixing the \( \alpha \) to be 1 and \( 1/2 \), respectively. More examples for \( f \) can be found in [24, 35].

| \( f(\lambda|\tau = 1, \xi) \) | Shape parameter \( \xi \) |
|-----------------------------|---------------------|
| 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 | \( (1 + \xi\lambda)^{-(1/\xi+1)} \) | \( \xi \) |

One of the practical bottlenecks in employing the new prior form (3) – (5) is a fully Bayesian estimation for the \( \xi \) [2]. Analytically, this is because the \( \xi \) participates in the local-scale density \( f \) which is heavy-tailed (since \( \xi > 0 \)) through an exponent. In the perspective of Bayesian learning, this is because the \( \xi \) is farthest from the data \( y \), hence, uncertainty (information) propagated from the \( y \) to the \( \xi \) can be attenuated when passing through the hierarchy, thereby requiring a special care in choosing a prior for the \( \xi \) and a nice sampling technique.

3.3 The GLT prior

We suggest a member of the global-local-tail form (3) – (5). We use the generalized Pareto distribution (GPD) [55], \( f(\lambda_j) = GDP(\lambda_j|\tau, \xi) = (1/\tau) \cdot (1 + \xi\lambda_j/\tau)^{-(1/\xi+1)} \) for the \( p \) local-scale parameters \( \lambda_j \) (\( j = 1, \cdots, p \)) in (4), and a truncated inverse-gamma-lognormal joint density for \( (\tau, \xi) \) in (5), \( g(\tau, \xi) = IG(\tau|p/\xi + 1, 1)I_{(0,\infty)}(\tau) \cdot \{\log N(\xi|\mu, \rho^2)I_{(1/2,\infty)}(\xi)\}/D, \) where
$D = D(\mu, \rho^2) = \int_{1/2}^{\infty} \log \mathcal{N}(\xi|\mu, \rho^2) d\xi$ is the normaliser of $g(\tau, \xi)$:

$$
\beta_j | \lambda_j, \sigma^2 \sim \mathcal{N}_1(0, \lambda^2_j \sigma^2), \quad \sigma^2 \sim \pi(\sigma^2) \propto 1/\sigma^2, \quad (j = 1, \cdots, p),
$$

$$
\lambda_j | \tau, \xi \sim \mathcal{GPD}(\tau, \xi), \quad (j = 1, \cdots, p),
$$

$$
\tau | \xi \sim \mathcal{IG}(p/\xi + 1, 1),
$$

$$
\xi \sim \log \mathcal{N}(\mu, \rho^2)|_{(1/2, \infty)}, \quad \mu \in \mathbb{R}, \quad \rho^2 > 0.
$$

Figure 1: DAG representation of $y | \beta, \sigma^2 \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$ and $\beta \sim \pi_{\text{GLT}}(\beta)$.

We call this specific hierarchical form (6) – (9) “the GLT prior”, denoted as $\beta \sim \pi_{\text{GLT}}(\beta)$. A directed asymmetric graphical (DAG) representation of a Bayesian linear model, $y | \beta, \sigma^2 \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$, $\pi(\sigma^2) \propto 1/\sigma^2$, and $\beta \sim \pi_{\text{GLT}}(\beta)$, is shown in Figure 1. Note that the $\mu$ and $\rho^2$ in (9) are hyper-parameters. Full description of posterior computation is provided in Appendix in Supplemental Material: the proposed sampling algorithm is a Gibbs sampler [15] endowed with an automatic hyper-parameter tuning enabled by a joint technique of elliptical slice sampler [51] and Hill estimator [35].

Some fundamental properties of the GLT prior (6) – (9) are investigated in Section 5. Among them, one of the notable characteristics is that the GLT prior asymptotically behaves like a two-group prior [27, 38, 39]; refer to Proposition 3 and Corollary 4. As indicated in (8), the GLT prior depends on the number of covariates $p$, and the prior mean of the $\tau$ given $\xi$ is $\mathbb{E}[\tau | \xi] = \xi/p$, which resembles the sparsity level (2), where only the nominator $q$ is replaced by the shape $\xi > 0$.

Main motivation of adopting the log-normal distribution $\log \mathcal{N}(x|\mu, \rho^2)$ as a prior for the shape parameter $\xi$ (9) is an apprehension that information from the data $y$ can be attenuated when passing
through the hierarchy and reaching to the $\xi$ located at the deepest level. To properly learn such a weaken information, we choose to use a sub-exponential density \cite{33} to induce a reasonably strong posterior concentration in learning the $\xi$; refer to Lemma 6. The reason for the truncation of the support of the log-normal prior by $(1/2, \infty)$ is to produce a “horseshoe” shape for the density of random shrinkage coefficient \cite{14}; refer to Proposition 5.

4. FIXED TAIL RULE– GLOBAL-LOCAL SHRINKAGE PRIORS

4.1 Global-local shrinkage priors

Most of the continuous shrinkage priors proposed and studied in the literature can be represented as global-local scale mixtures of Gaussian distribution:

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

where $f$, $g$, and $h$ are densities supported on $(0, \infty)$. Different choices of $f$ and $g$ for the top-level scale parameters lead to different class of priors \cite{9}. In the high-dimensional setting, the choices of $f$ and $g$ play a key role in controlling the effective sparsity and concentration of the prior and posterior distributions \cite{3, 45, 54, 57, 61, 77}.

It is important to note that the global-local hierarchy (10) – (11) can be regarded as a special case of the global-local-tail hierarchy (3) – (5) with a fixed shape parameter $\xi$. To see this, first, (a) choose a local-scale density $f(\cdot) = f(\cdot | \tau, \xi)$ in (4), and (b) fix the shape parameter $\xi$ with some positive value, and finally (c) bring up the scale parameter $\tau$ under the coefficient $\beta_j$ via re-parameterizing $\lambda_j/\tau$ for each $j$. Therefore, members of the global-local form (10) – (11) \cite{9, 14, 53} can be thought as members of the global-local-tail form (3) – (5) with the $\xi$ to be non-stochastic.

Although both prior formulations commonly belong to one-group continuous shrinkage prior formulation \cite{2, 11, 29, 31, 53}, they are based on different tail rules: \textit{varying tail rule} versus \textit{fixed tail rule}. A prior with varying tail rule may retain a great flexibility in the shape of the marginal density for the coefficients $\beta$. We exemplify the argument via the GLT prior; refer the Figure 2.
It is the shape parameter $\xi$ that tunnels a path to control a tail behaviour of a prior adaptive to the sparsity level (2): a prior with fixed tail rule abandons this tail controlling mechanism.

4.2 The Horseshoe

The Horseshoe [14] can be obtained by choosing the unit-scaled half-Cauchy densities, $C^+(x|0,1) = 2/\{\pi(1 + x^2)\}$, $x > 0$, for the $f$ and $g$ in (11) under the global-local form (10) – (11):

$$
\beta_j | \lambda_j, \tau, \sigma^2 \sim N_1(0, \lambda_j^2 \tau^2 \sigma^2), \quad \sigma^2 \sim \pi(\sigma^2) \propto 1/\sigma^2, \quad (j = 1, \cdots, p),
$$

(12)

$$
\lambda_j \sim C^+(0,1), \quad \tau \sim C^+(0,1), \quad (j = 1, \cdots, p).
$$

(13)

An equivalent hierarchy of the Horseshoe (12) – (13) can be derived from the global-local-tail form (3) – (5) as follows. First, (a) choose the half-$\alpha$-stable density for the local-scale density $f$ with the scale $\tau$ distributed according to $C^+(0,1)$ and shape $\xi$ in (4), and (b) fix the shape parameter to be $\xi = 1$ so that the $f$ becomes the half-Cauchy density scaled by $\tau$. Finally, (c) bring up the $\tau$ under the coefficient $\beta_j$ by re-parameterizing $\lambda_j/\tau$ for each $j$.

Among the continuous shrinkage priors, the Horseshoe (12) – (13) [14] is possibly the most studied member in the recent literature. Under the sparsity assumption $s \to 0$ as $n, p \to \infty$, it is known that the posterior mean of Horseshoe, the Horseshoe estimator, possesses many nice theoretical properties [3, 9, 57, 61, 70]. For instance, the Horseshoe estimator is robust and attains the minimax-optimal rate for squared error loss up to a multiplicative constant under certain conditions [67, 71]. Highly scalable algorithms are recently proposed for the Horseshoe [10, 37].

4.3 Restricted tail-heaviness of the Horseshoe

In the following, we show that the tail-heaviness of the Horseshoe (12) – (13) is fixed. For simplicity, we consider a univariate form of the Horseshoe, given as $\beta | \lambda, \tau \sim N_1(0, \lambda^2 \tau^2)$, and $\lambda \sim C^+(0,1)$ with fixed $\tau > 0$, and measure the tail-index $\alpha$ of the marginal density of the Horseshoe conditioned on $\tau$, $\pi_{HS}(\beta|\tau) = \int N_1(\beta|0, \lambda^2 \tau^2)C^+(\lambda|0,1)d\lambda$, $\beta \in \mathbb{R}, \tau > 0$. (Refer to (S.1) in Supplemental Material for the closed form expression of the density $\pi_{HS}(\beta|\tau)$.)
Proposition 2. Assume \( \beta | \lambda, \tau \sim N_1(0, \tau^2 \lambda^2) \), \( \lambda \sim C^+(0, 1) \), and \( \tau > 0 \). Then the tail-index of \( \pi_{HS}(\beta | \tau) \) is \( \alpha = 1 \) for any \( \tau > 0 \).

Proposition 2 is proved in Subsection S.4.2 in Supplemental Material. In general, it is known that the shape parameter of half-Cauchy density is \( \xi = 1 \) [24]. Proposition 2 implies that the tail-heaviness of the marginal density \( \pi_{HS}(\beta | \tau) \) inherits that of the local-scale density \( f(\lambda) = C^+(\lambda |0, 1) \) (13), and is fixed for any choice of the global-scale parameter \( \tau > 0 \). Although the density \( \pi_{HS}(\beta | \tau) \) is heavy-tailed due to the positive \( \xi = 1 > 0 \), the fixed tail-heaviness suggests an absence of the tail-controlling mechanism in accordance with the sparsity level, which is also pointed out by [56].

The deficiency of the tail-controlling mechanism may be troublesome in dealing with various sparsity regimes. Horseshoe is designed to perform well in an ultra sparse regime where the value \( \xi = 1 \) is sufficiently large to put an enough mass on the tail region of density \( \pi_{HS}(\beta | \tau) \). However, as the sparsity level \( s (2) \) increases, the need of placing more mass in the tail region may also increase to accommodate growing number of signals: in this case, the \( \xi = 1 \) may not be a sufficiently large value to allow an enough mass on the tail region, which can provoke anomalous inference results such as the collapsing behaviour. In Section 7 and 8, we will observe the collapsing behaviour of the Horseshoe estimator when sparsity level \( s (2) \) is moderately large.

5. PROPERTIES OF THE GLT PRIOR

5.1 Marginal density of the GLT prior

For simplicity we shall work with a univariate form of the GLT prior (6) – (9), given by \( \beta | \lambda \sim N_1(0, \lambda^2) \), and \( \lambda | \tau, \xi \sim GPD(\tau, \xi) \), with fixed \( \tau > 0 \) and \( \xi > 1/2 \). We investigate the marginal densities of the coefficient \( \beta \) and random shrinkage coefficient [14] (\( \kappa = 1/(1 + \lambda^2) \)), conditioned on the \((\tau, \xi)\); the marginal density of the \( \beta \) is \( \pi(\beta | \tau, \xi) = \int N_1(\beta |0, \lambda^2)GPD(\lambda |\tau, \xi) d\lambda \).

Proposition 3. Suppose \( \beta | \lambda \sim N_1(0, \lambda^2) \), \( \lambda \sim GPD(\tau, \xi) \), \( \tau > 0 \) and \( \xi > 1/2 \). Then:
(a) density of $\beta$ given $\tau$ and $\xi$ is

$$\pi(\beta|\tau, \xi) = \sum_{k=0}^{\infty} a_k \{\psi^S_k(\beta) + \psi^R_k(\beta)\}, \quad (14)$$

where $K = 1/(\tau^2 3/2 \pi^{1/2})$, $Z(\beta) = \frac{1}{\tau^2} \left(1 - \frac{\xi}{\tau^2}\right)^{1/\xi + k}$, and $\psi^S_k(\beta) = E_{k/2+1}(Z(\beta))$, and $\psi^R_k(\beta) = E_{k/2+1}(Z(\beta))^{-1 + \xi + k} \gamma((1 + 1/\xi + k)/2, Z(\beta))$. The superscripts on $\psi^S_k$ and $\psi^R_k$ represent “(noise) shrinkage” and “(tail) robustness”, respectively.

(b) density of $\kappa = 1/(1 + \lambda^2)$ given $\tau$ and $\xi$ is

$$\pi(\kappa|\tau, \xi) = \frac{\tau^{1/\xi}}{2} \cdot \frac{\kappa^{1/(2\xi) - 1} (1 - \kappa)^{-1/2}}{\{\tau \kappa^{1/2} + \xi (1 - \kappa)^{1/2}\}^{(1+1/\xi)}}. \quad (15)$$

Proposition 3 is proved in Subsection S.5.1 in Supplemental Material. Note that the marginal density $\pi(\beta|\tau, \xi)$ (14) is analytically expressed as an alternating series whose summands are separated into two terms: $\{\psi^S_k(\beta)\}_{k=0}^{\infty}$ and $\{\psi^R_k(\beta)\}_{k=0}^{\infty}$. There are two special functions participate in the density: (i) the generalized exponential-integral function of real order [18, 49] $E_s(x) = \int_1^{\infty} e^{-xt} t^{-s} dt$ ($x > 0, s \in \mathbb{R}$), and (ii) the incomplete lower gamma function $\gamma(s, x) = \int_0^x t^{s-1} e^{-t} dt$ ($s, x \in \mathbb{R}$). These special functions $E_s(x)$ and $\gamma(s, x)$ participate to the $\pi(\beta|\tau, \xi)$ through the sequences of functions $\{\psi^S_k(\beta)\}_{k=0}^{\infty}$ and $\{\psi^R_k(\beta)\}_{k=0}^{\infty}$, respectively. The generalized binomial coefficient $\binom{1/\xi + k}{k}$ is defined by $(1/\xi + k)(1/\xi + k - 1) \cdots (1/\xi + 1)/k!$ if $k \in \{1, 2, \ldots\}$, and zero if $k = 0$.

Analytically, the marginal density of the Horseshoe $\pi_{\text{HS}}(\beta|\tau)$ discussed in Subsection 4.3 is participated by one special function $E_1(x)$, namely the exponential integral function, while the Horseshoe has no functional component induced from the incomplete lower gamma function $\gamma(s, x)$ which the GLT prior possesses.

Figure 2 displays the marginal densities of the univariate coefficient $\beta$ obtained from the Horseshoe $\pi_{\text{HS}}(\beta|\tau)$ ($\tau > 0$) (S.1) and the GLT prior $\pi(\beta|\tau, \xi)$ ($\tau > 0, \xi > 1/2$) (14) for different values of $\tau$ and $\xi$. The tail-heaviness of the GLT prior (varying tail rule) gets thicker as the shape $\xi$.
increases, while that of the Horseshoe (fixed tail rule) is fixed.

![Graph showing comparison between two densities](image)

Figure 2: Comparison between two densities, $\pi_{HS}(\beta|\tau)$ ((S.1) in the Supplemental Material) and $\pi(\beta|\tau,\xi)$ ((a) in Proposition 3): $\tau = 1$ (top panels) and $\tau = 0.001$ (bottom panels). The density $\pi_{HS}(\beta|\tau)$ is colored in black, and densities $\pi(\beta|\tau,\xi)$ are colored in red ($\xi = 1$), green ($\xi = 1.5$), blue ($\xi = 2$), and violet ($\xi = 3$), respectively.

Although the GLT prior is a one-group prior [29], the following Corollary 4 suggests that the GLT prior asymptotically behaves like a two-group prior (so-called the “spike-and-slab prior” [27, 38, 39]) due to the two sequences of functions, $\{\psi^S_k(\beta)\}_{k=0}^{\infty}$ and $\{\psi^R_k(\beta)\}_{k=0}^{\infty}$. Roughly speaking, the roles of $\{\psi^S_k(\beta)\}_{k=0}^{\infty}$ and $\{\psi^R_k(\beta)\}_{k=0}^{\infty}$ are similar to those of ‘spike’ and ‘slab’ distributions of a two-group prior in asymptotic senses $|\beta| \to 0$ and $|\beta| \to \infty$, respectively:

**Corollary 4.** Suppose $\beta|\lambda \sim N_1(0, \lambda^2)$, $\lambda \sim GP\mathcal{D}(\tau, \xi)$, $\tau > 0$, and $\xi > 1/2$. Let $k \in \{0\} \cup \{1, 2, \ldots\}$. Then:

(a) If $k = 0$, then $\lim_{|\beta| \to 0} \psi^S_k(\beta) = \infty$; if $k \in \{1, 2, \ldots\}$, then $\lim_{|\beta| \to 0} \pi^S_k(\beta) = 2/k < \infty$.

(b) If $k \in \{0\} \cup \{1, 2, \ldots\}$, then $\lim_{|\beta| \to \infty} \psi^S_k(\beta) = 0$ with squared exponential rate.
(c) If \( k \in \{0\} \cup \{1, 2, \ldots \} \), then \( \lim_{|\beta| \to 0} \psi^R_k(\beta) = 2/(1 + 1/\xi + k) < \infty \).

(d) If \( k \in \{0\} \cup \{1, 2, \ldots \} \), then \( \psi^R_k(\beta) \) is regularly varying with index \( 1 + 1/\xi + k \).

Corollary 4 is proved in Subsection S.5.2 in Supplemental Material. Interpretations of the Corollary 4 are as follows. (a) implies that the marginal density \( \pi(\beta|\tau, \xi) \) (14) retains the infinite spike at origin for any \( \tau > 0, \xi > 1/2 \), as seen in the Figure 2, which is a common feature of the Horseshoe [14]. Technically, this infinite spike is caused by the exponential integral function \( E_1(x) \) (\( \lim_{x \to 0^+} E_1(x) = \infty \) [18]), allowing a very strong pulling of the \( \beta \) towards zero. By (a) and (c) of Corollary 4, it holds \( \lim_{|\beta| \to 0} \pi^S_k(\beta) = 2/k > \lim_{|\beta| \to 0} \psi^R_k(\beta) = 2/(1 + 1/\xi + k) k \in \{1, 2, \ldots \} \), which implies that the contribution of \( \{\pi^S_k(\beta)\}_{k=0}^\infty \) is larger than that of \( \{\psi^R_k(\beta)\}_{k=0}^\infty \) in shrinking the \( \beta \) towards zero. By (b), the squared exponential decay rates of the functions in \( \{\pi^S_k(\beta)\}_{k=0}^\infty \) as \( |\beta| \to \infty \) indicates that the contribution of \( \{\pi^S_k(\beta)\}_{k=0}^\infty \) in controlling the tail region of the density \( \pi(\beta|\tau, \xi) \) gets negligible as \( |\beta| \) goes to infinity. Finally, (d) implies the density \( \pi(\beta|\tau, \xi) \) possesses a systematic mechanism to control the tail region by controlling the \( \xi \) via the sequence of functions \( \{\psi^R_k(\beta)\}_{k=0}^\infty \).

We call the sequence of functions \( \{\psi^R_k(\beta)\}_{k=0}^\infty \) ‘tail lifters’ as their main roles are to lift the tail part of the density \( \pi(\beta|\tau, \xi) \) (14) by increasing the value of \( \xi \). The presence of tail lifters in the marginal prior \( \pi(\beta|\tau, \xi) \) provides a great flexibility to the shape of the density as shown in the panels in Figure 2. This may be particularly useful to handle various sparsity regimes.

In contrast, the marginal density of the Horseshoe \( \pi_{HS}(\beta|\tau) \) (S.1) does not have such a tail controlling mechanism; refer to Corollary 2. This is particularly problematic when \( \tau \) is estimated to be very small (say \( \tau = 0.001 \)). Panels in the second row in Figure 2 show a mismatch between the theoretical support \( \mathbb{R} \) and numerical support \( (-\epsilon, \epsilon) \), \( \epsilon \approx 0 \) of the density \( \pi_{HS}(\beta|\tau = 0.001) \) (1). If \( \tau \) is extremely small, say, \( \tau = 10^{-10} \), then the density \( \pi_{HS}(\beta|\tau = 10^{-10}) \) numerically degrades into the Dirac-delta function, possibly causing the collapsing behaviour.

Analytic characteristics of the density for the random shrinkage coefficient \( \kappa \) (15) are:

**Proposition 5.** Suppose \( \lambda \sim \mathcal{GPD}(\tau, \xi), \kappa = 1/(1 + \lambda^2) \in (0, 1), \tau > 0 \) and \( \xi > 1/2 \). Then:

(a) \( \lim_{\kappa \to 1^-} \pi(\kappa|\tau, \xi) = \infty \) and \( \lim_{\kappa \to 0^+} \pi(\kappa|\tau, \xi) = \infty \).
(b) \( \pi(\kappa|\tau = 1, \xi = 1) = \frac{\kappa^{-1/2}(1 - \kappa)^{-1/2}}{[2 \cdot \{\kappa^{1/2} + (1 - \kappa)^{1/2}\}]^2} \).

Probabilities of the regions \((1 - \epsilon, 1)\) and \((0, \epsilon)\), \(\epsilon \approx 0\) under the density \(\pi(\kappa|\tau, \xi)\) (15) are related with the shrinkage and the robustness, respectively [14]. The infinite spikes of \(\pi(\kappa|\tau, \xi)\) at \(k = 0\) and \(k = 1\) imply that the GLT prior has the desired shrinkage property. The density \(\pi(\kappa|\tau = 1, \xi = 1)\) is not a standardly known distribution, but resembles a ‘horseshoe’.

Figure 3: Comparison between two densities of random shrinkage coefficient \(\kappa\), \(\pi_{HS}(\kappa|\tau)\) (refer to (S.2) in Supplemental Material) and \(\pi(\kappa|\tau, \xi)\) ((b) in Proposition 3): \(\tau = 1\) (top panels) and \(\tau = 0.001\) (bottom panels). The density \(\pi_{HS}(\kappa|\tau)\) is colored in black, and densities \(\pi(\kappa|\tau, \xi)\) are colored in red (\(\xi = 1\)), green (\(\xi = 1.5\)), blue (\(\xi = 2\)), and violet (\(\xi = 3\)), respectively.

Figure 3 compares the densities of random shrinkage coefficient from the Horseshoe and GLT prior, \(\pi_{HS}(\kappa|\tau)\) (refer to (S.2) in Supplemental Material) and \(\pi(\kappa|\tau, \xi)\) (15), with different values of \(\tau\) and \(\xi\). When \(\tau = 1\), the top panels demonstrate Horseshoe-like shapes for both \(\pi_{HS}(\kappa|\tau = 1)\) and \(\pi(\kappa|\tau = 1, \xi)\). However, when \(\tau = 0.001\), the apparent difference is shown on the bottom-middle panel, where \(\pi_{HS}(\kappa|\tau = 0.001)\) places essentially zero-mass on \((0, \epsilon)\), \(\epsilon \approx 0\). This implies that the robustness property of the Horseshoe can be deteriorated when \(\tau\) is very small. On the other hand, the GLT prior \(\pi(\kappa|\tau = 0.001, \xi)\) still places a positive mass on \((0, \epsilon)\), \(\epsilon \approx 0\), and the
mass increases as the $\xi$ increases. This implies that the robustness property of the GLT prior can be maintained even when $\tau$ is very small, and is adjustable by controlling the $\xi$.

5.2 Tail learnability of the GLT prior

Consider a high-dimensional regression (1) where the coefficients is given by the GLT prior (6) – (9) as displayed in Figure 1. Since the shape parameter $\xi$ is the furthest from the response vector within the hierarchy, it is important to check the properness of the posterior distribution of $\xi$, $\pi(\xi|y) = f(y|\xi) \cdot \pi(\xi) / m(y)$ where $m(y) = \int \int \int f(y|\xi) \cdot \pi(\xi, \beta, \sigma^2, \lambda, \tau, \xi) d\beta d\sigma^2 d\lambda d\tau d\xi$. A rigorous way to show this is to prove that the marginal likelihood $m(y)$ is finite for all values $y \in \mathbb{R}^n$ [58], but this is challenging due to the complexity of the integrand. As a mild verification, we demonstrate properness of two posterior densities: (a) the full conditional posterior density $\pi(\xi|\lambda, \tau)$ used in a Gibbs sampler, and (b) the posterior density $\pi(\xi|y, \beta, \tau, \lambda)$ under a univariate hierarchy without covariates:

**Lemma 6.** [Tail learnability of the GLT prior]

(a) Assume $y \sim N_n(X\beta, \sigma^2 I_n)$ (1) and $\beta \sim \pi_{GLT}(\beta)$ (6) – (9). Then a proportional part of the full conditional posterior for $\xi$ is represented as:

$$
\pi(\xi|\lambda, \tau) \propto V_p(\xi) \cdot \log N_1(\xi|\mu, \rho^2) \cdot \mathcal{I}_{(1/2, \infty)}(\xi), \quad V_p(\xi) = \frac{\pi^{p/2}}{\Gamma(p/\xi + 1)} \prod_{j=1}^{p} r_j(\xi),
$$

where $\{r_j(\xi)\}_{j=1}^{p} = (\tau + \xi \lambda_j)^{-(1/\xi+1)}$. The density $\pi(\xi|\lambda, \tau)$ is proper on $(1/2, \infty)$. Here, $V$ stands for volume.

(b) Assume $y|\beta \sim N_1(\beta, 1)$, $\beta|\lambda \sim N_1(0, \lambda^2)$, $\lambda|\tau, \xi \sim GPD(\tau, \xi)$, and $\tau|\xi \sim IG(1/\xi + 1, 1)$. Let $\pi(\xi)$ be any proper density of $\xi$ supported on $(1/2, \infty)$, i.e., $\int_{1/2}^{\infty} \pi(\xi)d\xi = 1$. Then $\pi(\xi|y, \beta, \tau, \lambda)$ is proper on $(1/2, \infty)$.

Lemma 6 is proved in Subsection S.5.3 in Supplemental Material. Interestingly, the likelihood part of the full conditional posterior density $\pi(\xi|\lambda, \tau)$ (16) has a nice geometric interpretation: if
\[ \xi = 2 \] then the value of \( V_p(2) \) of the density is the volume of a \( p \)-dimensional ellipsoid with \( p \)-radii
\[ \{ r_j(2) = (\tau + 2\lambda_j)^{-3/2} \}^p_{j=1}. \]

6. EXAMPLE– PROSTATE CANCER DATA

6.1 Prostate cancer data

The prostate cancer data is downloadable from R package \texttt{sda}. (See p.272 of [23] for a detail
about the data.) The prostate cancer data is summarized in a 102-by-6033 dimensional matrix
\( X \in \mathbb{R}^{102 \times 6033} \) comprising of gene expression levels measured on microarrays from two classes.
The first 50 rows of \( X \), \( X[1 : 50, \cdot] \in \mathbb{R}^{50 \times 6033} \), correspond to healthy controls, and the remaining
52 rows, \( X[51 : 102, \cdot] \in \mathbb{R}^{52 \times 6033} \), correspond to cancer patients. The \( j \)-th column vector of \( X \),
\( X[\cdot, j] \in \mathbb{R}^{102}, j = 1, \ldots, 6,033 \), represents gene expression levels of the \( j \)-th gene.

The main goal of the study is to discover \( q \) number of interesting genes out of \( p = 6,033 \) genes
whose expression levels differ between the two groups [22]. Such genes are then investigated for
a causal link for the development of prostate cancer. Sparsity level \( s = q/p \) (2) is unknown.

![Figure 4: Histogram of \( z \)-values \( \{ y_{j} \}_{j=1}^{6033} \), one for each gene in the prostate cancer study (left panel) and the Q-Q plot (right panel)](image)

We follow the data transformation method proposed by [21]. Essentially, the objective of the
transformation is to indirectly solve a multiple hypothesis testing [5, 6] by estimating coefficients
for sparse normal mean model [11]:

\[ y_j = \beta_j + \sigma \varepsilon_j, \quad \varepsilon \sim \mathcal{N}(0, 1), \quad j = 1, \ldots, p, \quad (p = 6,033 \text{ genes}), \quad (17) \]

where \( \sigma \) is unknown.

The responses \( \{y_j\}_{j=1}^{p=6,033} \) are acquired as follows. First, for each \( j = 1, \ldots, 6,033 \), obtain \( t \)-test statistics \( t_j \) through a two-sample \( t \)-test statistic with 100 degrees of freedom based on the \( j \)-th vector \( X_{[\cdot, j]} \in \mathbb{R}_{102}^{102} \). Second, convert the acquired \( t \)-test statistics to \( z \)-test statistics using quantile transformation \( y_j = \Phi^{-1}(F_{100 \text{ d.f.}}(t_j)) \), where \( \Phi(\cdot) \) and \( F_{100 \text{ d.f.}}(\cdot) \) are distribution functions of \( \mathcal{N}(0, 1) \) and \( t_{100} \), respectively; refer to Section 2.1 of [22]. The histogram of \( \{y_j\}_{j=1}^{p=6,033} \) along with the standard normal density and its Q-Q plot are displayed in the left and right panels in the Figure 4, respectively.

Denote \( H_{0j} \) as the null hypothesis that posits no difference in the gene expression levels for the \( j \)-th gene between the healthy controls and cancer patients. If the global null hypothesis \( \cap_{j=1}^{6033} H_{0j} \) is true, the histogram of \( \{y_j\}_{j=1}^{p=6,033} \) should mimic a standard normal density closely. The presence of outliers is evident from the panels. Those outliers may correspond to an interesting gene (cancerous genes) that reject the null hypotheses [21].

### 6.2 Prostate cancer data analysis via the Horseshoe

In this example, we compare the performances of the Horseshoe (12) – (13) and the GLT prior (6) – (9) when they are used as a sparse-inducing prior for the coefficients for the sparse normal mean model (17) as the number of genes considered increases. To that end, we construct seven prostate datasets, denoted by \( \mathcal{P}_1 = \{y_j\}_{j=1}^{p=50}, \mathcal{P}_2 = \{y_j\}_{j=1}^{p=100}, \mathcal{P}_3 = \{y_j\}_{j=1}^{p=200}, \mathcal{P}_4 = \{y_j\}_{j=1}^{p=500}, \mathcal{P}_5 = \{y_j\}_{j=1}^{p=1000}, \mathcal{P}_6 = \{y_j\}_{j=1}^{p=3000}, \) and \( \mathcal{P}_7 = \{y_j\}_{j=1}^{p=6033} \). That way, it holds subset inclusions, \( \mathcal{P}_1 \subset \mathcal{P}_2 \subset \mathcal{P}_3 \subset \mathcal{P}_4 \subset \mathcal{P}_5 \subset \mathcal{P}_6 \subset \mathcal{P}_7 \), where \( \mathcal{P}_7 \) is the full dataset for the 6,033 genes. \( \mathcal{P} \) stands for prostate.

We show the posterior inference results by plotting ordered pairs \( \{(y_j, \hat{\beta}_j)\}_{j=1}^{p} \) such that \( \hat{\beta}_j \) represents the posterior mean of \( \beta_j \) for the \( j \)-th gene. Provided the (tail) robustness property holds...
we should see a reversed-S-shape curve formed by the pairs \(\{(y_j, \hat{\beta}_j)\}_{j=1}^p\) for each of the dataset [21? ]. (See Section 2 in [14] for a more detail about the robustness property.) The reversed-S-shape curve (like the left panel in Figure 5) is formed because of (i) \(\hat{\beta}_j \approx 0\) corresponding to a noise coefficient, (ii) \(\hat{\beta}_j \approx y_j\) corresponding to a signal coefficient, and (iii) continuity nature of a one-group Gaussian mixture prior.

![Figure 5: Idealistic reversed-S-shape curve (left panel) formed by pairs \(\{(y_j, \hat{\beta}_j)\}_{j=1}^p\) when a continuous shrinkage prior achieves the robustness property. Posterior inference results (right panel) obtained by the Horseshoe based on the dataset \(\mathcal{P}_l, l = 1, 2, 3, 7\). Dotted line is \(y = x\).](image)

Throughout the paper, we implement the Horseshoe (12) – (13) via the \texttt{R} function \texttt{Horseshoe} within the \texttt{R} package \texttt{Horseshoe}. More specifically, given a response vector \(y = (y_1, \cdots, y_p)^\top \in \mathbb{R}^p\) formulated from one of the seven prostate cancer dataset, we use \texttt{Horseshoe(y=y, X=I, method.tau="halfCauchy", method.sigma="Jeffreys", burn=10000, nmc=10000, thin=100)} where \(y = y\) and \(X = I_p\), to produce 100 thinned realizations from the posterior distribution \(\pi(\beta|y)\) via MCMC [1, 59].

Ordered pairs \(\{(y_j, \hat{\beta}_j)\}_{j=1}^p\) based on the datasets \(\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3,\) and \(\mathcal{P}_7\) are overlaid on the right panel of the Figure 5. (Results for others datasets are omitted.) The results suggests that the robustness property is manifested only when \(p = 50\), and the property disappears as \(p\) increases. When \(p = 200\) or more the posterior mean is numerically zero (that is, \(\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)^\top \approx 0\) =
$(0, \cdots, 0)^\top \in \mathbb{R}^p)$, hence, the plotted dots are essentially $\{(y_j, 0)\}_{j=1}^p$, displaying the collapsing behaviour.

In general, note that the ordinary estimates for the $\beta_j$'s are $y_j$ in the normal mean model (17), while Bayes estimates are typically biased, pulled towards a prior information on the $\beta_j$ [42]. What we observed is an anomalously strong pulling towards the prior mean. A similar real data example is discovered in analyzing gene expression data from breast cancer patients; refer to Section S.1 in Supplemental Material.

6.3 Prostate cancer data analysis via the GLT prior

Posterior inference results obtained using the GLT prior (6) – (9) are shown on the Figure 6: the left and right panels display the pairs $\{(y_j, \hat{\beta}_j)\}_{j=1}^p$, when applied to the four datasets $\mathcal{P}_l$, $l = 1, 2, 3, 4$, and the three datasets $\mathcal{P}_l$, $l = 5, 6, 7$, respectively. The desirable reversed-S-shape curves are formed in all datasets, which implies that the robustness property holds regardless of the number of genes used. It is important to emphasize that neither Horseshoe and GLT prior requires any tuning procedure, hence, the comparison of performances is fair.

Figure 6: Posterior inference results obtained by the GLT prior applied seven prostate cancer datasets $\mathcal{P}_l$, $l = 1, \cdots, 7$. Posterior means of $(\tau, \xi)$ corresponding to the seven datasets are $(0.0303, 1.620)$ ($\mathcal{P}_1$), $(0.0154, 1.662)$ ($\mathcal{P}_2$), $(0.0090, 1.789)$ ($\mathcal{P}_3$), $(0.0037, 1.905)$ ($\mathcal{P}_4$), $(0.0019, 1.991)$ ($\mathcal{P}_5$), $(0.0013, 2.760)$ ($\mathcal{P}_6$), and $(0.0013, 3.636)$ ($\mathcal{P}_7$), respectively.

Posterior means of the shape parameter $\xi$ for the seven datasets are $1.620$ ($\mathcal{P}_1$), $1.662$ ($\mathcal{P}_2$),
1.789 (P3), 1.905 (P4), 1.991 (P5), 2.760 (P6), and 3.636 (P7), respectively. This may suggest that the posterior tail-thickness gets heavier as the number of genes considered \( p \) increases to accommodate a growing number of interesting genes \( q \), which is unknown.

7. EXAMPLE– SIMULATION STUDY WITH VARIED SPARSITY LEVEL

7.1 Artificial high-dimensional data generator

We elaborate on the generation of high-dimensional data \((y, X) \in \mathbb{R}^n \times \mathbb{R}^{n \times p}\) from a high-dimensional regression (1) corresponding to a simulation environment \((n, p, q, \varrho, \text{SNR})\):

\[
(y, X) \sim p(y, X) = \mathcal{N}_n(y | X\beta_0, \sigma_0^2 I_n) \cdot \prod_{i=1}^n \mathcal{N}_p(x_i^T | 0, \Upsilon(\varrho)), \quad \Upsilon(\varrho) = \varrho J_p + (1 - \varrho) I_p,
\]

where \( p(y, X) \) is a true data generating process. The vector \( \beta_0 = (\beta_{0,1}, \ldots, \beta_{0,q}, \beta_{0,q+1}, \ldots, \beta_{0,p})^T \in \mathbb{R}^p \) is the true coefficients with \( q \) number of unit signals \( \beta_{0,1} = \cdots = \beta_{0,q} = 1 \) and \( p - q \) number of noises \( \beta_{0,q+1} = \cdots = \beta_{0,p} = 0 \). Sparsity level (2) is then \( s = q/p \). The matrices \( I \) and \( J \) indicate an identity matrix and a matrix whose elements are ones, respectively. The signal-to-noise ratio (SNR) is defined by \( \text{SNR} = \text{var}(X\beta_0) / \{ \text{SNR} \cdot \text{var}(\epsilon) \} \). The value \( \varrho \) is a number associated with column-wise correlations in the design matrix \( X \).

After specifying a simulation environment \((n, p, q, \varrho, \text{SNR})\), we use the following three steps to generate an artificial high-dimensional data \((y, X)\) distributed according to (18). (i) Generate a matrix \( X \in \mathbb{R}^{n \times p} \) where each row vector \( x_i \in \mathbb{R}^p \) is independently sampled from \( \mathcal{N}_p(0, \Upsilon(\varrho)) \).

Next, center the matrix \( X \) column-wisely so that each column vector \( X[:,j] \in \mathbb{R}^n \) has zero mean. After that, normalize each column vector to be of the unit Euclidean \( l_2 \)-norm. (ii) Generate \( n \)-dimensional Gaussian error \( \epsilon \sim \mathcal{N}_n(0, I_n) \). (iii) Add the mean part \( X\beta_0 \) and the error part \( \sigma_0 \epsilon \) to create the response vector \( y = X\beta_0 + \sigma_0 \epsilon \), where \( \sigma_0^2 = \text{var}(X\beta_0) / \{ \text{SNR} \cdot \text{var}(\epsilon) \} \), with \( \text{var}(z) = \sum_{i=1}^n (z_i - \bar{z})^2 / (n - 1) \) for \( z \in \mathbb{R}^n \).
7.2 The Horseshoe under varied sparsity level

To investigate the behaviour of the Horseshoe (12) – (13) when the sparsity level $s = q/p$ (2) increases, we generated four artificial datasets $A_l = (y, X) \in \mathbb{R}^{n} \times \mathbb{R}^{nxp}$, $l = 1, 2, 3, 4$, corresponding to four simulation environments ($n = 100, p = 500, q, \varrho = 0$, SNR = 5) with different values for $q = 2, 5, 8, 13$. The sparsity levels of the datasets are $2/500 = 0.004$ ($A_1$), $5/500 = 0.01$ ($A_2$), $8/500 = 0.016$ ($A_3$), and $13/500 = 0.026$ ($A_4$), respectively.

The results of posterior inference are displayed in Figure 7. Panels are arranged in a way that the sparsity level increases from the left to right. Panels in the first, second, and third rows in the Figure 7 display the 95% credible intervals for $\{\beta_j\}_{j=1}^{p}$, and those of $\{\lambda_j\}_{j=1}^{p}$, and posterior correlations $\{\text{cor}(\lambda_j, \tau | y)\}_{j=1}^{p}$, respectively. For the ease of visualization, results corresponding to only the first 50 coefficients of $\beta$ are plotted. The results corresponding to signals and noises are colored blue and red, respectively, and the true coefficient vector $\beta_0$ is colored green.

Results show that the Horseshoe works reasonably well on the datasets $A_1$, $A_2$, and $A_3$, but produces the collapsing behaviour ($\hat{\beta} \approx \mathbf{0}$) on the dataset $A_4$ as similar to what we observed in the prostate cancer dataset in the Figure 5. The posterior means of $\tau$ corresponding to the datasets are $1.41 \cdot 10^{-6}$ ($A_1$), $0.05$ ($A_2$), $0.13$ ($A_3$), and $6.53 \cdot 10^{-15}$ ($A_4$), respectively. Hence, the posterior mean of $\tau$ gradually increases as the sparsity level increases and after some threshold it suddenly drops to a very small number.
Figure 7: Results of posterior inference by using the Horseshoe under varying sparsity levels: $A_1$ (first column), $A_2$ (second column), $A_3$ (third column), and $A_4$ (fourth column). The results of posterior inference corresponding to signals and noises are colored in blue and red, respectively, and the truth $\beta_0$ is colored in green. Posterior means of $\tau$ corresponding to the four datasets are $1.41 \cdot 10^{-6}$ ($A_1$), $0.05$ ($A_2$), $0.13$ ($A_3$), and $6.53 \cdot 10^{-15}$ ($A_4$), respectively.

The relationship between the local $\{\lambda_j\}_{j=1}^p$ and the global $\tau$ scale parameters is the key to comprehend how the Horseshoe detects signals from a posteriori perspective. Observe that the $\tau$ is associated with the sparsity level [57], and is expected to be large in presence of a relatively high number of signals. The panels on the third row of the Figure 7 show weak negative posterior correlation between the $\lambda_j$ and $\tau$, $\text{cor}(\lambda_j, \tau|y)$, for each $j = 1, \cdots, p$. As seen on the panels on the first and third rows of the Figure 7, the selected signals among the $p$ coefficients $\{\beta_j\}_{j=1}^p$, saying $\{\beta_j\}_{j \in Q}$, $Q \subset P = \{1, \cdots, p\}$, are those whose corresponding posterior correlations $\{\text{cor}(\lambda_j, \tau|y)\}_{j \in Q}$ attain even stronger negative values than others $\{\text{cor}(\lambda_j, \tau|y)\}_{j \in P-Q}$. This implies that if there are no discriminable differences among the correlations $\{\text{cor}(\lambda_j, \tau|y)\}_{j \in P}$ then the Horseshoe loses its signal detection mechanism and produces the collapsing behaviour as seen on the panel of the fourth column.
It is important to emphasize that a collapsing behaviour for the Horseshoe (12) – (13) was pointed out by several authors but did not draw much attention in the literature. Recently, [76] in discussion of [70] noted a danger of collapse of marginal maximum-likelihood estimator for the global-scale parameter $\tau$ when the sparsity level $s$ (2) is very small. On the other hand, our research points out a collapsing behaviour of the fully Bayesian Horseshoe estimator where the $\tau$ is distributed according to $\mathcal{C}^+(0, 1)$ when the sparsity level is moderately large.

7.3 The GLT prior under varied sparsity level

We applied the GLT prior (6) – (9) to the same four artificial datasets $\mathcal{A}_l$ ($l = 1, 2, 3, 4$) used in the previous subsection. Figure 8 displays the results of posterior inference: posterior correlation between each of the local-scale parameters $\lambda_j$ and the shape parameter $\xi$ (that is, $\{\text{cor}(\lambda_j, \xi | y)\}_{j=1}^p$) are additionally plotted in the panels on the fourth row of the Figure 8. Note that the GLT prior can detect signals in the dataset $\mathcal{A}_4$ where the Horseshoe collapsed. The results suggest that the GLT prior works reasonably well across diverse sparse regimes. Thorough replicated simulation studies are conducted on Section 8.

As the sparsity level $s$ (2) is known, we can check the tail-adaptive shrinkage property. The posterior means of the shape parameter $\xi$ corresponding to the four datasets are 2.010 ($\mathcal{A}_1$), 2.134 ($\mathcal{A}_2$), 2.235 ($\mathcal{A}_3$), 2.347 ($\mathcal{A}_4$), respectively. The increasing monotonicity implies that the tail of the GLT prior is adaptive to the sparsity level a posteriori.

We describe a signal detection mechanism of the GLT prior (6) – (9). The GLT prior perceives the signal detection problem as a mirror image of the extreme value identification problem [57, 73]. Under the GLT prior formulation, a selected (signal) coefficient $\beta_j$ is the one whose corresponding local-scale parameter $\lambda_j$ is an extreme value possibly located at the tail part of the local-scale density $f(x|\tau, \xi) = \mathcal{GPD}(x|\tau, \xi)$ (7). As conventional distributional theory, role of the global-scale parameter $\tau$ is to scale the local-scale density $f$, and that of the shape parameter $\xi$ is to control the tail-heaviness of the $f$. As seen from panels on the third and the forth rows of the Figure 8, the estimates of $\tau$ and $\xi$ are nearly independent from those of $\{\lambda_j\}_{j=1}^p$ a posteriori.
Figure 8: Results of posterior inference obtained using the GLT prior when applied to the same four artificial datasets used in the Figure 7: $A_1$ (first column), $A_2$ (second column), $A_3$ (third column), and $A_4$ (fourth column). Posterior means of $(\tau, \xi)$ corresponding to the four datasets are $(0.003, 2.010)$ ($A_1$), $(0.004, 2.134)$ ($A_2$), $(0.004, 2.235)$ ($A_3$), and $(0.004, 2.347)$ ($A_4$), respectively.

8. SIMULATIONS

8.1 Outline

In Subsection 7.1, we illustrated how to generate an artificial high-dimensional data $(y, X) \in \mathbb{R}^n \times \mathbb{R}^{n \times p}$ given a simulation environment $(n, p, q, \varrho, \text{SNR})$ from a linear regression (1) when the truth $\beta_0 = (\beta_{0,1}, \ldots, \beta_{0,q}, \beta_{0,q+1}, \ldots, \beta_{0,p})^T \in \mathbb{R}^p$ is specified by $\beta_{0,1} = \cdots = \beta_{0,q} = 1$ and $\beta_{0,q+1} = \cdots = \beta_{0,p} = 0$. In the present section, we conduct a replicated study to compare the performances of the Horseshoe (12) – (13) and the GLT prior (6) – (9) under three different scenarios. Set the default environmental values by $(n, p) \in \{(100, 500), (200, 1000)\}$, $s = q/p = 0.01$, SNR = 5, and $\varrho = 0$, and then separately consider the following three scenarios by varying one environmental value while fixing others;
**Scenario 1**: varied sparsity level \( s = q/p \) from 0.001 to 0.1,

**Scenario 2**: varied \( \varrho \) from 0 to 0.5,

**Scenario 3**: varied SNR from 2 to 10.

We separately report the medians of mean squared errors (MSE) corresponding to signal and noise coefficients measured across the 50 replicated datasets. Let \( \hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^\top \in \mathbb{R}^p \) is the posterior mean obtained by using either the Horseshoe or the GLT prior: then, MSE corresponding to signals and noises are defined by

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

Note that when the posterior mean \( \hat{\beta} \) collapses, the two metrics \( \text{MSE}_S \) and \( \text{MSE}_N \) numerically become 1 and 0, respectively. The posterior computations for the Horseshoe and the GLT prior are fully automated and tuning-free, hence the comparison is fair.
8.2 Scenario 1: varied sparsity level $q/p$

Figure 9: Medians of $\text{MSE}_S$, $\text{MSE}_N$, and posterior means of $\tau$ and $\xi$ across different sparsity level $q/p$: $(n, p) = (100, 500)$ (top panel) and $(n, p) = (200, 1000)$ (bottom panel). Results from the GLT prior are marked with black circle dot •; black triangle dot ▲ (right panels) represents the posterior mean of $\xi$. Results from the Horseshoe are marked with blue square dot ■. The red dotted horizontal line represents the zeros.

Figure 9 displays the medians of $\text{MSE}_S$, $\text{MSE}_N$, and posterior means of $\tau$ and $\xi$ under Scenario 1. The top and bottom panels correspond to $(n, p) = (100, 500)$ and $(n, p) = (200, 1000)$, respectively. To be specific, the top panel corresponds to the setting $(n = 100, p = 500, q, \varrho = 0, \text{SNR} = 5)$ with $q \in \{1, 6, 11, 16, 22, 27, 32, 37, 43, 48\}$, and the bottom panel corresponds to $(n = 200, p = 1000, q, \varrho = 0, \text{SNR} = 5)$ with $q \in \{1, 11, 22, 32, 43, 53, 64, 74, 85, 95\}$ so that the sparsity level $q/p$ varies from 0.001 to 0.1.

The Horseshoe performs well if the sparsity level $q/p$ is less than $11/500 = 22/1000 = 0.022$, but suddenly collapsed beyond this due to a sharp decrease in the posterior mean of $\tau$. On the other hand, the GLT prior works reasonably well across diverse sparse regimes, and the posterior means of the $\tau$ are maintained at around 0.004, and the posterior means of $\xi$ increase as the sparsity level increases to $q/p = 0.022$. 

27
8.3 **Scenario 2: varied \( \varrho \)**

Figure 10: Medians of \( \text{MSE}_S \), \( \text{MSE}_N \), and posterior means of \( \tau \) and \( \xi \) across different sparsity level \( q/p \): \((n, p) = (100, 500)\) (top panel) and \((n, p) = (200, 1000)\) (bottom panel).

Figure 10 displays the medians of \( \text{MSE}_S \), \( \text{MSE}_N \), and posterior means of \( \tau \) and \( \xi \) under Scenario 2. MSE\(_S\) are MSE\(_N\) obtained from the both priors increase as \( \varrho \) increases. The GLT prior shows better signal detection whereas the Horseshoe shows slightly better noise shrinkage. When \((n, p) = (200, 1000)\) and \( \varrho = 0.5 \) the GLT prior outperforms the Horseshoe as seen on the bottom-middle panel.
8.4 Scenario 3: varied signal-to-noise ratio

Figure 11: Medians of $\text{MSE}_S$, $\text{MSE}_N$, and posterior means of $\tau$ and $\xi$ across different sparsity level $q/p$: $(n, p) = (100, 500)$ (top panel) and $(n, p) = (200, 1000)$ (bottom panel).

Figure 11 displays the medians of $\text{MSE}_S$, $\text{MSE}_N$, and posterior means of $\tau$ and $\xi$ under Scenario 3. $\text{MSE}_S$ are $\text{MSE}_N$ obtained from the both priors monotonically decreases as the value of SNR increases. Both priors show excellent performances in shrinkage noises as seen in the middle panels. When SNR = 2, the GLT prior shows better performance than the Horseshoe as seen in the left panels.

9. DISCUSSION

In this paper we have proposed a new shrinkage prior (GLT) for high-dimensional regression problems. The purpose of this article is not to criticize the existing continuous shrinkage priors or Horseshoe in particular, but is simply to recognize that these priors are devised to produce meaningful results only in the regime where there are a handful of true signals, an ultra sparse regime. However, as briefly motivated in Section 2, it is often necessary to devise a sparse estimation method which works reasonably well across diverse sparsity regimes. We proposed the GLT prior
with the tail-adaptive shrinkage property to address this gap. This was further supported by the
two real gene expression datasets and simulation studies. We also conducted simulation studies
for variant versions of the Horseshoe, namely truncated Horseshoe [66], Horseshoe-plus [8], and
regularized Horseshoe [56]; refer to Section S.2 in Supplemental Material. In majority of the
cases, we observed superior performance of the GLT prior in signal detection. We also explored an
application of the GLT prior to a curve fitting study; refer to Section S.3 in Supplemental Material.

We emphasize that delicate care is required to estimate the shape parameter $\xi$ within the global-
local-tail shrinkage framework and we regard this as one of the salient contributions of the paper.
For the GLT prior, we proposed an algorithm which combined the elliptical slice sampler [51]
and the Hill estimator [35] from the extreme value theory which obviates the need for tuning any
hyper-parameters. This automatic-tuning leads to learning the shape parameter $\xi$ adaptive to the
unknown sparsity level. Refer to the Appendix in Supplemental Material for more details.
REFERENCES

[1] ANDRIEU, C., DE FREITAS, N., DOUCET, A. & JORDAN, M. I. (2003). An introduction to mcmc for machine learning. *Machine learning* **50**, 5–43.

[2] ARMAGAN, A., DUNSON, D. & LEE, J. (2010). Bayesian generalized double pareto shrinkage. *Biometrika*.

[3] BAI, R. & GHOSH, M. (2018). High-dimensional multivariate posterior consistency under global–local shrinkage priors. *Journal of Multivariate Analysis* **167**, 157–170.

[4] BAI, R. & GHOSH, M. (2018). On the beta prime prior for scale parameters in high-dimensional bayesian regression models. *arXiv preprint arXiv:1807.06539*.

[5] BENJAMINI, Y. & HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57**, 289–300.

[6] BENJAMINI, Y. & YEKUTIELI, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association* **100**, 71–81.

[7] BERTRAM, J. S. (2000). The molecular biology of cancer. *Molecular aspects of medicine* **21**, 167–223.

[8] BHADRA, A., DATTA, J., POLSON, N. G., WILLARD, B. et al. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis* **12**, 1105–1131.

[9] BHADRA, A., DATTA, J., POLSON, N. G. & WILLARD, B. T. (2017). Lasso meets horseshoe. *arXiv preprint arXiv:1706.10179*.

[10] BHATTACHARYA, A., CHAKRABORTY, A. & MALLICK, B. K. (2016). Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, asw042.
[11] BHATTACHARYA, A., PATI, D., PILLAI, N. S. & DUNSON, D. B. (2015). Dirichlet–laplace priors for optimal shrinkage. *Journal of the American Statistical Association* **110**, 1479–1490.

[12] BÜHLMANN, P. & VAN DE GEER, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

[13] CARVALHO, C. M., POLSON, N. G. & SCOTT, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*.

[14] CARVALHO, C. M., POLSON, N. G. & SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480.

[15] CASELLA, G. & GEORGE, E. I. (1992). Explaining the gibbs sampler. *The American Statistician* **46**, 167–174.

[16] CASTILLO, I., SCHMIDT-HIEBER, J., VAN DER VAART, A. et al. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics* **43**, 1986–2018.

[17] CASTILLO, I., VAN DER VAART, A. et al. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics* **40**, 2069–2101.

[18] CHICCOLI, C., LORENZUTTA, S. & MAINO, G. (1992). Concerning some integrals of the generalized exponential-integral function. *Computers & Mathematics with Applications* **23**, 13–21.

[19] CLAMP, M., FRY, B., KAMAL, M., XIE, X., CUFF, J., LIN, M. F., KELLIS, M., LINDBLAD-TOH, K. & LANDER, E. S. (2007). Distinguishing protein-coding and non-coding genes in the human genome. *Proceedings of the National Academy of Sciences* **104**, 19428–19433.

[20] COLES, S., BAWA, J., TRENNER, L. & DORAZIO, P. (2001). *An introduction to statistical modeling of extreme values*, vol. 208. Springer.
[21] EFRON, B. (2010). The future of indirect evidence. *Statistical science: a review journal of the Institute of Mathematical Statistics* **25**, 145.

[22] EFRON, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, vol. 1. Cambridge University Press.

[23] EFRON, B. & HASTIE, T. (2016). *Computer age statistical inference*, vol. 5. Cambridge University Press.

[24] EMBRECHTS, P., KLÜPPELBERG, C. & MIKOSCH, T. (2013). *Modelling extremal events: for insurance and finance*, vol. 33. Springer Science & Business Media.

[25] ESPLIN, E. D., CHITTENDEN, A. B., STOPFER, J., SCHNEIDER, K., SACCA, R., STICKEVERS, S., KIPNIS, L., KOELLER, D. R., GAONKAR, S., SOTELO, J. et al. (2018). Beyond brca1/2: Clinician-reported utility 3 years post panel testing.

[26] FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2001). *The elements of statistical learning*, vol. 1. Springer series in statistics New York.

[27] GEORGE, E. I. & McCULLOCH, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

[28] GEORGE, E. I. & McCULLOCH, R. E. (1995). Stochastic search variable selection. *Markov chain Monte Carlo in practice* **68**, 203–214.

[29] GHOSH, P., CHAKRABARTI, A. et al. (2017). Asymptotic optimality of one-group shrinkage priors in sparse high-dimensional problems. *Bayesian Analysis* **12**, 1133–1161.

[30] GIERZ, G., HOFMANN, K., KEIMEL, K., LAWSON, J., MISLOVE, M. & SCOTT, D. (2003). Encyclopedia of mathematics and its applications.

[31] GRIFFIN, J., BROWN, P. et al. (2017). Hierarchical shrinkage priors for regression models. *Bayesian Analysis* **12**, 135–159.
[32] GRIFFIN, J. E., BROWN, P. J. et al. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* 5, 171–188.

[33] GULISASHVILI, A., TANKOV, P. et al. (2016). Tail behavior of sums and differences of log-normal random variables. *Bernoulli* 22, 444–493.

[34] HASTIE, T., TIBSHIRANI, R. & WAINWRIGHT, M. (2015). Statistical learning with sparsity: the lasso and generalizations. Chapman and Hall/CRC.

[35] HILL, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The annals of statistics*, 1163–1174.

[36] JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 186, 453–461.

[37] JOHNDROW, J. E., ORENSTEIN, P. & BHATTACHARYA, A. (2017). Bayes shrinkage at gwas scale: Convergence and approximation theory of a scalable mcmc algorithm for the horseshoe prior. *arXiv preprint arXiv:1705.00841*.

[38] JOHNSON, V. E. & ROSSELL, D. (2010). On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, 143–170.

[39] JOHNSON, V. E. & ROSSELL, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association* 107, 649–660.

[40] KARAMATA, J. (1933). Sur un mode de croissance régulière. théorèmes fondamentaux. *Bulletin de la Société Mathématique de France* 61, 55–62.

[41] LEE, S. & KIM, J. H. (2018). Exponentiated generalized pareto distribution: Properties and applications towards extreme value theory. *Communications in Statistics-Theory and Methods*, 1–25.
[42] Lindley, D. V. & Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 1–18.

[43] Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V. et al. (2018). An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416.

[44] Maric, V. (2000). *Regular variation and differential equations*, vol. 1726. Springer Science & Business Media.

[45] Martin, R., Mess, R., Walker, S. G. et al. (2017). Empirical bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* **23**, 1822–1847.

[46] McCullagh, P. & Polson, N. G. (2018). Statistical sparsity. *Biometrika* **105**, 797–814.

[47] McNeil, A. J., Frey, R. & Embrechts, P. (2015). *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press.

[48] Mikosch, T. (1999). *Regular variation, subexponentiality and their applications in probability theory*. Eindhoven University of Technology.

[49] Milgram, M. (1985). The generalized integro-exponential function. *Mathematics of computation* **44**, 443–458.

[50] Mitchell, T. J. & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83**, 1023–1032.

[51] Murray, I., Prescott Adams, R. & MacKay, D. J. (2010). Elliptical slice sampling.

[52] Narod, S. A. & Foulkes, W. D. (2004). Brca1 and brca2: 1994 and beyond. *Nature Reviews Cancer* **4**, 665.

[53] Park, T. & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association* **103**, 681–686.
[54] Pati, D., Bhattacharya, A., Pillai, N. S., Dunson, D. et al. (2014). Posterior contraction in sparse bayesian factor models for massive covariance matrices. *The Annals of Statistics* **42**, 1102–1130.

[55] Pickands III, J. et al. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics* **3**, 119–131.

[56] Pironen, J., Vehtari, A. et al. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics* **11**, 5018–5051.

[57] Polson, N. G. & Scott, J. G. (2010). Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics* **9**, 501–538.

[58] Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media.

[59] Robert, C. & Casella, G. (2013). *Monte Carlo statistical methods*. Springer Science & Business Media.

[60] Rouleau, M., Patel, A., Hendzel, M. J., Kaufmann, S. H. & Poirier, G. G. (2010). Parp inhibition: Parp1 and beyond. *Nature reviews cancer* **10**, 293.

[61] Song, Q. & Liang, F. (2017). Nearly optimal bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964* .

[62] Stephens, P. J., Tarpey, P. S., Davies, H., Van Loo, P., Greenman, C., Wedge, D. C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G. R. et al. (2012). The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400.

[63] Stratton, M. R., Campbell, P. J. & Futreal, P. A. (2009). The cancer genome. *Nature* **458**, 719.

[64] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* , 267–288.
[65] TIPPING, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research* **1**, 211–244.

[66] VAN DER PAS, S., KLEIJN, B., VAN DER VAART, A. et al. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics* **8**, 2585–2618.

[67] VAN DER PAS, S., SALOMOND, J.-B., SCHMIDT-HIEBER, J. et al. (2016). Conditions for posterior contraction in the sparse normal means problem. *Electronic journal of statistics* **10**, 976–1000.

[68] VAN DER PAS, S., SZABÓ, B. & VAN DER VAART, A. (2016). How many needles in the haystack? adaptive inference and uncertainty quantification for the horseshoe. *arXiv preprint arXiv:1607.01892*.

[69] VAN DER PAS, S., SZABÓ, B., VAN DER VAART, A. et al. (2017). Adaptive posterior contraction rates for the horseshoe. *Electronic Journal of Statistics* **11**, 3196–3225.

[70] VAN DER PAS, S., SZABÓ, B., VAN DER VAART, A. et al. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis* **12**, 1221–1274.

[71] VAN DER PAS, S. L., KLEIJN, B. J., VAN DER VAART, A. W. et al. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics* **8**, 2585–2618.

[72] WEINSTEIN, J. N., COLLISSON, E. A., MILLS, G. B., SHAW, K. R. M., OZENBERGER, B. A., ELLROTT, K., SHMULEVICH, I., SANDER, C., STUART, J. M., NETWORK, C. G. A. R. et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics* **45**, 1113.

[73] WEST, M. (1984). Outlier models and prior distributions in bayesian linear regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, 431–439.
[74] Willyard, C. (2018). New human gene tally reignites debate. *Nature* **558**, 354.

[75] Yang, Y., Wainwright, M. J., Jordan, M. I. et al. (2016). On the computational complexity of high-dimensional bayesian variable selection. *The Annals of Statistics* **44**, 2497–2532.

[76] Yoo, W. W. (2017). Contributed discussion to uncertainty quantification for the horseshoe by st\'ephanie van der pas, botond szab\'o and aad van der vaart. *arXiv preprint arXiv:1710.05987*.

[77] Zhang, R. & Ghosh, M. (2019). Ultra high-dimensional multivariate posterior contraction rate under shrinkage priors. *arXiv preprint arXiv:1904.04417*. 
Supplemental Material to

*Tail-adaptive Bayesian shrinkage*

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, 3143 TAMU, College Station, TX, U.S.A.

APPENDIX

A.1 Gibbs sampler

Consider \( y = X\beta + \sigma \epsilon \), \( \epsilon \sim N_n(0, I_n) \) (1), \( \sigma^2 \sim \pi(\sigma^2) \propto 1/\sigma^2 \), and \( \beta \sim \pi_{GLT}(\beta) \) (6) – (9). Let \( \Omega = (\beta, \sigma^2, \lambda, \tau, \xi) \in \mathbb{R}^p \times (0, \infty) \times (0, \infty)^p \times (0, \infty) \times (1/2, \infty) \) denotes all the latent random variables. Current Section provides a full description for posterior computation to sample from the full joint posterior distribution, \( \pi(\Omega|y) \).

The full joint density \( \pi(\Omega|y) \) is proportional to

\[
N_n(y|X\beta, \sigma^2 I_n)N_p(\beta|0, \sigma^2 \Lambda)\pi(\sigma^2)\left\{ \prod_{j=1}^p \pi(\lambda_j|\tau, \xi) \right\} \pi(\tau, \xi), \quad \Lambda = \text{diag}(\lambda^2_1, \cdots, \lambda^2_p) \in \mathbb{R}^{p \times p}
\]

\[
\propto N_n(y|X\beta, \sigma^2 I_n)N_p(\beta|0, \sigma^2 \Lambda)\pi(\sigma^2)\left\{ \prod_{j=1}^p \mathcal{GPD}(\lambda_j|\tau, \xi) \right\} \mathcal{IG}(\tau|p/\xi + 1, 1) \log N(\xi|\mu, \rho^2)\mathcal{I}(1/2, \infty)(\xi).
\]

Note that the \( \mu \in \mathbb{R} \) and \( \rho^2 > 0 \) are hyper-parameters, which typically requires an expert-tuning. Section A.4 provides an automated hyper-parameter turning algorithm.

Since full joint posterior distribution \( \pi(\Omega|y) \) is not in a closed form, we develop a Markov chain Monte Carlo sampling algorithm to sample \( \Omega \) from the distribution. The following algorithm provides a Gibbs sampler [20]
**Step 1.** Sample $\beta$ from conditional posterior

$$\pi(\beta|\cdot) \sim \mathcal{N}_p(\Sigma X^\top y, \sigma^2 \Sigma), \quad \Sigma = (X^\top X + \Lambda^{-1})^{-1} \in \mathbb{R}^{p \times p}.$$ 

**Step 2.** Sample $\sigma^2$ from conditional posterior

$$\pi(\sigma^2|\cdot) \sim \mathcal{IG}\left(\frac{n+p}{2}, \frac{\|y - X\beta\|^2 + \beta^\top \Lambda^{-1} \beta}{2}\right).$$

**Step 3.** Update $\lambda_j$, $j = 1, \cdots, p$, independently using slice sampler [27] within the Gibbs sampler. Proportional part of full conditional posterior is

$$\pi(\lambda_j|\cdot) \propto \frac{1}{\lambda_j} \exp\left(-\frac{\beta_j^2}{2\sigma^2 \lambda_j^2}\right) \cdot \left(1 + \frac{\xi \lambda_j}{\tau}\right)^{-\left(1/\xi+1\right)}. \quad (A.1)$$

**Step 4.** Update $\tau$ using slice sampler [27] within the Gibbs sampler. Proportional part of full conditional posterior is

$$\pi(\tau|\cdot) \propto \tau^{-2} \exp(-1/\tau) \cdot \prod_{j=1}^p (\tau + \xi \lambda_j)^{-\left(1/\xi+1\right)}. \quad (A.2)$$

**Step 5.** Update $\xi$ using elliptical slice sampler [26] after variable change $\eta = \log \xi$ within the Gibbs sampler. Proportional part of full conditional posterior is

$$\pi(\xi|\cdot) \propto \mathcal{V}_p(\xi) \cdot \log \mathcal{N}_1(\xi|\mu, \rho^2) \cdot \mathcal{I}_{(1/2,\infty)}(\xi), \quad (A.3)$$

where $\mathcal{V}_p(\xi) = \{\Gamma(p/\xi+1)\}^{-1} \pi^{p/2} \prod_{j=1}^p r_j(\xi)$ with $r_j(\xi) = (\tau + \xi \lambda_j)^{-\left(1/\xi+1\right)}$, $j = 1, \cdots, p$.

A.2 **Slice sampler implementation in Step 3 and Step 4**

Slice sampler [27] is a popular technique to adapt the step-size of a MCMC algorithm and is based on the local property of the target density. The basic idea is parameter expansion which involves intentional introduction of auxiliary variables [7]. Finding an appropriate parameter expansion
depends on the functional form of the target density.

Let \( j \in \{1, \cdots, p\} \). To implement the slice sampler in the Step 3 (A.1), first use change of variable, \( \gamma_j = \lambda_j^2 \), to get

\[
\pi(\gamma_j|\cdot) \propto \gamma_j^{-1} \exp (-m_j/\gamma_j) \cdot (\tau + \xi \sqrt{\gamma_j})^{-\left(1/\xi+1\right)}
\]
\[
= \gamma_j^{-1} \exp (-m_j/\gamma_j) \cdot (\sqrt{\gamma_j})^{-\left(1/\xi+1\right)} \cdot (\tau + \xi \sqrt{\gamma_j})^{-\left(1/\xi+1\right)}
\]
\[
= \gamma_j^{-\left(1/\xi+1\right)/2-1} \exp (-m_j/\gamma_j) \cdot (\xi + \tau \cdot \gamma_j^{-1/2})^{-\left(1/\xi+1\right)}
\]
\[
\propto IG\{\gamma_j|(1/\xi + 1)/2, m_j\} \cdot g(\gamma_j), \tag{A.4}
\]

where \( m_j = \beta_j^2/(2\sigma^2) \) and \( g(\gamma_j) = (\xi + \tau \cdot \gamma_j^{-1/2})^{-\left(1/\xi+1\right)} \). Note that the function \( u_j = g(\gamma_j) \) is increasing on \((0, \infty)\), and its inverse function is \( \gamma_j = g^{-1}(u_j) = [\tau/\{u_j^{-\left(\xi/(1+\xi)\right)} - \xi\}]^2 \). Now, consider a density, \( \pi(\gamma_j, u_j|\cdot) \propto IG\{\gamma_j|(1/\xi + 1)/2, m_j\} \cdot \mathcal{I}_{(0,g(\gamma_j))}(u_j) \). Then we can show that \( \int \pi(\gamma_j, u_j|\cdot) du_j = \pi(\gamma_j|\cdot) \), which means that \( \pi(\gamma_j, u_j|\cdot) \) is a valid parameter expansion of (A.4). Actual sampling is executed on \( \pi(\gamma_j, u_j|\cdot) \) using the Gibbs sampler: (i) \( u_j|\gamma_j, \cdot \sim \pi(u_j|\gamma_j, \cdot) = U(0, g(\gamma_j)) \) and (ii) \( \gamma_j|u_j, \cdot \sim \pi(\gamma_j|u_j, \cdot) = IG\{\gamma_j|(1/\xi + 1)/2, m_j\} \cdot \mathcal{I}_{(g^{-1}(u_j),0)}(\gamma_j) \).

After the Gibbs sampling, transform back to \( \lambda_j = \sqrt{\gamma_j} \).

To implement the slice sampler in the Step 4, note from (A.2):

\[
\pi(\tau|\cdot) \propto IG(\tau|1, 1) \cdot \prod_{j=1}^{p} g_j(\tau), \tag{A.5}
\]

where \( g_j(\tau) = (\tau + \xi \lambda_j)^{-\left(1/\xi+1\right)}, j = 1, \cdots, p \). Note that \( p \)-functions \( v_j = g_j(\tau), j = 1, \cdots, p \), are decreasing on \((0, \infty)\), and their inverse functions are \( \tau = g_j^{-1}(v_j) = v_j^{-\left(\xi/(1+\xi)\right)} - \xi \lambda_j, j = 1, \cdots, p \). Now, consider a density: \( \pi(\tau, v_1, \cdots, v_p|\cdot) \propto IG(\tau|1, 1) \cdot \prod_{j=1}^{p} \mathcal{I}_{(0,g_j(\tau))}(v_j) \). Then we have \( \int \cdots \int \pi(\tau, v_1, \cdots, v_p|\cdot) dv_1 \cdots dv_p = \pi(\tau|\cdot) \) and hence \( \pi(\tau, v_1, \cdots, v_p|\cdot) \) is a valid parameter expansion of (A.5). Actual sampling is executed on \( \pi(\tau, v_1, \cdots, v_p|\cdot) \) using the Gibbs
A.3 Summary of the Hill estimator

We briefly explain the Hill estimator which plays a central role in hyper-parameter specification of the \( \mu \). For notational coherence, we use the Greek letter \( \lambda \) to describe a random quantity. Suppose that \( \lambda = (\lambda_1, \cdots, \lambda_p)^\top \in (0, \infty)^p \) is \( p \)-dimensional random variables from a strongly stationary process whose marginal distribution is \( F \) such that its tail distribution is regularly varying with the tail-index \( 1/\xi \) with \( \xi > 0 \) (hence, the corresponding shape parameter is \( \xi \)). By the Karamata’s characterization theorem [18], the tail distribution (survival function) is described as \( \bar{F}(\lambda) = 1 - F(\lambda) = L(\lambda) \cdot \lambda^{-1/\xi} \) for some \( \xi > 0 \) where \( L \) is a slowly varying function [8, 31]. Denote its order statistics with \( \lambda_{(1)} \geq \cdots \geq \lambda_{(p)} \).

The Hill estimator [13] is a well-known estimator of shape parameter \( \xi \) principled on the peaks-over-threshold methods. Hill estimator is obtained from the \( k \) upper order statistics:

\[
\hat{\xi}_k(\lambda) = \frac{1}{k-1} \sum_{j=1}^{k-1} \log \left( \frac{\lambda_{(j)}}{\lambda_{(k)}} \right), \quad \text{for } 2 \leq k \leq p. \tag{A.7}
\]

It is known that the Hill estimator (A.7) is a consistent estimator for \( \xi \), i.e., \( \hat{\xi}_k(\lambda) \to \xi \) in probability, if \( p \to \infty, k \to \infty, \) and \( k/p \to 0 \) [8, 11, 31]. Empirically it is known that the Hill estimator may work effectively when \( F \) is of a Pareto type [8, 19]. (See Fig 1 in [8].)

Suppose we have \( p \) number of observations \( \lambda = (\lambda_1, \cdots, \lambda_p)^\top \), possibly generated from the aforementioned heavy distribution \( F \). In practice, the Hill estimator is used as follows. First,
calculate the estimator \( \hat{\xi}_k(\lambda) \) at each integer \( k \in \{2, \cdots, p\} \), and then plot the ordered pairs \( \{(k, \hat{\xi}_k(\lambda))\}_{k=2}^p \): the resulting plot is called the Hill plot (See the Figure 6.4.3 of [11]). Then, select value(s) from the set of Hill estimators \( \{\hat{\xi}_k(\lambda)\}_{k=2}^p \) which are stable (roughly constant) with respect to \( k \): then, such stable value(s) are regarded as reasonable estimate(s) for the shape parameter \( \xi \) [8]. Typically, the Hill plot may display high variability when \( k \) is close to 2 or \( p \). As a practical remedy, one may disregard the first or last few of the estimates: the values \( \hat{\xi}_k(\lambda) \) that are evaluated at integers \( k \in \{k_L, \cdots, k_U\} \), \( 2 < k_L < k_U < p \), are considered to be monitored where the integers \( k_L \) and \( k_U \) are designated by user.

A.4 Hyper-parameter specification of \( \mu \) and \( \rho^2 \)

Suppose we are at the Step 5 of the \( s \)-th iteration of the Gibbs sampler described in Subsection A.1. At this moment, we have already acquired posterior realizations, \( \lambda^{(s+1)} = (\lambda_1^{(s+1)}, \cdots, \lambda_p^{(s+1)})^\top \) and \( \tau^{(s+1)} \), that had been sampled from the previous steps, Step 3 and Step 4, respectively.

By treating the indicator \( I_{(1/2, \infty)}(\xi) \) in (A.3) as a part of likelihood, we consider sampling \( \xi^{(s+1)} \) from the density;

\[
\xi^{(s+1)} \sim \pi(\xi|\cdot) = \pi(\xi|\lambda^{(s+1)}, \tau^{(s+1)}) \propto \mathcal{L}(\xi) \cdot \log \mathcal{N}_1(\xi|\mu, \rho^2), \quad \mathcal{L}(\xi) = \mathcal{V}_p(\xi) I_{(1/2, \infty)}(\xi).
\]

(A.8)

Henceforth, the basic idea is to strictly obey the philosophy of Gibbs sampler: as long as we are to sample \( \xi^{(s+1)} \sim \pi(\xi|\cdot) \) (A.8), every latent variables except for the target variable \( \xi \) are treated as observed variables, including \( \lambda^{(s+1)} \) and \( \tau^{(s+1)} \).

To start with, we choose a small value of the hyper-parameter \( \rho^2 \) so that the prior part in (A.8), that is, \( \pi(\xi) = \log \mathcal{N}_1(\xi|\mu, \rho^2) \), is highly concentrated around its prior mean \( \mathbb{E}[\xi] = \exp(\mu + \rho^2/2) \approx \exp(\mu) \). That way, a future state \( \xi^{(s+1)} \) is highly probable to be sampled around the value \( \exp(\mu) \), leading to an approximate relationship between the future state \( \xi^{(s+1)} \) and hyper-parameter \( \mu \), described by \( \xi^{(s+1)} \approx \exp(\mu) \), or equivalently, \( \mu \approx \log(\xi^{(s+1)}) \). This approximation will be utilized shortly later. Throughout this paper, we use \( \rho^2 = 0.001 \) as the default value for \( \rho^2 \).

Now, we are in a position to describe how to calibrate the hyper-parameter \( \mu \) via the Hill
estimator (A.7). We start with ordering the realizations of the local-scale parameters $\lambda^{(s+1)} = (\lambda^{(s+1)}_1, \ldots, \lambda^{(s+1)}_p)^\top$ to obtain $\lambda^{(s+1)}_1 \geq \cdots \geq \lambda^{(s+1)}_p$. The Hill estimator based on $\lambda^{(s+1)}$ is then

$$
\hat{\xi}_k(\lambda^{(s+1)}) = \frac{1}{k - 1} \sum_{j=1}^{k-1} \log \left( \frac{\lambda^{(s+1)}_j}{\lambda^{(s+1)}_k} \right), \quad \text{for } k_L \leq k \leq k_U,
$$

where $k_L = \lfloor p/10 \rfloor$ and $k_U = \lfloor 9p/10 \rfloor$, with $\lfloor \cdot \rfloor$ is the floor function, where $p$ is the number of covariates. In high-dimensional setting, the number of the elements of the set $\{k_L, \ldots, k_U\} = \{\lfloor p/10 \rfloor, \ldots, \lfloor 9p/10 \rfloor\} \subset \{2, \ldots, p\}$ is still large, approximately, $\lfloor 4p/5 \rfloor$, enough to retain the consistency of the Hill estimator. Note that estimates in (A.9) depend on $k$. To eliminate dependency on $k$, first, we average out the Hill estimators (A.9) over $k$, and then use the approximation $\mu \approx \log \xi^{(s+1)}$, to get:

$$
\hat{\mu}(\lambda^{(s+1)}) = \log \left\{ \hat{\xi}(\lambda^{(s+1)}) \right\} = \log \left\{ \frac{1}{k_U - k_L + 1} \sum_{k=k_L}^{k_U} \hat{\xi}_k(\lambda^{(s+1)}) \right\}. \quad (A.10)
$$

Note that the value of $\hat{\mu}(\lambda^{(s+1)})$ (A.10) changes at every cycle of the Gibbs sampler, and tuned by $\lambda^{(s+1)}$ through the Hill estimator. In other words, $\hat{\mu}(\lambda^{(s+1)})$ can be thought as a calibrated hyper-parameter adapted via the $p$ local-scale realizations $\lambda^{(s+1)}$. By replacing $\mu$ with $\hat{\mu}(\lambda^{(s+1)})$ and substituting $\rho^2 = 0.001$ in the full conditional posterior density $\pi(\xi|\cdot)$ (A.8), the Step 5 within the Gibbs sampler is tuning-free.

Finally, we explain how to sample from the density $\pi(\xi|\cdot)$ (A.8). For that, first, use a change of variable $\eta = \log \xi$ and sample from

$$
\eta^{(s+1)} \sim \pi(\eta|\cdot) = \pi(\eta|\lambda^{(s+1)}, \tau^{(s+1)}) \propto \mathcal{L}(\eta) \cdot \mathcal{N}_1(\eta|\hat{\mu}(\lambda^{(s+1)}), \rho^2 = 0.001), \quad (A.11)
$$

where $\mathcal{L}(\eta) = \psi_p(e^\eta) \mathcal{I}_{(\log 1/2, \infty)}(\eta) = \{(\Gamma(p/e^\eta + 1))^{-1} \pi^{p/2} \prod_{j=1}^p (\tau^{(s+1)}_j + e^\eta \lambda^{(s+1)}_j)^{-1/(e^\eta + 1)}\} \mathcal{I}_{(\log 1/2, \infty)}(\eta)$. Once we obtain a sample $\eta^{(s+1)} \sim \pi(\eta|\cdot)$, then $\xi^{(s+1)} \sim \pi(\xi|\cdot)$ is obtained via the inverse transformation through $\xi^{(s+1)} = \exp \eta^{(s+1)}$.

We use the elliptical slice sampler (ESS) [26] to sample from $\eta^{(s+1)} \sim \pi(\eta|\cdot)$ (A.11) that
exploits the Gaussian prior measure. Conceptually, ESS and Metropolis-Hastings (MH) algorithm are similar in that both comprises two steps: proposal step and criterion step. A difference between the two algorithms arises in the criterion step. If a new candidate does not pass the criterion, then MH takes the current state as the next state: whereas, ESS re-proposes a new candidate until rejection does not take place, rendering the algorithm rejection-free. Further information for ESS is referred to the original paper [26]. By adopting a jargon from their paper, as the calibrated $\mu$, $\hat{\mu}(\lambda^{(s+1)})$, is positioned at the center of an ellipse [26, 28], hence, we refer to the following Algorithm 1 to implement the Step 5 as elliptical slice sampler centered by the Hill estimator.
Algorithm 1: Elliptical slice sampler centered by the Hill estimator

Circumstance:  At the Step 5 of the s-th iteration of the Gibbs sampler in Subsection A.1.

Input:  Current state $\xi^{(s)}$, and posterior realizations $\lambda^{(s+1)}$ and $\tau^{(s+1)}$ obtained from the Step 3 and Step 4, respectively.

Output:  A new state $\xi^{(s+1)}$.

1. Calibration of $\mu$: obtain $\hat{\mu}(\lambda^{(s+1)}) = \log \{\hat{\xi}(\lambda^{(s+1)})\}$ (A.10).
2. Variable change ($\eta = \log \xi$): $\eta^{(s)} = \log \xi^{(s)}$.
3. Implement elliptical slice sampler to (A.11);
   a. Choose ellipse centered by the Hill estimator: $\nu \sim N_1(\hat{\mu}(\lambda^{(s+1)}), \rho^2 = 0.001)$.
   b. Define a criterion function: $\alpha(\eta, \eta^{(s)}) = \min \{\mathcal{L}(\eta)/\mathcal{L}(\eta^{(s)}), 1\} : (\log 1/2, \infty) \to [0, 1]$, where $\mathcal{L}(\eta) = [(\Gamma(p/e^{\eta} + 1))^{-1}\pi^{p/2} \prod_{j=1}^{p}(\tau^{(s+1)} + e^{\eta}\lambda_j^{(s+1)})^{-(1/e^{\eta}+1)}] \cdot \mathcal{I}(\log 1/2, \infty)(\eta)$.
   c. Choose a threshold and fix: $u \sim U[0, 1]$.
   d. Draw an initial proposal $\eta^*$:
      $$\theta \sim U(-\pi, \pi)$$
      $$\eta^* = \{\eta^{(s)} - \hat{\mu}(\lambda^{(s+1)})\} \cos \theta + \{\nu - \hat{\mu}(\lambda^{(s+1)})\} \sin \theta + \hat{\mu}(\lambda^{(s+1)})$$
   e. if ($u < \alpha(\eta^*, \eta^{(s)})$) { $\eta^{(s+1)} = \eta^*$ } else {
      Define a bracket: $(\theta_{\min}, \theta_{\max}) = (-\pi, \pi)$.
      while ($u \geq \alpha(\eta^*, \eta^{(s)})$) {
         Shrink the bracket and try a new point:
         if ($\theta > 0$) $\theta_{\max} = \theta$ else $\theta_{\min} = \theta$
         $\theta \sim U(\theta_{\min}, \theta_{\max})$
         $\eta^* = \{\eta^{(s)} - \hat{\mu}(\lambda^{(s+1)})\} \cos \theta + \{\nu - \hat{\mu}(\lambda^{(s+1)})\} \sin \theta + \hat{\mu}(\lambda^{(s+1)})$
      }
   }
   $\eta^{(s+1)} = \eta^*$

4. Variable change ($\xi = e^\eta$): $\xi^{(s+1)} = \exp \eta^{(s+1)}$. 

S.2
S.1. EXAMPLE– BREAST CANCER DATA

S.1.1 Breast cancer data

Breast cancer data is downloadable from R package TCGA2STAT. The breast cancer data is composed of a health response vector and a design matrix, $(y, X) \in \mathbb{R}^n \times \mathbb{R}^{n \times p}$, obtained from $n = 729$ breast cancer patients and $p = 3,250$ genes. The $i$-th response value $y_i \in \mathbb{R}$, $i = 1, \cdots, n$, is the log-transformed overall survival (OS) time of the $i$-th subject such that all responses $\{y_i\}_{i=1}^n$ were quality assessed, integrated and processed with the help from disease experts and TCGA Biospecimen Core Resource [21]. Following a guideline from [21], subjects who have moderately long OS are considered in our study. A detailed clinical information of the dataset can be found in [21]. The minimum, mean, and maximum of OS are 84 days, 1,000 days (2.7 years), and 8,605 days (23 years), respectively. $X$ is a column-standardized design matrix such that the $ij$-th element $x_{ij}$ represents gene expression levels of the $j$-th gene obtained from the $i$-th patient.

National Cancer Institute (NCI) defines OS as the length of time from either the date of diagnosis or the start of treatment for a disease, such as cancer, that patients diagnosed with the disease are still alive. In clinical cancer trials, measuring the OS is one way to see how well a new treatment works. Therefore, the value of OS is an indirect evidence of measurement about how strong the immune system of the patients. The histogram of $\{y_i\}_{i=1}^n$ and its Q-Q plot are displayed on the Figure 1. The Q-Q plot shows small deviation of the responses $\{y_i\}_{i=1}^n$ from normality.

After centering the response vector $y$ to avoid introducing an intercept term, our goal is to estimate the coefficients $\beta$ from the high-dimensional regression ((1) in the main paper). The primary objective of this study is then to discover two categories of small number of interesting genes: (i) beneficial genes that may enhance the immune system of breast cancer patients (positive sign of $\beta_j$), and (ii) risky genes that may undermine the immune system of breast cancer patients (negative sign of $\beta_j$). Such beneficial genes can be further investigated by genetic scientist in immunotherapy, whereby scientists are attempting to harness the body’s own immune system to fight and prevent malignancies [32].
S.1.2 Breast cancer data analysis via the Horseshoe

As similar to the prostate cancer data study in Section 6 in the main paper, we shall compare the performance of the Horseshoe ((12) – (13) in the main paper) and GLT prior (6) – (9) in the main paper) by applying to nested datasets to see how the two priors behave over different (unknown) sparse regimes. We constructed four nested datasets, $B_1 = (y, X(:, 1 : 500))$, $B_2 = (y, X(:, 1 : 1000))$, $B_3 = (y, X(:, 1 : 2000))$, and $B_4 = (y, X(:, 1 : 3250) = X)$. It holds the subset inclusion $B_1 \subset B_2 \subset B_3 \subset B_4$. The dataset $B_4$ is the full dataset, and $B$ stands for breast. The four datasets share the same response vector $y \in \mathbb{R}^n (n = 729)$, but the number of genes used in the design matrix are different; $B_1, B_2, B_3$, and $B_4$ consider 500, 1,000, 2,000, and 3,250 genes, respectively.

In general, a sparse linear model (1) applied to an actual gene expression datasets needs to overcome intrinsic colinearity in the high-dimensional gene matrix $X \in \mathbb{R}^{n \times p}$. Figure 2 displays the stacked histograms of the column-wise correlations obtained from the design matrices from the four datasets. Left and right histograms are obtained by confining the correlations to intervals $[-0.6, -0.4]$ and $[0.8, 1]$, respectively. We note from Figure 2 that as the number of genes used increases, the genome-wise correlations get substantially intensified, elucidating a significant increase in genetic-association [10, 14, 23]. (As the correlation quantifies a linear relationship between a “pair” of genes in terms of the gene expression, the panels in the Figure 2 only show a one
Given dataset $B_l$ ($l = 1, 2, 3, 4$), we implemented the Horseshoe by using `horseshoe(y=y, X=X, method.tau="halfCauchy", method.sigma="Jeffreys", burn = 10000, nmc=10000, thin=100)` where $y = y$ and $X = X$. We report the results of the posterior inference by displaying the gene ranking plot, where the coefficients in $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ are ranked based on the absolute values of the posterior mean $\{\hat{\beta}_j\}_{j=1}^p$, ordered from largest to smallest.

Figure 3 displays the top 50 genes obtained by using the Horseshoe for each dataset $B_l$, $l = 1, 2, 3, 4$. Table S.2 summarizes top 10 genes along with their names, and directions which have been taken from the signs of the posterior means. The results are reasonable for $B_1$ and $B_2$. However, the Horseshoe produces the collapsed inference where the posterior means numerically become the null vector ($\hat{\beta} \approx 0$) when applied to $B_3$ and $B_4$ as similar to what we have observed in the prostate cancer data example (refer to the Figure 5 in the main paper). Based on the Table S.2, it turns out that the genes NGEF and FAM138F are found to be the most significant for the datasets $B_1$ and $B_2$, respectively, and both genes have negative effects on the response OS. Figure 3 can be used for uncertainty quantification associated with the coefficients.
Figure 3: Top 50 genes obtained by the Horseshoe: $B_1$ (top-left panel), $B_2$ (top-right panel), $B_3$ (bottom-left panel), and $B_4$ (bottom-right panel). The dots \( \bullet \) and vertical bars represent the posterior means and 95\% credible intervals, respectively. The colors blue and red represent plus and negative signs of posterior mean of $\beta_j$, respectively. Posterior means of $\tau$ corresponding to the four datasets are $0.10839 (B_1)$, $0.06145 (B_2)$, $3.65 \cdot 10^{-8} (B_3)$, and $3.19 \cdot 10^{-9} (B_4)$, respectively.

Table S.2: Top 10 interesting genes selected by the Horseshoe when applied to $B_l$, $l = 1, 2, 3, 4$

|   | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     | 10    |
|---|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $B_1$ | NGEF\((-\)) | PLN\((-\)) | C3orf59\(+(+)\) | C21orf63\(+(+)\) | LOC100130331\((-\)) | FCGR2A\((-\)) | HES4\(+(+)\) | BCAP31\((-\)) | GSTM1\(+(+)\) | TOB2\((-\)) |
| $B_2$ | FAM138F\((-\)) | SLC39A4\((-\)) | PLN\((-\)) | NGEF\((-\)) | PCGF5\((+\)) | NA | NA | NA | NA | NA |
| $B_3$ | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |
| $B_4$ | NA | NA | NA | NA | NA | NA | NA | NA | NA | NA |

NOTE: Contents of table is (gene name, direction). Genes with positive sign \((+)\) may enhance the immune system of patients; however, genes with minus \((-\)) may damage the immune system of patients. When the Horseshoe is applied to the datasets $B_3$ and $B_4$, genes are unranked because the Horseshoe estimator collapsed (the posterior mean numerically becomes the null vector, $\hat{\beta} \approx 0.$)
S.1.3 Breast cancer data analysis via the GLT prior

The GLT prior is applied to the same four breast cancer data $B_l$, $l = 1, 2, 3, 4$, constructed in the previous subsection. Recall that the Horseshoe collapsed when applied to $B_l$, $l = 3, 4$: see the bottom panels in the Figure 3. The Figure 4 and the Table S.3 show the top 50 gene ranking plots and top 10 interesting genes obtained by using the GLT prior when applied to the four breast cancer datasets. Posterior means of the shape parameters $\xi$ corresponding to the four datasets are $2.188 (B_1)$, $2.230 (B_2)$, $2.382 (B_3)$, and $2.922 (B_4)$, respectively. The monotonicity suggests that it holds the *tali-adaptive shrinkage property* of the GLT prior.

Table S.4 summarizes the top 13 interesting genes selected by the GLT prior when applied to the full breast cancer dataset $B_4$, and some references from the literature on oncology and genetics. The GLT prior discovered LOC150776 that has been less studied in the literature. As the direction of LOC150776 is positive (+), an over expression of LOC150776 may enhance the immune system of breast cancer patients. Interestingly, the GLT prior identified the famous superman gene BHLHE41: it is known that the genetic variant of BHLHE41 provides a greater resistance to the effects of sleep deprivation, possibly enhancing the immune system [29].
Figure 4: Top 50 gene ranking plots obtained by the GLT prior: $B_1$ (top-left panel), $B_2$ (top-right panel), $B_3$ (bottom-left panel), and $B_4$ (bottom-right panel). Posterior means of $(\tau, \xi)$ corresponding to the four datasets are $(0.00436, 2.188)$ ($B_1$), $(0.00221, 2.230)$ ($B_2$), $(0.00135, 2.382)$ ($B_3$), and $(0.00135, 2.922)$ ($B_4$), respectively.

Table S.3: Top 10 interesting genes selected by the GLT prior when applied to $B_l$, $l = 1, 2, 3, 4$

|   | 1    | 2    | 3    | 4    | 5    |
|---|------|------|------|------|------|
| $B_1$ | NGEF(−) | C2orf63(+) | PLN(−) | C3orf59(+) | FCGR2A(−) |
| $B_2$ | FAM138F(−) | SLC39A4(−) | NGEF(−) | PCGF5(+) | PLN(−) |
| $B_3$ | FAM138F(−) | SLC39A4(−) | NGEF(−) | PLN(−) | COL7A1(−) |
| $B_4$ | FAM138F(−) | NSUN4(−) | COL7A1(−) | LOC150776(+) | NGEF(−) |

|   | 6    | 7    | 8    | 9    | 10   |
|---|------|------|------|------|------|
| $B_1$ | BCAP31(−) | GSTM1(+) | LOC100130331(−) | TOB2(−) | ABCA17P(+) |
| $B_2$ | FCGR2A(−) | CRHR1(+) | TOB2(−) | GSTM1(+) | LOC150776(+) |
| $B_3$ | CRHR1(+) | FCGR2A(−) | RPLP1(+) | HES4(+) | TOB2(−) |
| $B_4$ | SMCHD1(+) | RPLP1(+) | HES4(+) | SLC37A2(−) | SLC39A4(−) |
Table S.4: Top 13 interesting genes selected by the GLT prior when applied to $B_4$

| Rank | Gene (direction) | Note | References |
|------|------------------|------|------------|
| 1    | FAM138F(−)       | Increasing a risk of breast and ovarian cancer | [12, 33] |
| 2    | NSUN4(−)         | Related with ovarian and prostate cancer | [17] [H] |
| 3    | COL7A1(−)        | Related with cell migration (metastasis) | [39] [H] |
| 4    | LOC150776(+)     | Less studied in oncology and genetics | |
| 5    | NGEF(−)          | Related with obesity-related diseases | [37] [H] |
| 6    | SMCHD1(+)        | Important in regulation | [16] [H] |
| 7    | RPLP1(+)         | Important in protein synthesis | [9] [H] |
| 8    | HES4(+)          | Gene knockdown increases a brain disease | [1] [H] |
| 9    | SLC37A2(−)       | Negatively related with survival probability | [H] |
| 10   | SLC39A4(−)       | Negatively related with survival probability | [16] [H] |
| 11   | MFRP(−)          | Related with ovarian cancer | |
| 12   | ARSA(+)          | Positively related with survival probability | [H] |
| 13   | BHLHE41(+)       | High recovery from fatigue or short sleep | [29] [H] |

NOTE: [H] is linked to The Human Protein Atlas (https://www.proteinatlas.org).

S.2. SIMULATION STUDIES WITH VARIANTS OF THE HORSESHOE

S.2.1 Outline

Consider a high-dimensional linear regression (1):

$$y = X\beta + \sigma \epsilon, \quad \epsilon \sim \mathcal{N}_n(0, I_n), \quad X \in \mathbb{R}^{n \times p}, \quad p \gg n,$$

where the $p$ coefficients $\beta = (\beta_1, \cdots, \beta_p)^\top \in \mathbb{R}^p$ and the error variance $\sigma^2$ is unknown. Our particular interest is on the estimation of the $\beta$.

Here, we compare the performance of the Horseshoe [5] (given as the hierarchy (12) – (13) in the main paper) and the GLT prior (given as the hierarchy (6) – (9) in the main paper), and three variant versions of the Horseshoe, namely truncated horseshoe [35], horseshoe-plus [3], and regularized horseshoe [30].

We conducted replicated study under the high-dimensional regression with $n = 100$ responses and $p = 500$ covariates. Simulation environments coincides with the three scenarios described in Subsection 8.1 in the main paper. (Remember that the Scenarios 1, 2, and 3 are characterized by varied sparsity level, varied correlation $\varrho$ associated with design matrix, and varied signal-to-noise
(SNR) ratio, respectively.)

For evaluation criteria, as same with how we reported in the Section 8.1 in the main paper, provided the truth \( \beta_0 = (\beta_{0,1}, \cdots, \beta_{0,q}, \beta_{0,q+1}, \cdots, \beta_{0,p})^T \) specified by \( \beta_{0,1} = \cdots = \beta_{0,q} = 1 \) (\( q \) unit signals) and \( \beta_{0,q+1} = \cdots = \beta_{0,p} = 0 \) (\( p - q \) noises), we separately report median of the following quantities obtained by 50 replicates:

\[
MSE = \frac{1}{p} \sum_{j=1}^{p} (\widehat{\beta}_j - \beta_{0,j})^2, \quad MSE_S = \frac{1}{q} \sum_{j=1}^{q} (\widehat{\beta}_j - 1)^2, \quad \text{and} \quad MSE_N = \frac{1}{p-q} \sum_{j=q+1}^{p} (\widehat{\beta}_j)^2.
\]

The MSE measures overall accuracy of estimation for the coefficients induced by a prior, which can be dissected by two components: (1) MSE for signal part (MSE\(_S\)) measuring signal recovery ability and (2) MSE for noise parts (MSE\(_N\)) measuring noise shrinking ability. We emphasize that the Bayes estimate collapses (that is, the posterior means \( \widehat{\beta}_j \) are nearly zeros), then MSE\(_S\) and MSE\(_N\) will be close to 1 and 0, respectively. In this circumstance, the total MSE is not a reasonable evaluation criteria.

**S.2.2 Three variants of the Horseshoe**

The followings are hierarchies of the truncated horseshoe [35], horseshoe-plus [3], and regularized horseshoe [30]:

**Truncated horseshoe** [35].

\[
\beta_j | \lambda_j, \tau, \sigma^2 \sim N_1(0, \lambda_j^2 \tau^2 \sigma^2), \quad \lambda_j \sim C^+(0,1), \quad \tau \sim T\mathcal{C}^+(0,1)_{(1/p, \infty)}, \quad (j = 1, \cdots, p).
\]

The \( T\mathcal{C}^+(0,1)_{(1/p, \infty)} \) is the unit-scaled half-Cauchy distribution truncated from below by \( 1/p \). The \( \text{R} \) function \texttt{horseshoe} within the \( \text{R} \) package \texttt{horseshoe} provides an option to use this setting by specifying \texttt{method.tau = "truncatedCauchy"}.

**Horseshoe-plus** [3].

\[
\beta_j | \lambda_j, \sigma^2 \sim N_1(0, \lambda_j^2 \sigma^2), \quad \lambda_j | \eta_j, \tau \sim \mathcal{C}^+(0, \eta_j \tau), \quad \eta_j, \tau \sim \mathcal{C}^+(0,1), \quad (j = 1, \cdots, p).
\]
Note that the Horseshoe-plus is characterized by a further half-Cauchy mixing variable \( \eta_j \) embedded to the local-scales \( \lambda_j \).

**Regularized horseshoe** [30].

\[
\beta_j | \bar{\lambda}_j, \tau \sim N_1(0, \tau^2 \bar{\lambda}_j^2), \quad \bar{\lambda}_j^2 = \frac{c^2 \lambda_j^2}{c^2 + \tau^2 \lambda_j^2}, \quad (j = 1, \cdots, p),
\]

\[
\lambda_j, \tau \sim C^+(0, 1), \quad c^2 \sim IG(\nu/2, \nu s^2/2), \quad (j = 1, \cdots, p),
\]

where \( \nu, s^2 > 0 \) are hyper-parameters: we shall simply fix them to be 1.

S.2.3 **Simulation results**

Figure 5 displays the simulation results: **Scenario 1** (top three panels); **Scenario 2** (middle three panels); and **Scenario 3** (bottom three panels). The followings are summaries based on the results:

1. Under the **Scenario 1**, we see that the truncated horseshoe prior [36] suffers from the similar collapse observed in the Horseshoe [5] when sparsity level is larger than certain threshold. The total MSE does not bring out this phenomenon, motivating separate reports for the MSE\(_S\) and MSE\(_N\).

2. Under the **Scenario 1** with ultra sparsity regime (where the sparsity level \( q/p \) is between 0.002 and 0.024), all considered priors perform reasonably well, while the signal recovery ability of the GLT prior is marginally getting better as the sparsity level increases.

3. Under the **Scenario 1** with moderate sparsity regime (where the sparsity level \( q/p \) is between 0.034 and 1), (i) the regularized horseshoe [30] outperforms others in terms of MSE\(_S\), while (ii) the GLT prior outperforms others in terms of MSE\(_N\).

4. Under the **Scenario 2**, the Horseshoe [5] and the truncated horseshoe [36] outperform other priors in terms of MSE, while the GLT prior outperforms in terms of MSE\(_S\).

5. Under the **Scenario 3**, the Horseshoe [5] and the truncated horseshoe [36] outperform other priors in terms of MSE, while the GLT prior outperforms in terms of MSE\(_S\) when SNR is 2.
Figure 5: Simulation results under the three scenarios: Scenario 1 (top panels); Scenario 2 (middle panels); and Scenario 3 (bottom panels). Metrics measured are MSE (left panels), MSE_S (center panels), and MSE_N (right panels).
S.3. CURVE FITTING STUDY

S.3.1 Simulated curves

Consider two functions $f$ on domain $D$ from what data is generated: (i) sinc curve $f(x) = \text{sinc}(x) = \frac{\sin x}{x}$ on $D = (-20, 20)$, and (ii) flat curve $f(x) = (5x - 3)^3 \cdot \mathbf{1}(x > 3/5)$ on $D = (0, 1)$. To fabricate perturbed functional responses, first, we uniformly sampled $n$-inputs $\{x_i\}_{i=1}^n$ from domain $D$, and then let $y_i = f(x_i) + \sigma_0 \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, 1)$, $i = 1, \ldots, n$, with $\sigma_0 = 0.15$, to generate $n$-pair $\{(y_i, x_i)\}_{i=1}^n$. Goal is to infer the true function $f$ out of the $n$ noised observations.

To estimate $f$ given the $n$-pair $\{(y_i, x_i)\}_{i=1}^n$, we use the sparse Gaussian kernel regression [4, 34] where the true unknown function $f$ is approximated by a kernel-based function:

$$y_i = f_n(x_i) + \sigma \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1), \quad i = 1, \ldots, n$$

$$f_n(\cdot) = \alpha + \sum_{j=1}^n \beta_j K(\cdot, x_j) : D \rightarrow \mathbb{R},$$

where $\alpha \in \mathbb{R}$ is intercept term, and $n$ weights $\beta = (\beta_1, \cdots, \beta_n)^T \in \mathbb{R}^n$ are assumed to be sparse. We use the Gaussian kernel for $K$ [4].

We use the Horseshoe (given as the hierarchy (12) – (13) in the main paper) and the GLT prior (given as the hierarchy (6) – (9) in the main paper) to impose shrinkage on the $n$ coefficients $\beta$. For the intercept $\alpha$, we use the flat prior for $\alpha$ [22]. For each test curve, we generated $n = 100$ observations (that is, $\{(y_i, x_i)\}_{i=1}^{100}$), and report the median of average mean squared error (AMSE) [38] obtained from 100 replications. AMSE is defined by $\sum_{i=1}^n (\hat{f}_n(x_i) - f(x_i))^2 / n$, where $\hat{f}_n(x) = \mathbb{E}[\alpha + \sum_{j=1}^n \beta_j K(x, x_j) | y]$ is a posterior mean of $f_n(x)$ at $x$.

For the (i) sinc test curve, the median AMSE obtained by the Horseshoe and the GLT priors are 0.00393 and 0.00385, respectively. For the (ii) flat test curve, the median of AMSE obtained by using the Horseshoe and the GLT prior are 0.00490 and 0.00382, respectively. See Figure 6 for one of the 100 replicates.
Figure 6: Two simulated curves fitted by the sparse Gaussian kernel regression. The red dot and red dotted curve represent observation and the truth $f$. The black curve and blue dotted curve represent the posterior mean of $f_n(x)$ at $x$ obtained by using the GLT prior and the Horseshoe, respectively. The shaded region depicts the pointwise 95% credible interval obtained by using the GLT prior.

S.3.2 Real curves

The sparse Gaussian kernel regression is applied to four example curves: circadian rhythm curve of gene expression of PER2 from colon tissue, light-curve from an eclipsing binary star system, fossil data, and LIDAR data. The number of observations for the four data are 100, 377, 106, and 221, respectively. The circadian rhythm data and light-curve data can be obtained from the website http://circadb.hogeneschlab.org and https://www.eso.org, respectively. The fossil data and the LIDAR data can be downloaded from R package SemiPar. See Figure 7 for the results: the results are virtually indistinguishable.
Figure 7: Four real curves fitted by the sparse Gaussian kernel regression.
S.4. PROPERTIES OF THE HORSESHOE

S.4.1 Prior analysis of the Horseshoe

Lemma 1 (Marginal density and random shrinkage coefficient of the Horseshoe).

(a) Assume $\beta | \lambda, \tau \sim N_1(0, \tau^2 \lambda^2)$, $\lambda \sim C^+(0, 1)$, and $\tau > 0$. Then:

$$\pi_{HS}(\beta | \tau) = \int N_1(\beta | 0, \tau^2 \lambda^2) \pi(\lambda) d\lambda = K_{HS} e^{Z_{HS}(\beta)} E_1\{Z_{HS}(\beta)\}, \quad (S.1)$$

where $K_{HS} = 1/(\tau^2 \pi^{3/2})$ and $Z_{HS}(\beta) = \beta^2/(2\tau^2)$. $E_1(x) = \int_1^\infty e^{-xt} t^{-1} dt$, $x \in \mathbb{R}$, is the exponential integral function.

(b) Assume $\lambda \sim C^+(0, 1)$, $\kappa = 1/(1 + \tau^2 \lambda^2) \in (0, 1)$, and $\tau > 0$. Then:

$$\pi_{HS}(\kappa | \tau) = \frac{\tau \cdot \kappa^{-1/2}(1 - \kappa)^{-1/2}}{1 - (1 - \tau^2) \kappa}. \quad (S.2)$$

S.4.2 Proof– Restricted tail-heaviness of Horseshoe

Function $e^x E_1(x)$ satisfies tight upper and lower bounds [5]:

$$\frac{1}{2} \cdot \log \left( \frac{x + 2}{x} \right) < e^x E_1(x) < \log \left( \frac{x + 1}{x} \right), \quad x > 0. \quad (S.3)$$

Replacing $x$ with $Z_{HS}(\beta) = \beta^2/(2\tau^2)$ and multiplying $K_{HS} = 1/(\tau^2 \pi^{3/2})$ to the both sides of the inequalities (S.3) lead to;

$$L(\beta) < \pi_{HS}(\beta | \tau) < U(\beta), \quad \beta \in \mathbb{R}, \quad \tau > 0, \quad (S.4)$$

where $L(\beta) = (K_{HS}/2) \cdot \log \{(Z_{HS}(\beta) + 2)/Z_{HS}(\beta)\}$ and $U(\beta) = K_{HS} \cdot \log \{(Z_{HS}(\beta) + 1)/Z_{HS}(\beta)\}$.

Now, denote the tail (survival) function of the random variable $\beta | \tau$ given $\tau > 0$ by $\bar{F}_{HS}(\beta | \tau) = 1 - F_{HS}(\beta | \tau)$: then, it holds $(d/d\beta)F_{HS}(\beta | \tau) = \pi_{HS}(\beta | \tau)$. Then to show that the tail-index of $\pi_{HS}(\beta | \tau)$ is $\alpha = 1$ for any $\tau > 0$, we will prove that the it holds $\lim_{\beta \to \infty} \bar{F}_{HS}(c\beta | \tau)/\bar{F}_{HS}(\beta | \tau) = c^{-1}$ for any $c > 0$ and $\tau > 0$. (Because $\pi_{HS}(\beta | \tau)$ is a symmetric density, showing one-directional
limit $\beta \to \infty$ is sufficient.) By the L'Hôpital's Rule, it holds $\lim_{\beta \to \infty} \bar{F}_{\text{HS}}(c\beta|\tau) / \bar{F}_{\text{HS}}(\beta|\tau) = c \cdot \lim_{\beta \to \infty} \pi_{\text{HS}}(c\beta|\tau) / \pi_{\text{HS}}(\beta|\tau)$, hence, our eventual goal is to prove

$$\lim_{\beta \to \infty} \frac{\pi_{\text{HS}}(c\beta|\tau)}{\pi_{\text{HS}}(\beta|\tau)} = c^{-2}, \quad c > 0, \quad \tau > 0.$$ 

Now, use inequality (S.4) to upper and lower bound the function $\pi_{\text{HS}}(c\beta|\tau) / \pi_{\text{HS}}(\beta|\tau)$;

$$\frac{l(c\beta)}{u(\beta)} < \frac{\pi_{\text{HS}}(c\beta|\tau)}{\pi_{\text{HS}}(\beta|\tau)} < \frac{u(c\beta)}{l(\beta)}, \quad c > 0, \quad \beta, \tau > 0. \quad (S.5)$$

First, calculate the limit of the upper bound in the inequality (S.5) at infinity by using L'Hôpital's Rule;

$$\lim_{\beta \to \infty} \frac{u(c\beta)}{l(\beta)} = 2 \lim_{\beta \to \infty} \frac{\log \left\{ \frac{Z_{\text{HS}}(c\beta) + 1}{Z_{\text{HS}}(c\beta)} \right\}}{\log \left\{ \frac{Z_{\text{HS}}(\beta) + 2}{Z_{\text{HS}}(\beta)} \right\}} = 2 \lim_{\beta \to \infty} \frac{\{Z_{\text{HS}}(c\beta)/(Z_{\text{HS}}(c\beta) + 1)\} \cdot (-c^2/Z_{\text{HS}}(c\beta)^2)}{\{Z_{\text{HS}}(\beta)/(Z_{\text{HS}}(\beta) + 2)\} \cdot (-2/Z_{\text{HS}}(\beta)^2)} = c^2 \cdot \lim_{\beta \to \infty} \frac{Z_{\text{HS}}(\beta) \cdot (Z_{\text{HS}}(\beta) + 2)}{Z_{\text{HS}}(c\beta) \cdot (Z_{\text{HS}}(c\beta) + 1)} = c^{-2}, \quad c > 0.$$

By the same way, we can show $\lim_{\beta \to \infty} l(c\beta) / u(\beta) = c^{-2}, \quad c > 0$. Use the squeeze theorem to the inequality (S.5) to finish the proof.

S.5. PROOF– PROPERTIES OF THE GLT PRIOR

S.5.1 Proof– Proposition 3

(a) Clearly,

$$\pi(\beta|\tau, \xi) = \frac{1}{\tau \sqrt{2\pi}} \int_{0}^{\infty} \frac{1}{\lambda} \exp \left( -\frac{\beta^2}{2\lambda^2} \left( 1 + \frac{\xi \lambda}{\tau} \right) \right)^{-\left(1/\xi + 1\right)} d\lambda.$$ 

Let $x = \xi \lambda / \tau$. Then

$$\pi(\beta|\tau, \xi) = \frac{1}{\tau \sqrt{2\pi}} \int_{0}^{\infty} \exp \left( -\frac{\beta^2 \xi^2}{2\tau^2 x^2} \right) x^{-1}(1 + x)^{-\left(1/\xi + 1\right)} dx,$$
or equivalently, for \( t = 1/x^2 \):

\[
\pi(\beta|\tau, \xi) = K \int_0^\infty e^{-Zt(t^{1/2})-1+1/\xi}(1 + t^{1/2})^{-(1+1/\xi)} dt,
\]

where \( K = 1/(\tau^{3/2} \pi^{1/2}) \) and \( Z(\beta) = \beta^2 \xi^2 / (2 \tau^2) \). Use \( Z = Z(\beta) \) to avoid notation clutter. To utilize the Newton’s generalized binomial theorem;

\[
(x + y)^r = \sum_{k=0}^{\infty} \binom{r}{k} x^{r-k} y^k, \quad |x| > |y|, r \in \mathbb{C},
\]

we divide the integral in (S.6) into two parts. Then we have

\[
\pi(\beta|\tau, \xi) = K \left\{ \int_0^1 e^{-Zt(t^{1/2})-1+1/\xi}(1 + t^{1/2})^{-(1+1/\xi)} dt + \int_1^\infty e^{-Zt(t^{1/2})-1+1/\xi}(1 + t^{1/2})^{-(1+1/\xi)} dt \right\}.
\]

(S.7)

The first integral of (S.7) is

\[
\int_0^1 e^{-Zt(t^{1/2})-1+1/\xi}(1 + t^{1/2})^{-(1+1/\xi)} dt = \int_0^1 e^{-Zt(t^{1/2})-1+1/\xi} \sum_{k=0}^{\infty} \left( -\frac{1}{k} - \frac{1}{\xi} \right) (t^{1/2})^k dt
\]

\[
= \sum_{k=0}^{\infty} \left( -\frac{1}{k} - \frac{1}{\xi} \right) \int_0^1 e^{-Zt(1+1/\xi+k/2)/2} dt
\]

\[
= \sum_{k=0}^{\infty} \left( -\frac{1}{k} - \frac{1}{\xi} \right) Z^{-(1+1/\xi+k)/2} \gamma\left(\frac{1+1/\xi+k}{2}, Z\right),
\]

(S.8)
where \( \gamma(s, x) = \int_0^x t^{s-1}e^{-t} dt \) \((s, x \in \mathbb{R})\), is the incomplete lower gamma function.

The second integral of (S.7) is

\[
\int_1^\infty e^{-Z(t^{1/2})-(1+1/\xi)(1+t^{1/2})} dt = \int_1^\infty e^{-Z(t^{1/2})-(1+1/\xi)} \sum_{k=0}^{\infty} \left(-\frac{1}{k}\right)(t^{1/2})^{-1-1/\xi-k} dt
\]

\[
= \sum_{k=0}^{\infty} \left(-\frac{1}{k}\right) \int_0^1 e^{-Z(t^{1/2})-(1-k)} dt
\]

\[
= \sum_{k=0}^{\infty} \left(-\frac{1}{k}\right) E_{k/2+1}(Z),
\]

(S.9)

where \( E_s(x) = \int_1^\infty e^{-xt-x} dt \) \((s, x \in \mathbb{R})\) is the generalized exponential-integral function of real order \([6, 25]\). Use \((-1-1/\xi) = (-1)^k(1/\xi+k)\), (S.8), and (S.9) to conclude the proof.

(b) Prove by using the change of variable;

\[
\pi(\kappa|\tau, \xi) = \mathcal{GP}(\lambda|\tau, \xi) \bigg|_{\lambda=\sqrt{(1-\kappa)/\kappa}} \cdot \left| \frac{d\lambda}{d\kappa} \right| = \frac{1}{\tau} \left( \frac{\xi}{\tau} \sqrt{\frac{1-\kappa}{\kappa}} \right)^{-(1+1/\xi)} \cdot \frac{1}{2\kappa^2} \left( \frac{1}{\kappa} \right)^{-1/2}
\]

\[
= \frac{1}{2\tau} (\tau \sqrt{\kappa + \xi \sqrt{1-\kappa}})^{(1+1/\xi)} \cdot \frac{1}{\kappa^{1/2}} \left( \frac{1}{\kappa} \right)^{-1/2}
\]

\[
= \tau^{1/\xi} \cdot \frac{\kappa^{1/(2\xi)-1}}{(\tau \kappa^{1/2} + \xi (1-\kappa)^{1/2})^{(1+1/\xi)}}.
\]

S.5.2 Proof—Corollary 4

(a) In general, the generalized exponential-integral function has the following property: \( \lim_{x \to 0^+} E_1(x) = \infty \) and \( \lim_{x \to 0^+} E_s(x) = 1/(s-1) \) for \( s > 1 \) \([6]\). Using this property, if \( k = 0 \), then \( \lim_{|\beta| \to 0} \psi^S_{k=0}(\beta) = \lim_{|\beta| \to 0} E_1\{Z(\beta)\} = \infty \) because \( Z(\beta) = \beta^2\xi^2/(2\tau^2) \). If \( k \in \mathbb{N} \), then \( \lim_{|\beta| \to 0^+} \psi_k^S(\beta) = \lim_{|\beta| \to 0^+} E_{k/2+1}\{Z(\beta)\} = 2/k < \infty \).

(b) In general, the incomplete gamma function has the following property: \( \lim_{x \to 0^+} x^{-a} \cdot \gamma(a, x) = a^{-1} \) for \( a > 0 \) \([15]\). Using this property, \( \lim_{|\beta| \to 0} \psi^R_k(\beta) = \lim_{|\beta| \to 0} Z(\beta)^{-(1+1/\xi+k)} \cdot \gamma\{(1+1/\xi+k)/2, Z(\beta)\} = 2/(1+1/\xi+k) < \infty \) for all \( k \in \{0\} \cup \mathbb{N} \).

(c) In general, the generalized exponential-integral function has the following property: \( e^{-x}/(x+s) \leq E_s(x) \leq e^{-x}/(x+s-1) \) for \( x > 0 \) and \( s \geq 1 \) \([6]\). Using this property, we obtain an inequality \( e^{-Z(\beta)}/\{Z(\beta) + s\} \leq E_s(Z(\beta)) \leq e^{-Z(\beta)}/\{Z(\beta) + s - 1\} \) for \(|\beta| > 0 \) and \( s \geq 1 \). As
\( |\beta| \to \infty \), both bounds of \( E_s(Z(\beta)) \) converges to zero with squared exponential rate, and hence, \( E_s(Z(\beta)) \) also do for any \( s \geq 1 \).

(d) For fixed \( k \in \{0\} \cup \mathbb{N} \) and \( \xi \), we have \( \lim_{|\beta| \to \infty} \gamma \{(1+1/\xi+k)/2, Z(\beta)\} = \Gamma((1+1/\xi+k)/2) \), where \( \Gamma \) is the gamma function, and hence, the function \( \gamma \{(1+1/\xi+k)/2, Z(\beta)\} \) is a slowly varying function \([24]\). Using this we can re-express \( \psi_R^k(\beta) = Z(\beta) - (1+1/\xi+k)/2 \cdot \gamma \{(1+1/\xi+k)/2, Z(\beta)\} \) by \( \psi_R^k(\beta) = \beta - (1+1/\xi+k) \cdot L(\beta) \), where \( L \) is a slowly varying function. This implies that the tail-index of function \( \psi_R^k(\beta) \) is \( 1 + 1/\xi + k \).

S.5.3 Proof– Lemma 6

(a) Under the formulation (6) – (9), i.e., \( \beta \sim \pi_{\text{GLT}}(\beta) \), we have

\[
\pi(\xi\mid |-) \propto \left\{ \prod_{j=1}^{p} \mathcal{GPD}(\lambda_j | \tau, \xi) \right\} \cdot \mathcal{IG}(\tau | p/\xi + 1, 1) \cdot \log \mathcal{N}_1(\xi | \mu, \rho^2) \cdot \mathcal{I}_{(1/2, \infty)}(\xi)
\]

\[
= \left\{ \prod_{j=1}^{p} \frac{1}{\tau} \left( 1 + \frac{\xi \lambda_j}{\tau} \right)^{-(1/\xi + 1)} \right\} \cdot \frac{\tau^{-p/\xi - 2} e^{-1/\tau}}{\Gamma(p/\xi + 1)} \cdot \log \mathcal{N}_1(\xi | \mu, \rho^2) \cdot \mathcal{I}_{(1/2, \infty)}(\xi)
\]

\[
\propto \left\{ \tau^{p/\xi} \cdot \prod_{j=1}^{p} \left( \tau + \xi \lambda_j \right)^{-(1/\xi + 1)} \right\} \cdot \frac{\tau^{-p/\xi - 2}}{\Gamma(p/\xi + 1)} \cdot \log \mathcal{N}_1(\xi | \mu, \rho^2) \cdot \mathcal{I}_{(1/2, \infty)}(\xi)
\]

\[
\propto \frac{\pi^{p/2}}{\Gamma(p/\xi + 1)} \prod_{j=1}^{p} \left( \tau + \xi \lambda_j \right)^{-(1/\xi + 1)} \cdot \log \mathcal{N}_1(\xi | \mu, \rho^2) \cdot \mathcal{I}_{(1/2, \infty)}(\xi).
\]

Now, our goal is to show

\[
m(\lambda, \tau) = \int_{1/2}^{\infty} \frac{\pi^{p/2}}{\Gamma(p/\xi + 1)} \prod_{j=1}^{p} \left( \tau + \xi \lambda_j \right)^{-(1/\xi + 1)} \cdot \log \mathcal{N}_1(\xi | \mu, \rho^2) d\xi < \infty, \quad \lambda \in (0, \infty)^p, \tau \in (0, \infty).
\]
Let \( x = 1/\xi \). Then

\[
m(\lambda, \tau) = \int_0^\pi \frac{\pi^{p/2}}{\Gamma(px + 1)} \prod_{j=1}^p \left( \frac{x}{\lambda_j + \tau x} \right)^{\lambda_j + \tau x} \cdot \log N_1(1/x|\mu, \rho^2) \cdot -\frac{1}{x^2} dx
\]

\[
= \pi^{p/2} \cdot \int_0^2 \frac{(1/\tau)^p(x+1)}{\Gamma(px + 1)} \prod_{j=1}^p \left( \frac{\tau x}{\lambda_j + \tau x} \right)^{\lambda_j + \tau x} \cdot \log N_1(1/x|\mu, \rho^2) \cdot \frac{1}{x^2} dx
\]

\[
\leq \pi^{p/2} \cdot \int_0^2 r(x) \cdot \log N_1(1/x|\mu, \rho^2) \cdot \frac{1}{x^2} dx, \tag{S.10}
\]

where \( r(x) = (1/\tau)^p(x+1)/\Gamma(px + 1) \). Since \( r(x) \) is continuous on a closed interval \([0, 2]\), there exists \( x_0 \in [0, 2] \) such that \( r(x_0) = \sup_{x \in [0,2]} r(x) = B \). Using this bound \( B \) to (S.10), we have

\[
m(\lambda, \tau) \leq \pi^{p/2} \cdot B \cdot \int_0^\pi \log N_1(1/x|\mu, \rho^2) \cdot \frac{1}{x^2} dx
\]

\[
\leq \pi^{p/2} \cdot B \cdot \int_0^\pi \log N_1(1/x|\mu, \rho^2) \cdot \frac{1}{x^2} dx = \pi^{p/2} \cdot B < \infty, \quad \lambda \in (0, \infty)^p, \tau \in (0, \infty).
\]

(b) Start with a likelihood part:

\[
f(y|\xi) = \int_0^\infty \int_0^\infty \int_{-\infty}^\infty N_1(y|\beta, 1) \cdot N_1(\beta|0, \lambda^2) \cdot \mathcal{GPD}(\lambda|\tau, \xi) \cdot \mathcal{IG}(1/\xi + 1, 1)d\beta d\lambda d\tau
\]

\[
= \int_0^\infty \int_0^\infty \int_{-\infty}^\infty N_1(y|0, 1 + \lambda^2) \cdot \mathcal{GPD}(\lambda|\tau, \xi) \cdot \mathcal{IG}(\tau|1/\xi + 1, 1)d\lambda d\tau
\]

\[
= \frac{1}{\sqrt{2\pi}} \int_0^\infty \left( \int_0^\infty \frac{1}{\sqrt{1 + \lambda^2}} \cdot \exp \left\{ -\frac{y^2}{2(1 + \lambda^2)} \right\} \cdot \frac{1}{\tau} \left( 1 + \frac{\xi \lambda}{\tau} \right)^{-\frac{1}{\xi+1}} d\lambda \right) \cdot \mathcal{IG}(\tau|1/\xi + 1, 1)d\tau
\]

\[
\leq \frac{1}{\sqrt{2\pi}} \int_0^\infty \left( \int_0^\infty \frac{1}{\sqrt{1 + \lambda^2}} \cdot \exp \left\{ -\frac{y^2}{2(1 + \lambda^2)} \right\} \cdot \frac{1}{\tau + \xi \lambda} d\lambda \right) \cdot \mathcal{IG}(\tau|1/\xi + 1, 1)d\tau
\]

\[
= \frac{1}{\sqrt{2\pi}} \int_0^\infty \left( \int_0^1 g(y, \lambda, \tau, \xi) d\lambda + \int_1^\infty g(y, \lambda, \tau, \xi) d\lambda \right) \cdot \mathcal{IG}(\tau|1/\xi + 1, 1)d\tau, \tag{S.11}
\]

where \( g(y, \lambda, \tau, \xi) = \{1/\sqrt{1 + \lambda^2} \cdot \exp [-y^2/(2(1 + \lambda^2))] \cdot 1/(\tau + \xi \lambda) \}, y \in \mathbb{R} \) and \( \lambda, \tau > 0 \).

Because \( g(y, \lambda, \tau, \xi) \) is continuous on a closed interval \([0, 1]\) as a function of \( \lambda \), by mean value
theorem for integral [2], there exists $c \in (0, 1)$ such that

$$
\int_0^1 g(y, \lambda, \tau, \xi) d\lambda = g(y, c, \tau, \xi) = \frac{1}{\sqrt{1 + c^2}} \cdot \exp \left\{ - \frac{y^2}{2(1 + c^2)} \right\} \cdot \frac{1}{\tau + \xi c}
$$

$$
\leq \left[ \frac{1}{\sqrt{1 + c^2}} \exp \left\{ - \frac{y^2}{2(1 + c^2)} \right\} \right] \cdot \frac{1}{\tau} = A \cdot \frac{1}{\tau} \leq \frac{1}{\tau}, \quad \tau \in (0, \infty), \quad (S.12)
$$

where $A = A(y, c) = \{1/(\sqrt{1 + c^2})\} \cdot \exp \left\{ -y^2/\{2(1 + c^2)\} \right\}$, which is upper bounded by 1 on $\mathbb{R} \times (0, 1)$. Also, we have

$$
\int_1^\infty g(y, \lambda, \tau, \xi) d\lambda = \int_1^\infty \frac{1}{\sqrt{1 + \lambda^2}} \cdot \exp \left\{ - \frac{y^2}{2(1 + \lambda^2)} \right\} \cdot \frac{1}{\tau + \xi \lambda} d\lambda
$$

$$
\leq \int_1^\infty \frac{1}{\lambda} \cdot \frac{1}{\xi \lambda} d\lambda = \int_1^\infty \frac{1}{\lambda^2} d\lambda \cdot \frac{1}{\xi} = \frac{1}{\xi}, \quad \xi \in (1/2, \infty). \quad (S.13)
$$

Using the upper bounds (S.12) and (S.13) to (S.11), then we have

$$
f(y|\xi) \leq \frac{1}{\sqrt{2\pi}} \int_0^\infty \left( \frac{1}{\tau} + \frac{1}{\xi} \right) \cdot IG(\tau|1/\xi + 1, 1) d\tau = \frac{1}{\sqrt{2\pi}} \left( \int_0^\infty \frac{1}{\tau} \cdot IG(\tau|1/\xi + 1, 1) d\tau + \frac{1}{\xi} \right)
$$

$$
= \frac{1}{\sqrt{2\pi}} \left\{ \left( \frac{1}{\xi} + 1 \right) + \frac{1}{\xi} \right\} = \frac{1}{\sqrt{2\pi}} \left( \frac{2}{\xi} + 1 \right) \leq \frac{5}{\sqrt{2\pi}} < \infty, \quad y \in \mathbb{R}, \: \xi \in (1/2, \infty).
$$

Therefore, trivially for any proper prior $\pi(\xi)$ on $(1/2, \infty)$, we have

$$
m(y) = \int_{1/2}^\infty f(y|\xi) \cdot \pi(\xi) d\xi \leq \frac{5}{\sqrt{2\pi}} \int_{1/2}^\infty \pi(\xi) d\xi = \frac{5}{\sqrt{2\pi}} < \infty, \quad y \in \mathbb{R}.
$$
REFERENCES

[1] Bai, G., Cheung, I., Shulha, H. P., Coelho, J. E., Li, P., Dong, X., Jakovcevski, M., Wang, Y., Grigorenko, A., Jiang, Y. et al. (2014). Epigenetic dysregulation of hairy and enhancer of split 4 (hes4) is associated with striatal degeneration in postmortem huntington brains. *Human molecular genetics* **24**, 1441–1456.

[2] Bartle, R. G. & Sherbert, D. R. (2011). *Introduction to real analysis*. Hoboken, NJ: Wiley.

[3] Bhadra, A., Datta, J., Polson, N. G., Willard, B. et al. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis* **12**, 1105–1131.

[4] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

[5] Carvalho, C. M., Polson, N. G. & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480.

[6] Chiccoli, C., Lorenzutta, S. & Maino, G. (1992). Concerning some integrals of the generalized exponential-integral function. *Computers & Mathematics with Applications* **23**, 13–21.

[7] Damlen, P., Wakefield, J. & Walker, S. (1999). Gibbs sampling for bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**, 331–344.

[8] Drees, H., De Haan, L., Resnick, S. et al. (2000). How to make a hill plot. *The Annals of Statistics* **28**, 254–274.

[9] Du, Y.-J., Luo, X.-Y., Hao, Y.-Z., Zhang, T. & Hou, W.-R. (2007). cdna cloning and overexpression of acidic ribosomal phosphoprotein p1 gene (rplp1) from the giant panda. *International journal of biological sciences* **3**, 428.
[10] EFRON, B. & HASTIE, T. (2016). *Computer age statistical inference*, vol. 5. Cambridge University Press.

[11] EMBRECHTS, P., KLÜPPELBERG, C. & MIKOSCH, T. (2013). *Modelling extremal events: for insurance and finance*, vol. 33. Springer Science & Business Media.

[12] GORRINGE, K. L., GEORGE, J., ANGLESIO, M. S., RAMAKRISHNA, M., ETEMADMOGHADAM, D., COWIN, P., SRIDHAR, A., WILLIAMS, L. H., BOYLE, S. E., YANAIHARA, N. et al. (2010). Copy number analysis identifies novel interactions between genomic loci in ovarian cancer. *PloS one* 5, e11408.

[13] HILL, B. M. (1975). A simple general approach to inference about the tail of a distribution. *The annals of statistics*, 1163–1174.

[14] HIRSCHHORN, J. N., LOHMUELLER, K., BYRNE, E. & HIRSCHHORN, K. (2002). A comprehensive review of genetic association studies. *Genetics in medicine* 4, 45.

[15] JAMESON, G. (2016). The incomplete gamma functions. *The Mathematical Gazette* 100, 298–306.

[16] JANSZ, N., KENRY, A., TRUSSART, M., BILDSOE, H., BECK, T., TONKS, I. D., MOULD, A. W., HICKEY, P., BRESLIN, K., IMINITOFF, M. et al. (2018). SmcHd1 regulates long-range chromatin interactions on the inactive X chromosome and at hox clusters. *Nature structural & molecular biology* 25, 766.

[17] KAR, S. P., BEESLEY, J., AL OLAMA, A. A., MICHAILIDOU, K., TYRER, J., KOTETJARAI, Z., LAWRENSON, K., LINDSTROM, S., RAMUS, S. J., THOMPSON, D. J. et al. (2016). Genome-wide meta-analyses of breast, ovarian, and prostate cancer association studies identify multiple new susceptibility loci shared by at least two cancer types. *Cancer discovery* 6, 1052–1067.

[18] KARAMATA, J. (1933). Sur un mode de croissance régulière. théorèmes fondamentaux. *Bulletin de la Société Mathématique de France* 61, 55–62.
[19] Lee, S. & Kim, J. H. (2018). Exponentiated generalized pareto distribution: Properties and applications towards extreme value theory. *Communications in Statistics-Theory and Methods*, 1–25.

[20] Lee, S. Y. (2020). A note on gibbs sampler and coordinate ascent variational inference.

[21] Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatch, A. J., Benz, C. C., Levine, D. A., Lee, A. V. et al. (2018). An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416.

[22] Makalic, E. & Schmidt, D. F. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters* **23**, 179–182.

[23] Mallick, B. K., Gold, D. & Baladandayuthapani, V. (2009). *Bayesian analysis of gene expression data*, vol. 131. John Wiley & Sons.

[24] Mikosch, T. (1999). *Regular variation, subexponentiality and their applications in probability theory*. Eindhoven University of Technology.

[25] Milgram, M. (1985). The generalized integro-exponential function. *Mathematics of computation* **44**, 443–458.

[26] Murray, I., Prescott Adams, R. & MacKay, D. J. (2010). Elliptical slice sampling.

[27] Neal, R. M. (2003). Slice sampling. *Annals of statistics*, 705–741.

[28] Nishihara, R., Murray, I. & Adams, R. P. (2014). Parallel mcmc with generalized elliptical slice sampling. *The Journal of Machine Learning Research* **15**, 2087–2112.

[29] Pellegrino, R., Kavaki, I. H., Goel, N., Cardinale, C. J., Dingès, D. F., Kuna, S. T., Maislin, G., Van Dongen, H. P., Tufik, S., Hogenesch, J. B. et al. (2014). A novel bhlhe41 variant is associated with short sleep and resistance to sleep deprivation in humans. *Sleep* **37**, 1327–1336.
[30] Piironen, J., Vehtari, A. et al. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics* **11**, 5018–5051.

[31] Resnick, S. & Stáricá, C. (1995). Consistency of hill’s estimator for dependent data. *Journal of Applied Probability* **32**, 139–167.

[32] Schmidt, C. (2015). Another shot at cancer. *Nature* **527**, S105.

[33] Shaw, J. A., Page, K., Blighe, K., Hava, N., Guttery, D., Ward, B., Brown, J., Ruangpratheep, C., Stebbing, J., Payne, R. et al. (2012). Genomic analysis of circulating cell-free dna infers breast cancer dormancy. *Genome research* **22**, 220–231.

[34] Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research* **1**, 211–244.

[35] van der Pas, S., Kleijn, B., Van Der Vaart, A. et al. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics* **8**, 2585–2618.

[36] van der Pas, S. L., Kleijn, B. J., Van Der Vaart, A. W. et al. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics* **8**, 2585–2618.

[37] Wang, Q.-X., Chen, E.-D., Cai, Y.-F., Li, Q., Jin, Y.-X., Jin, W.-X., Wang, Y.-H., Zheng, Z.-C., Xue, L., Wang, O.-C. et al. (2016). A panel of four genes accurately differentiates benign from malignant thyroid nodules. *Journal of Experimental & Clinical Cancer Research* **35**, 169.

[38] Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

[39] Wu, D.-M., Liu, T., Deng, S.-H., Han, R. & Xu, Y. (2017). Slc39a4 expression is
associated with enhanced cell migration, cisplatin resistance, and poor survival in non-small cell lung cancer. *Scientific reports* 7, 7211.