JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
2021, VOL. 91, NO. 2, 396–419
https://doi.org/10.1080/00949655.2020.1815199

The adaptive normal-hypergeometric-inverted-beta priors for sparse signals

Hanjun Yu[a], Xinyi Xu[b] and Di Cao[c]

[a]School of Statistics, Capital University of Economics and Business, Beijing, People's Republic of China; [b]Department of Statistics, The Ohio State University, Columbus, OH, USA; [c]Leyi Network, Shenzhen, People's Republic of China

ABSTRACT

We investigate the estimation of high-dimensional normal mean under sparsity. Most shrinkage priors in the literature are based on certain assumptions of sparsity levels and signal sizes. Violation of these assumptions can lead to unsatisfactory estimation. In this paper, we propose a new class of flexible priors, the adaptive normal-hypergeometric-inverted-Beta (ANHIB) priors, which generalize several popular shrinkage priors without requiring prior knowledge of data sparsity levels and signal sizes, and thus can be used as good default priors in a large variety of situations. We show that the ANHIB estimators provide strong suppression to noises and little shrinkage to large signals, and have consistently superior estimation performance under various sparsity levels and signal sizes.

ARTICLE HISTORY
Received 2 September 2019
Accepted 23 August 2020

KEYWORDS
Sparsity; global-local shrinkage; adaptive normal-hypergeometric-inverted-beta prior; super efficiency; tail robustness

2010 MATHEMATICS SUBJECT CLASSIFICATION
62F15

1. Introduction

Identifying and estimating signals in the presence of a large amount of noise is a challenging problem in high-dimensional statistical inference. In the classic normal model setup, we observe data for a p-dimensional multivariate normal variable $\mathbf{X}$ with an unknown mean vector $\boldsymbol{\theta}$, which contains non-zero components corresponding to signals and potentially many zero components corresponding to noises. The coordinates of $\mathbf{X}$ are assumed to be conditionally independent with a common unknown variance σ^2, that is,

$$\mathbf{X} \mid \boldsymbol{\theta}, \quad \sigma \sim N(\boldsymbol{\theta}, \sigma^2 \mathbf{I}_p), \tag{1}$$

where $\mathbf{I}_p$ is the $p \times p$ identity matrix. The goal is to estimate the high-dimensional mean vector $\boldsymbol{\theta}$ under the squared error loss $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2$.

The naive maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \mathbf{X}$ performs poorly when the parameter dimension is high and the mean vector $\boldsymbol{\theta}$ is sparse [1]. The frequentist approaches for improving the MLE are usually based on penalized likelihoods. The estimators, such as the ridge estimator [2,3], the lasso estimator [4] and the elastic net estimator

CONTACT Xinyi Xu xinyi@stat.osu.edu Department of Statistics, The Ohio State University, 1958 Neil Ave., Columbus, OH 43210, USA

[5], are derived as the minimizers of penalized likelihoods under various penalty functions. Recent work often embeds local parameters as well as global parameters in the penalty functions to better separate signals from noises. For example, the adaptive lasso method proposed by Zou [6] uses different weights for penalizing the coefficients of different coordinates.

Under the Bayesian framework, the penalized likelihood estimators can often be interpreted as the posterior modes of certain posterior distributions, and their penalty functions are the negative logarithm of the corresponding priors. For example, the ridge estimator can be viewed as the posterior mode (mean) under a conjugate normal prior, and the lasso estimator can be viewed as the posterior mode under a double-exponential prior [4]. The choice of prior distributions is critical in high-dimensional Bayesian analysis. A popular class of priors for inducing sparsity are the so-called 'spike-and-slab' priors [7]. They consist of a point mass at 0 that provides substantial shrinkage for noises, and a (heavy-tailed) continuous distribution that gives little shrinkage to large signals. Although such a two-group model has conceptual and theoretical appeals, its computational complexity is exponential in parameter dimension, and so the implementation on massive datasets is usually impractical.

In the past decade, increasing attention has been drawn to continuous one-group priors that feature a peak around zero and heavy tails on the two sides. These priors, while far from exhaustive, include the Strawderman–Berger prior [8,9], the normal-Jeffreys prior [10], the normal-exponential-gamma (NEG) prior [11], the normal-gamma (NG) prior [12], the horseshoe (HS) prior [13], the three-parameter beta prior [14], the generalized double Pareto (GDP) prior [15], the Dirichlet–Laplace prior [16] and the horseshoe+ prior [17]. They often contain both global and local scale parameters. In these priors, the prior variance of each coordinate is represented by the product of a global scale parameter, which is common across all coordinates, and a local scale parameter, which is specific to that coordinate. The shared global scale parameter represents the overall variation in $\boldsymbol{\theta}$, and is usually tied with the likelihood variance σ^2, while the local scale parameters allow better separation of signals and noises through different shrinkage degrees for different coordinates. These global and local scale parameters are assumed to follow another layer of hyper-priors. The shape of a hyper-prior can have large impact on its shrinkage profile, and is usually assumed to be a specific form based on certain assumption of the sparsity levels or the signal sizes. These distributions are attractive for their computational tractability and excellent empirical and theoretical properties when their sparsity assumptions are satisfied [18,19]. However, when their sparsity assumptions are violated, they leave much to be desired. For example, the Bayes estimator under the horseshoe prior closely attains the oracle risk for highly sparse data, but its performances for non-sparsity scenarios leave room for improvement [18].

The major thrust of this paper is to develop a new class of global-local priors, the adaptive normal-hypergeometric-inverted-Beta (ANHIB) priors, which generalize several popular shrinkage priors in the literature and retain the desirable theoretical properties. These priors do not require prior knowledge of data sparsity levels and signal sizes, and thus can be used as good default priors in a large variety of situations. We prove that the Bayes estimator under the ANHIB priors provides strong suppression to noises and essentially no shrinkage to large signals. Moreover, we demonstrate through simulation studies and empirical analysis that the ANHIB estimator with the default configuration consistently

provides superior estimation performance under various sparsity levels and signal sizes, and substantially improves some common shrinkage estimators.

The rest of this paper is organized as follows: Section 2 describes the construction of the ANHIB prior, shows the properties of its marginal density and compares its shrinkage profile with some commonly used priors in the literature. Section 3 establishes the theoretical properties of the Bayes estimator under the ANHIB prior. The empirical performances of the ANHIB prior is evaluated and compared with other shrinkage priors through simulation studies in Section 4. Its usage in wavelet de-noising and linear regression is further demonstrated using an electrocardiogram data set and a prostate cancer data set in Section 5. Finally, Section 6 concludes with discussions on the results and future work.

2. The adaptive normal-hypergeometric-inverted-beta priors

2.1. Construction of the adaptive normal-hypergeometric-inverted-Beta priors

Among the Bayesian approaches for estimating the mean of a high-dimensional normal distribution, a very successful strand of methods is through using priors that are scale mixtures of normals, that is, the probability density function $p(\boldsymbol{\theta})$ is

$$p(\boldsymbol{\theta}) = \int N(\boldsymbol{\theta} \mid \mathbf{0}, \boldsymbol{\psi}^2) G(\mathrm{d}\boldsymbol{\psi}^2), \tag{2}$$

where $N(\boldsymbol{\theta} \mid \mathbf{0}, \boldsymbol{\psi}^2)$ represents the normal density function with mean vector $\mathbf{0}$ and $p \times p$ diagonal covariance matrix $\mathrm{Diag}(\boldsymbol{\psi}^2) = \mathrm{Diag}(\psi_1^2, \ldots, \psi_p^2)$, and $G(d\boldsymbol{\psi}^2)$ is the mixing distribution. The choice of the mixing distribution $G(d\boldsymbol{\psi}^2)$ largely determines the shape of $p(\boldsymbol{\theta})$, especially the peak height at the origin and the tail heaviness at large values, and thus the shrinkage profile of the corresponding Bayes estimator. Various forms of the mixing distribution have been proposed in the literature (e.g. [8–13,15]).

To obtain accurate and robust signal estimation, we would like our prior distribution $p(\boldsymbol{\theta})$ to have the following three properties:

(1) *Scale invariance*. The prior $p(\boldsymbol{\theta})$ is desired to incorporate the scale of the observations, so that the Bayes estimator is invariant to measurement units;
(2) *Flexibility*. The prior $p(\boldsymbol{\theta})$ is desired to have a flexible form. Also, it should contain local parameters, so that it can provide strong shrinkage for noises and little shrinkage for signals.
(3) *Adaptivity*. The shape of the prior $p(\boldsymbol{\theta})$ is desired to be adaptive to a wide range of sparsity levels and signal sizes, so that the Bayes estimator can have robust performances in various scenarios.

For achieving the scale invariance property, we follow the common practice and let

$$\psi_i^2 = \sigma^2 \lambda_i^2, \quad i = 1, \ldots, p, \tag{3}$$

where σ^2 is the sampling variance of X_i in (1) and λ_i^2's are the local scale parameters. In this way, the prior belief about $\boldsymbol{\theta}$ is calibrated by the scale of measurement of $\mathbf{X}$. As σ can be regarded as a scale parameter, the distribution of λ_i^2 plays an important role in determining

Table 1. Priors for λ_l^2 and $\kappa_l = 1/(1+\lambda_l^2)$ under several common scale mixtures of normal priors.

Prior for θ_l	Density for λ_l^2	Density for κ_l
Normal-HIB($a, b, \tau = 1, s = 0$)	$(\lambda_l^2)^{b-1}(1+\lambda_l^2)^{-(a+b)}$	$\kappa_l^{a-1}(1-\kappa_l)^{b-1}$
Horseshoe	$(\lambda_l^2)^{-\frac{1}{2}}(1+\lambda_l^2)^{-1}$	$\kappa_l^{-\frac{1}{2}}(1-\kappa_l)^{-\frac{1}{2}}$
Strawderman-Berger	$(1+\lambda_l^2)^{-\frac{3}{2}}$	$\kappa_l^{-\frac{1}{2}}$
NEG($\gamma, \delta = 1$)	$(1+\lambda_l^2)^{-(\gamma+1)}$	$\kappa_l^{\gamma-1}$
Double-exponential	$\exp(-\lambda_l^2/2)$	$\kappa_l^{-3}\exp[-(1-\kappa_l)/(2\kappa_l)]$

Note: The densities are given up to constants.

the shape of the probability density function $p(\theta_i)$. For example, when $\sigma^2 = 1$,

$$p(\theta_i) = \int_0^\infty N(\theta_i \mid 0, \lambda_i^2) p(\lambda_i^2)\, \mathrm{d}\lambda_i^2.$$

Table 1 summarizes a collection of popular scale mixtures of normal priors. Note that the mixing densities $p(\lambda_i^2)$ in the horseshoe prior, the Strawderman-Berger prior and the normal-exponential-gamma (NEG) priors have a common form $\lambda_i^{2(b-1)}(1+\lambda_i^2)^{-(a+b)}$ for some constants a and b. They can be encompassed by the hypergeometric inverted-Beta distribution HIB(a, b, τ, s) with the density

$$p(\lambda_i^2) = C^{-1}(\lambda_i^2)^{b-1}(1+\lambda_i^2)^{-(a+b)} \exp\left\{-\frac{s}{1+\lambda_i^2}\right\}\left\{\frac{1}{\tau^2} + \left(1-\frac{1}{\tau^2}\right)\frac{1}{1+\lambda_i^2}\right\}^{-1}, \quad (4)$$

where $a, b, \tau > 0$ and $s \in \mathbb{R}$. The normalizing constant C can be expressed as

$$C = \exp(-s)\mathrm{Beta}(a, b)\Phi_1(b, 1, a+b, s, 1-1/\tau^2),$$

where $\mathrm{Beta}(a, b)$ is the beta function and Φ_1 is the degenerate hypergeometric function of two variables [20, 9.261], that is,

$$\Phi_1(\alpha, \beta, \gamma, x, y) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \frac{(\alpha)_{m+n}(\beta)_n}{(\gamma)_{m+n} m!n!} y^n x^m,$$

where $(c)_n$ is the rising factorial. The hypergeometric inverted-Beta distribution HIB(a, b, τ, s) is also known as Pearson's Type VI distribution, and is contained in the class of compound confluent hypergeometric distributions [21]. When $\tau = 1$ and $s = 0$, it reduces to an inverted-Beta distribution. In addition, as both a and b approach 0, this prior converges to the Jeffreys prior. The normal scale mixtures with HIB(a, b, τ, s) as mixing distribution can be called the normal-hypergeometric-inverted-Beta prior.

Polson and Scott [22,23] studied the effects of the four hyper-parameters a, b, τ and s on $p(\theta)$ in details. The parameters a and b control the shape of the distribution. Smaller values of a encourage heavier tails of $p(\theta)$, and smaller values of b encourage $p(\theta)$ to place more prior mass around the origin. Specifically, $a = 1/2$ leads to $p(\theta)$ with Cauchy-like tails, and $b \leq 1/2$ leads to $p(\theta)$ with an unbounded peak at the origin. Figure 1 shows the effects of a and b on the shape of $p(\lambda^2)$ and $p(\theta)$ when $\tau = 1$ and $s = 0$. The upper two panels illustrate the role of a when b is fixed at 2. Smaller values of a yield less mass of $p(\lambda^2)$ near the origin,

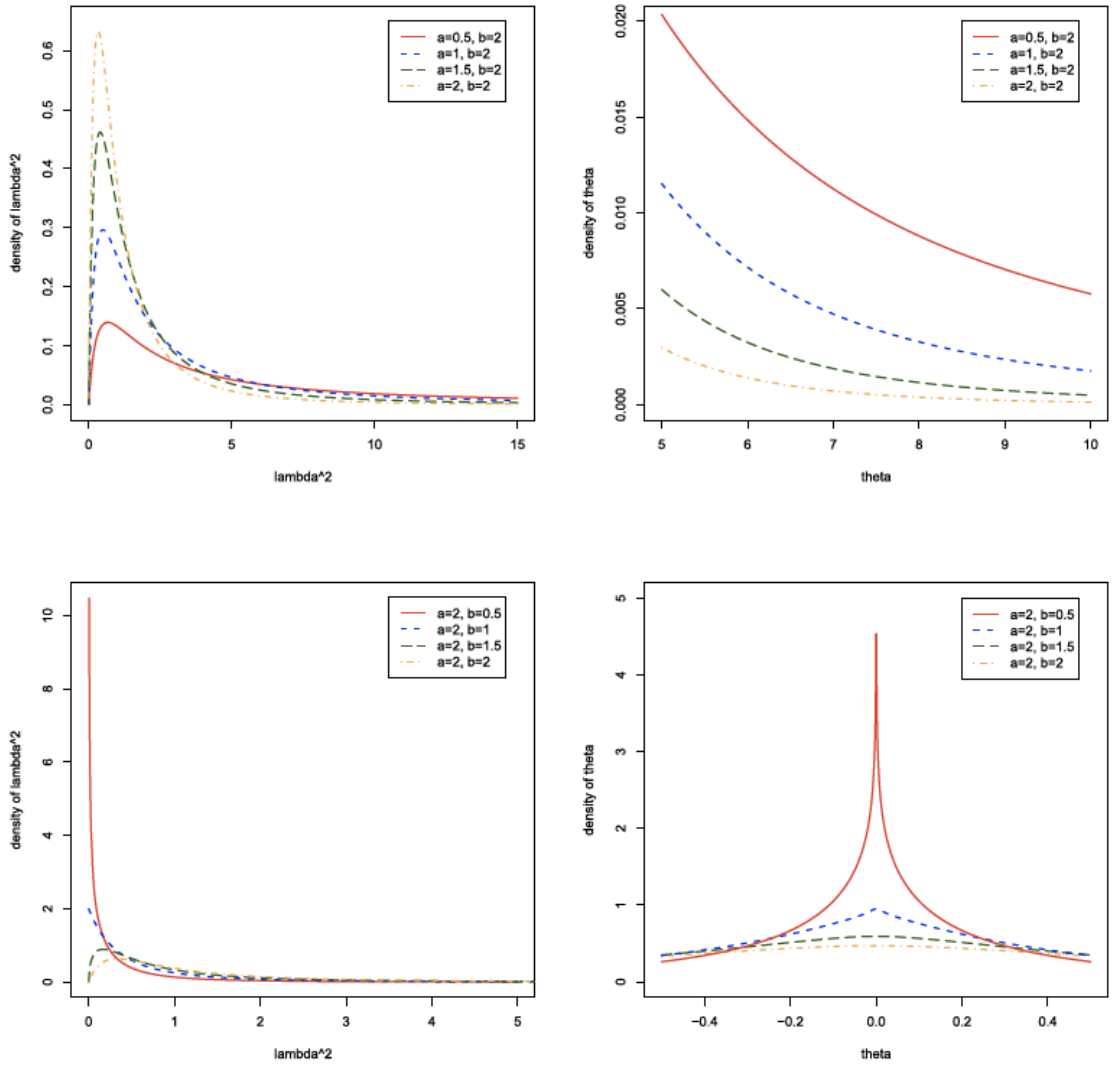

Figure 1. Effects of a and b on $p(\lambda^2)$ and $p(\theta)$ with $\tau = 1$ and $s = 0$. The upper two compare $p(\lambda^2)$ and $p(\theta)$ at tails when b is fixed and a varies. The lower two compare $p(\lambda^2)$ and $p(\theta)$ around the origin when a is fixed and b varies.

which encourages heavier tails of $p(\theta)$. The bottom two panels present the role of b when a is fixed at 2. Smaller values of b encourage more mass of $p(\lambda^2)$ near the origin, which leads to more prior mass of $p(\theta)$ around the origin. The parameters τ and s are global scale parameters. As discussed in [22,23], it is hard to separate their roles, because similar global shrinkage behaviours can be obtained through tuning either τ or s.

We consider constructing a class of flexible and adaptive priors based on the hypergeometric inverted-Beta distribution (4). Without sacrificing much flexibility, we set $s = 0$ and adopt the hypergeometric inverted-Beta distributions with τ as the only global scale parameter. To allow the shape of the prior $p(\theta)$ to be adaptive to various sparsity levels and signal sizes in the data, we place another layer of priors on these hyper-parameters in HIB$(a, b, \tau, 0)$. Ideally, these hyper-priors should reflect our belief about the data sparsity and signal sizes. Under the likelihood function (1) and the scale mixtures of normal

prior (2), the posterior mean of θ_i is

$$E(\theta_i \mid X_i) = \left(1 - E\left(\frac{1}{1+\lambda_i^2}\middle| X_i\right)\right) X_i = (1 - E(\kappa_i \mid X_i))X_i,$$

and thus $\kappa_i = 1/(1+\lambda_i^2)$'s can be viewed as the shrinkage coefficients. It is desirable to have large κ_i values for noises and little κ_i for signals. When $\lambda_i^2 \sim \text{HIB}(a, b, \tau, 0)$, the implied density for κ_i takes the form

$$p(\kappa_i) \propto \kappa_i^{a-1}(1-\kappa_i)^{b-1}\left\{\frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right)\kappa_i\right\}^{-1},$$

and as $\tau^2 = 1$, this is reduced to a Beta(a, b) distribution with mean $a/(a+b)$ and variance $ab/(a+b)^2(a+b+1)$. The existing priors such as the horseshoe prior and the Strawderman–Berger prior with the HIB distribution as mixing distribution have fixed a and b values, which determine fixed shrinkage profiles of the prior, so that the corresponding Bayesian estimators have good performance for certain sparsity levels and signal sizes. To make the prior adaptive to various sparsity levels and signal sizes, instead of choosing fixed values for a and b, we let them be estimated from the data by placing hyper-priors on a and b. The roles of a and b can be better reflected through the transformation form $M = a + b$ and $N = a/(a+b)$. The parameter N reflects the mean of the shrinkage coefficients κ_i's. The more sparse the data is, the larger N should be to provide substantial shrinkage to the dominating noises; on the other hand, the less sparse the data is, the smaller N should be to provide little shrinkage to the signals. The parameter M helps to control the variance of the shrinkage coefficients κ_i's, and thus the separation of the Bayes estimators for different coordinates. We assume

$$\begin{aligned} M = a + b &\sim Gamma(a_0, b_0), \\ N = \frac{a}{a+b} &\sim Beta(\alpha_0, \beta_0), \end{aligned} \tag{5}$$

where a_0, b_0, α_0 and β_0 are pre-specified constants. Finally, to complete the prior specification, we follow the suggestion from [24] and use the following priors for the sampling variance σ^2 and the global scale parameter τ

$$\begin{aligned} p(\sigma^2) &\propto \frac{1}{\sigma^2}, \\ \tau &\sim C^+(0, 1), \end{aligned} \tag{6}$$

where C^+ represents a half-Cauchy distribution. We call the class of hierarchical prior defined jointly by (2), (3), (4), (5) and (6) the adaptive normal-hypergeometric-inverted-Beta (ANHIB) prior and denote it by ANHIB$(a_0, b_0, \alpha_0, \beta_0)$.

Under the ANHIB prior, the marginal density of θ_i is symmetric and has the peak at 0. Since the hyper-prior (5) places positive prior mass on $b \leq 1/2$, this peak is unbounded, which allows strong shrinkage for small noises close to 0. Moreover, the hyper-prior (5) also places positive prior mass on small a values, which leads to heavy tails at large θ values and provides little shrinkage for large signals.

For the default values of the hyper-parameters a_0, b_0, α_0 and β_0, to allow for reasonable shrinkage profiles, noting that most of the common priors in Table 1 have M values between 1/2 and 3/2, and therefore we recommend using $a_0 = b_0 = 20$ as the default values, so that most of the prior mass for M falls in that range. Also, when there is no strong information about the data sparsity a priori, we use $\alpha_0 = \beta_0 = 1$, which indicates that N follows a uniform distribution between 0 and 1.

2.2. Marginal density and shrinkage profile

As many of scale mixtures of normal priors, our ANHIB prior does not have a closed form representation for the marginal density. However, we can provide tight bounds for the density function through the following theorem.

Theorem 2.1: *Assume* $\sigma^2 = \tau^2 = 1$. *Then the marginal density of the ANHIB prior* $p_{\text{ANHIB}}(\theta)$ *satisfies*:

(1) *For* $\theta \neq 0$,

$$C_2 \log\left(1 + \frac{4}{\theta^2}\right) < p_{\text{ANHIB}}(\theta) < \frac{C_1}{|\theta|}, \tag{7}$$

where $C_1 = \int_{a=0}^{\infty}\int_{b=0}^{\infty} C(a,b)\,\mathrm{d}\pi(a,b)$, $C_2 = (1/2\sqrt{2\pi}) * \int_{a=0}^{1/2}\int_{b=0}^{1/2} C(a,b)\,\mathrm{d}\pi(a,b)$, $C(a,b) = \text{Beta}(a,b)\Phi_1(b,1,a+b,0,0)$ *and* $\pi(a,b)$ *is the joint distribution of* a *and* b.

(2) *As* $\theta \to 0$, *for any constant* $0 < \delta_0 < 1/2$,

$$p_{\text{ANHIB}}(\theta) > C_3|\theta|^{2\delta_0 - 1} + O(1) \to \infty, \tag{8}$$

where $C_3 = 2^{-\delta_0}\Gamma(1/2 - \delta_0)/\sqrt{\pi} * \int_{a=0}^{\infty}\int_{b=0}^{\delta_0} C(a,b)\,\mathrm{d}\pi(a,b)$.

Proof: See Section A.1 in the Appendix. ■

Remark: For comparison, recall that the marginal density of the horseshoe prior satisfies the bounds

$$\frac{K}{2}\log\left(1 + \frac{4}{\theta^2}\right) < p_{HS}(\theta) < K\log\left(1 + \frac{2}{\theta^2}\right),$$

where $K = 1/(2\pi^3)^{1/2}$. Theorem 2.1 reveals our ANHIB prior decreases at a comparable rate as the horseshoe prior at the tails, while goes to infinity at a faster rate than the horseshoe prior as $\theta \to 0$. This is because in the construction of the ANHIB prior, a positive probability is placed on the hyper-parameter values $b < 1/2$, which encourages more prior mass around the origin. Therefore, the ANHIB prior implies that it could provide more suppression for noises.

To visually compare the priors for all θ values, we approximate them using a Monte Carlo method that averages the normal priors on θ_i given a set of κ_i values, which are sampled from the corresponding hyper priors.

The upper left panel of Figure 2 illustrates the central parts of the marginal priors of θ_i's under ANHIB(20, 20, 1, 1) and the priors listed in Table 1 with $\sigma^2 = 1$, and the upper

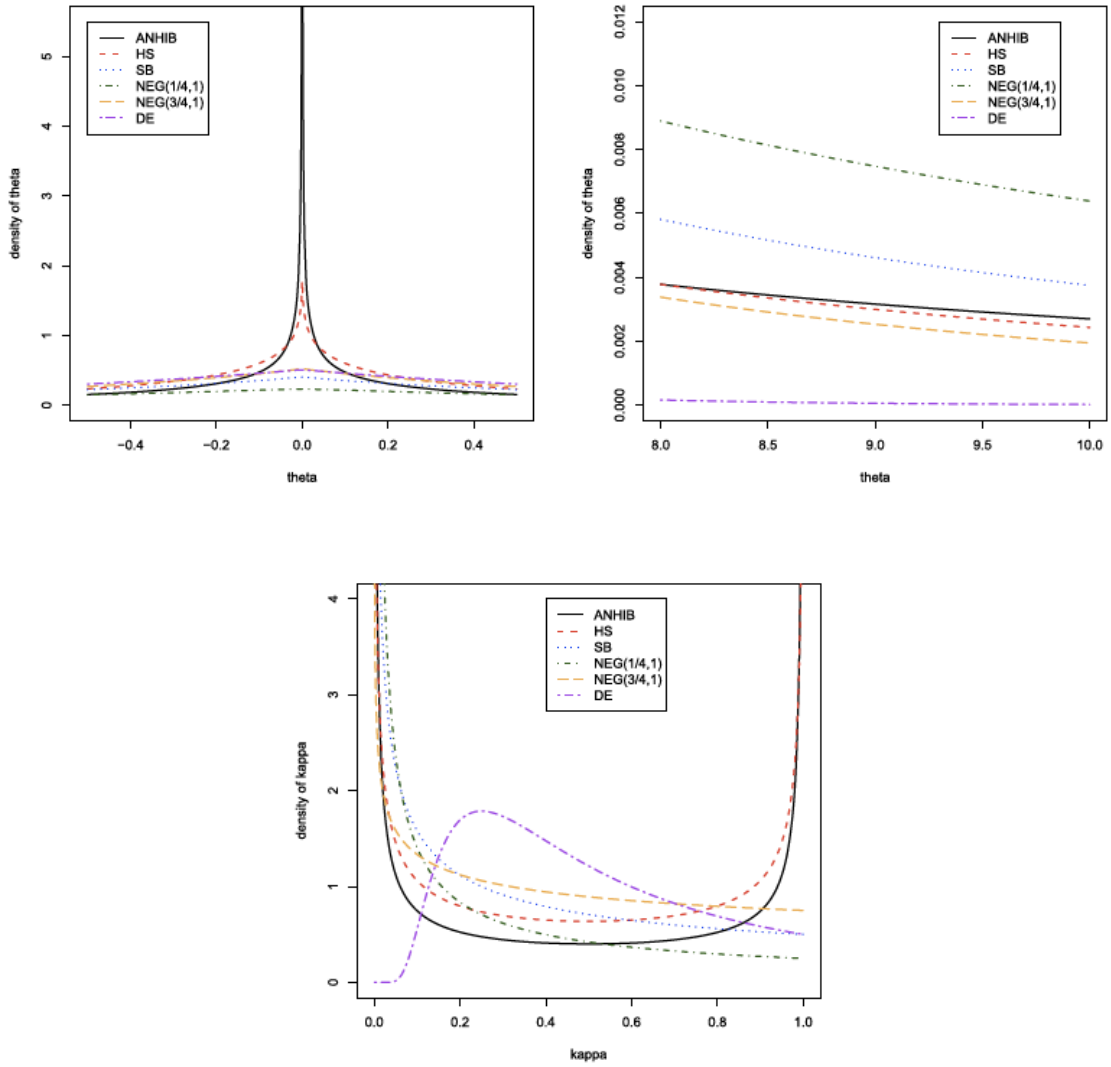

Figure 2. Comparison of the densities $p(\theta)$ and $p(\kappa)$ in the one-dimensional case under the ANHIB(20, 20, 1, 1) prior, the horseshoe prior, the Strawderman–Berger prior, the NEG(1/4,1) prior, the NEG(3/4,1) prior and the double-exponential prior. The upper left panel compares the densities of θ around 0, the upper right panel compares the densities of θ at tails, and the bottom panel compares the densities of κ.

right panel shows the tails of the marginal priors. It can be easily seen that the marginal prior densities of θ_i's under ANHIB(20, 20, 1, 1) and horseshoe have unbounded peaks at the origin, while those under Strawderman-Berger, normal-exponential-gamma(1/4, 1), normal-exponential-gamma(3/4, 1) and double-exponential are bounded. Also, the marginal prior densities under ANHIB(20, 20, 1, 1), horseshoe, Strawderman-Berger, normal-exponential-gamma(1/4, 1) and normal-exponential-gamma(3/4, 1) all have heavy tails. In particular, it is worth noting that ANHIB(20, 20, 1, 1) places more prior mass than horseshoe in a small neighbourhood around the origin and has heavier tails. This suggests that the ANHIB(20, 20, 1, 1) prior favours the case where the signals and the noises are separated. It is consistent with our simulation results in Section 4.

The bottom panel of Figure 2 illustrates the shrinkage profiles under these priors through the densities of κ_i's. Those under ANHIB(20, 20, 1, 1) and horseshoe have

unbounded peaks near 1, allowing strong shrinkage for noises close to 0, while those under Strawderman-Berger, normal-exponential-gamma(1/4, 1), normal-exponential-gamma(3/4, 1) and double-exponential tend to have fixed constants around 1 and thus limit the shrinkage power. The heavy-tailed priors, including the ANHIB(20, 20, 1, 1), horseshoe, normal-exponential-gamma(1/4, 1), normal-exponential-gamma(3/4, 1) and Strawderman-Berger have unbounded peaks near $\kappa_i = 0$, which provides essentially no shrinkage to large signals. Furthermore, the double-exponential prior has vanishing prior mass around $\kappa_i = 0$ that can lead to biased estimates of signals. Note that the ANHIB(20, 20, 1, 1) prior has higher peaks near $\kappa_i = 0$ and 1 compared to the horseshoe prior, reflecting the higher peak at $\theta_i = 0$ and heavier tails at large θ_i values, and still leaves sizable mass in other regions.

3. Theoretical properties of the Bayes estimator under the ANHIB prior

In this section, we investigate the theoretical properties of the Bayes estimator under the ANHIB prior. We show that its Kullback-Leibler risk is bounded and converges to 0 as the sample size grows. In particular, it converges to the true distribution at a super-efficient rate at the origin, while provides little shrinkage for large signals.

3.1. Kullback–Leibler risk bounds

We study the risk of the Bayes estimator under the ANHIB prior measured by the Kullback–Leibler divergence. Let $f_\theta = f(x \mid \theta)$ be a sampling model of X with parameter θ, and let $\hat{f}_n$ be the Bayes estimator of f_{θ_0} based on a sample of size n, that is, $\hat{f}_n = \int f_\theta \pi_n(\mathrm{d}\theta \mid X)$, where $\pi_n(\mathrm{d}\theta \mid X)$ is the posterior distribution. The estimation performance of the Bayes estimator is measured by the Cesàro-average risk

$$R_n(\theta_0) = n^{-1} \sum_{j=1}^{n} L(f_{\theta_0}, \hat{f}_j),$$

where θ_0 is the true value of the parameter and $L(f_{\theta_0}, \hat{f}_j) = E_{f_{\theta_0}} \log(f_{\theta_0}/\hat{f}_j)$ is the Kullback–Leibler divergence of $\hat{f}_j$ from f_{θ_0}.

The following result from [25] provides a useful upper bound for the Cesàro-average risk of a Bayes estimator.

Lemma 3.1 ([25]): *Let $A_\epsilon = \{\theta : L(f_{\theta_0}, f_\theta) \le \varepsilon\} \subset \mathbb{R}$ denote the Kullback–Leibler information neighbourhood of size ε, centred at θ_0 and assume that the prior probability of this neighbourhood $\nu(A_\varepsilon) > 0$ for all $\varepsilon > 0$. Then the Cesàro-average risk of the Bayesian density estimator $\hat{f}_n$ is bounded by*

$$R_n(\theta_0) \le \epsilon - n^{-1} \log \nu(A_\epsilon),$$

for all $\varepsilon > 0$.

By this lemma, the larger the prior probability $\nu(A_\varepsilon)$ is, the smaller the Cesàro-average risk bound would be. The next theorem establishes the Cesàro-average risk bound of the

Bayesian density estimator under the ANHIB prior at any θ_0 value where the prior $p(\theta_0)$ is bounded, that is, at $\theta_0 \neq 0$.

Theorem 3.2: *Suppose that the true sampling model f_{θ_0} is $X \sim N(\theta_0, \sigma^2)$. At any θ_0 value where the prior $p(\theta_0)$ is bounded in a neighbourhood, the Cesàro-average risk of the Bayesian density estimator under the ANHIB prior satisfies*

$$R_{n,\,\mathrm{ANHIB}}(\theta_0) \leq \frac{\log n}{2n} + O\left(\frac{1}{n}\right).$$

Proof: See Section A.2 in the Appendix. ■

3.2. Super-efficiency for sparsity

When the true parameter value $\theta_0 = 0$, according to Lemma 3.1, the unbounded density around 0 under the ANHIB prior induces large prior probabilities in the neighbourhoods A_ε. Therefore, as shown in the next theorem, the corresponding Bayes estimator has a super-efficient rate of convergence and yields more shrinkage power at the origin.

Theorem 3.3: *Suppose that the true sampling model f_{θ_0} is $X \sim N(\theta_0, \sigma^2)$. At $\theta_0 = 0$, the Cesàro-average risk of the Bayesian density estimator under the ANHIB prior satisfies*

$$R_{n,\,\mathrm{ANHIB}}(0) \leq \frac{\delta_0 \log n}{n} + O\left(\frac{1}{n}\right), \tag{9}$$

for any constant $0 < \delta_0 < 1/2$.

Proof: See Section A.3 in the Appendix. ■

The above result shows that at the origin, the Bayes estimator under the ANHIB prior has the Cesàro-average risk converging to 0 at a faster rate than that of the MLE, which has the rate $O(n^{-1}\log n)$, and so is super-efficient in this sense. Moreover, Carvalho et al. [13] showed that the horseshoe estimator is also super-efficient at the origin with the following risk bound:

$$R_{n,HS}(0) \leq \frac{\log n}{2n} - \frac{\log\log n}{n} + O\left(\frac{1}{n}\right).$$

Comparing this bound with (9) suggests that as $n \to \infty$, the bound for the ANHIB estimator is smaller for $0 < \delta_0 < 1/2$, which is due to its higher concentration around 0.

3.3. Robustness to large signals

In the presence of large signals, that is, when the observations would be very different from the prior mean, it is important that the prior has bounded influence on these estimates, so that the signals are not over-shrunk by the prior.

Let $p(|X-\theta|)$ be the standard normal density and $p(\theta)$ be a zero-mean scale mixture of normals with a proper mixing distribution, the posterior mean of θ can be represented by

$$E(\theta \mid x) = x + \frac{\mathrm{d}}{\mathrm{d}x} \log m(x),$$

where $m(x)$ is the marginal density of X [13]. Therefore, the shrinkage of the Bayes estimator is controlled by the derivative of the log marginal likelihood. By Theorem 2.1, the ANHIB prior has polynomially heavy tails. The next result shows that its posterior mean achieves asymptotic tail robustness, in the sense that the impact of the prior vanishes as the observations go to infinity.

Theorem 3.4: *Assume that* $X \sim N(\theta, 1)$ *and* $\theta \sim \text{ANHIB}(a_0, b_0, \alpha_0, \beta_0)$. *Then the marginal density of X satisfies* $\lim_{|x|\to\infty} \mathrm{d} \log m_{\text{ANHIB}}(x)/\,\mathrm{d}x = 0$, *where* $m_{\text{ANHIB}}(x)$ *is the marginal likelihood of X under the ANHIB prior. Therefore, the Bayes estimator*

$$\hat{\theta}_{\text{ANHIB}} = E_{\text{ANHIB}}(\theta \mid x) \to x, \quad \text{as } |x| \to \infty.$$

Proof: See Section A.4 in the Appendix. ■

There are two more remarks for the above theorem. First, Carvalho et al. [13] provided similar tail robustness results for the horseshoe estimator. However, their results were based on a fixed scale parameter τ. But in Theorem 3.4, τ is integrated out, as well as the hyperparameters a and b, in the marginal density. Second, the above theorem is not restricted to the ANHIB prior. In fact, the proof holds for all priors with a hypergeometric inverted-Beta distribution on θ and a hyper prior on (a, b, τ) such that the marginal density $m(x)$ and its derivative are finite.

4. Simulation studies

To investigate the risk properties of the Bayes estimator under the ANHIB prior, we conduct simulation studies with a wide variety of sparsity levels and signal sizes. We consider the maximum likelihood estimator (MLE) as well as the Bayesian estimators under the following common priors:

- The ANHIB prior defined jointly by (2)–(6) with $M \sim Gamma(20, 20)$ and $N \sim Beta(1, 1)$ (ANHIB(20, 20, 1, 1));
- The Horseshoe (HS) prior;
- The Strawderman-Berger (SB) prior;
- The Normal-Exponential-Gamma($\gamma = 1/4, \delta = 1$) prior (NEG(1/4, 1));
- The Normal-Exponential-Gamma($\gamma = 3/4, \delta = 1$) prior (NEG(3/4, 1)); and
- The Double-Exponential (DE) prior.

As mentioned in Section 2.1, this ANHIB(20, 20, 1, 1) prior is a good default choice because its hyper-prior on M places most of the mass between 1/2 and 3/2, which is the range of M values in most common priors, and its hyper-prior on N is a noninformative uniform distribution between 0 and 1. Also note that the SB prior can be viewed

as a Normal-Exponential-Gamma prior NEG($\gamma = 1/2, \delta = 1$). In general the smaller the parameter γ is, the flatter the NEG prior is. All the above priors include a scale variance σ^2 in the normal kernels.

For each scenario, we generate 100 replicates of samples and run MCMC for 20,000 iterations with a burn-in of the first 10,000 iterations. Then we calculate the posterior means and report the averages and standard errors of the sum of squared error (SSE)

$$\text{SSE} = \sum_{i=1}^{p} (\hat{\theta}_i - \theta_i)^2$$

over the 100 data sets.

We first investigate performances of these estimators across various combinations of sparsity levels and signal sizes. The observations $\mathbf{X}$ are generated from a normal distribution $\text{N}_{100}(\boldsymbol{\theta}, \mathbf{I}_{100})$. To reflect different sparsity levels, we set the proportion of signals in $\boldsymbol{\theta}$ to be $q = 5\%$, 10% and 20%, respectively. Moreover, following the simulation setting in [16,17], we assume that all the non-zero signals are of the same size, which is set to be 2, 5 and 8, respectively, reflecting small, medium and large signals. These fixed signal sizes help to verify the theoretical results in Section 3. The super-efficiency property in Section 3.2 can be verified by the simulation analysis at small signal sizes (e.g. 2), and the tail-robustness property in Section 3.3 can be verified by the simulation analysis at large signal sizes (e.g. 8). The averaged SSEs of the estimates and their standard errors are reported in Table 2. The best methods (accounting for the standard errors) are highlighted in bold. Our simulation results suggest:

(1) For sparse data with $q = 5\%$ or 10%, no matter how large the signal sizes are, the Bayesian estimates under the ANHIB and horseshoe priors perform the best, because they provide the most suppression of noises.
(2) For non-sparse data with $q = 20\%$:
   - When the signal sizes are small, the observations of signals and noises are mixed. The double-exponential prior, which places the most prior mass at small non-zero values, has the smallest SSEs, but the ANHIB prior still yields much smaller SSEs than the NEG(1/4, 1) prior and the MLE;
   - When the signal sizes are medium or large, the double-exponential prior tends to over-shrink the large signals and thus yields inferior estimation performance, while the ANHIB prior yields the lowest SSEs (accounting for the standard errors). Furthermore, it is worth noting that the ANHIB estimator also outperforms the horseshoe estimator in these scenarios. In particular, when $q = 20\%$ and the signal size is 8, it could reduce the SSE of the horseshoe estimator by as much as 22%!

The theoretical properties in Section 3 suggest that the ANHIB estimator is super-efficient at suppressing noises and robust for large signals. Therefore, at small signal sizes (e.g. 2) and high sparsity levels (5–10%), a large number of noises are effectively suppressed while the shrinkage of small signals does not incur large errors. The small SSEs of the ANHIB estimator in these situations help to verify that it effectively suppresses the noises around 0. Similarly, at large signal sizes (e.g. 8) and medium sparsity levels (10–20%), the large signals are robustly identified and estimated, while the noises are separated from the

Table 2. Average SSEs and their standard errors of the MLEs and Bayesian estimates under the ANHIB, HS, SB, NEG(1/4, 1), NEG(3/4, 1) and DE priors with various sparsity levels and signal sizes.

Signal size	Small (2)			Medium (5)			Large (8)		
Signal percentage	5%	10%	20%	5%	10%	20%	5%	10%	20%
ANHIB	**16.49**	**30.64**	64.14	**23.14**	**36.00**	**55.31**	**21.49**	**31.40**	**51.15**
	0.32	0.49	0.88	0.86	0.90	1.05	1.01	0.97	1.28
HS	**16.30**	**29.85**	61.87	**23.03**	**36.75**	59.09	**22.59**	36.91	65.73
	0.34	0.49	0.71	0.85	0.90	1.09	0.99	0.98	1.19
SB	65.93	66.96	68.00	67.05	64.79	63.12	68.14	64.11	63.95
	0.98	0.96	0.99	0.98	0.98	0.97	1.11	0.96	1.01
NEG(1/4,1)	83.33	84.27	84.20	85.15	82.99	80.38	86.92	83.08	82.46
	1.15	1.11	1.17	1.14	1.18	1.15	1.33	1.22	1.24
NEG(3/4,1)	52.69	54.16	57.09	53.18	51.50	**53.36**	53.71	50.20	**51.59**
	0.83	0.83	0.87	0.84	0.84	0.93	0.95	0.79	0.91
DE	39.60	44.82	**50.98**	70.73	79.52	83.96	83.97	87.84	91.28
	1.17	1.06	0.91	1.18	1.08	1.10	1.25	1.22	1.31
MLE	98.73	99.54	99.53	100.98	99.16	97.12	102.82	99.96	99.86
	1.29	1.21	1.39	1.27	1.38	1.40	1.52	1.49	1.48

Notes: The data are generated from $N_{100}(\boldsymbol{\theta}, \mathbf{I}_{100})$, where all the non-zero signals are set to be of the same size. For each estimator, we use 100 replicates of samples. The upper row reports the averaged SSEs and the lower row reports their standard errors. The lowest SSEs (accounting for the standard errors) for each scenario are highlighted in bold.

signals and shrunk to zero by small local scale parameters. The superior performance of the ANHIB estimator with small SSEs in these situations also helps to verify that it provides robust estimation to the large signals with little shrinkage. On the other hand, when the signals are medium-sized (e.g. 5), the observations of the signals and noises tend to be mixed together. It is not as easy to distinguish the signals from noises as in the large signal situation, and the costs of over-shrinking the signals are not as negligible as in the small signal situation. Therefore, the relatively large SSEs at medium signals are consistent with the theoretical properties of the ANHIB estimator.

The Bayesian estimates under the three NEG(γ, 1) priors (including the Strawderman-Berger prior as a special case) have quite stable SSEs across various configurations of sparsity levels and signal sizes, and the performance improves as the shape parameter γ increases, which is consistent with the results in [13].

Next, we study the performance of the ANHIB estimator, compared with those of the competing Bayesian estimators and the MLE, in very high dimensional setups. Following the simulation setup of [16], we generate the observations $\mathbf{X}$ from $N_{1000}(\boldsymbol{\theta}, \mathbf{I}_{1000})$, where the mean $\boldsymbol{\theta}$ has 900 components set to be 0, 10 components set to be 10, and 90 components set to be a constant A. That is, the signal percentage is 10%. We let the constant A vary from 2 to 7, representing a range of signal sizes. Table 3 summarizes the averaged squared error losses and their standard errors of the estimates in this study. The best methods (accounting for the standard errors) are highlighted in bold. It is clear that the ANHIB and the horseshoe estimators provide substantially smaller risks than the other estimators in all considered scenarios. Furthermore, the ANHIB estimator is even better than the horseshoe estimator when the signal sizes are large (i.e. $A = 6, 7$), which is evidence of its adaptivity. As explained for Table 2, when A is small (e.g. 2), the SSEs of the ANHIB estimator are small because of its super-efficiency property; when A is large (e.g. 6, 7), the SSEs of the ANHIB estimator are small because of its tail-robustness property; when A is medium (e.g. 3–5), the SSEs are relatively large, which is due to mixing of the observations at the

Table 3. Average SSEs and their standard errors of the competing Bayesian estimators and the MLE in very high-dimensional setups.

A	2	3	4	5	6	7
ANHIB	**265.86**	**351.07**	393.86	**338.72**	**293.90**	**258.38**
	1.71	2.95	5.04	4.84	5.45	5.67
HS	276.30	**345.05**	**334.69**	355.48	368.88	356.48
	1.50	3.48	2.77	3.30	3.39	3.34
SB	671.54	673.20	661.52	661.21	660.92	643.73
	3.58	3.27	3.38	3.27	3.03	2.99
NEG(1/4,1)	845.11	851.20	842.32	845.77	845.25	827.30
	4.31	3.85	4.07	3.94	3.56	3.54
NEG(3/4,1)	541.94	544.89	530.77	525.79	523.69	506.44
	2.95	2.83	2.83	2.80	2.52	2.56
DE	694.43	828.57	907.35	954.57	968.26	957.76
	7.15	6.55	5.96	4.95	4.33	3.90
MLE	994.60	1007.61	998.76	1004.45	1003.93	984.79
	4.81	4.40	4.41	4.57	3.98	3.96

Notes: The data are generated from $N_{1000}(\boldsymbol{\theta}, \mathbf{I}_{1000})$, where the mean $\boldsymbol{\theta}$ has 900 components set to be 0, 10 components set to be 10, and 90 components set to be a constant A. The averages are over 100 replicates. For each estimator, the upper row reports the averaged SSEs and the lower row reports their standard errors. The lowest SSEs (accounting for the standard errors) for each scenario are highlighted in bold.

signals and noises. Similar patterns have been found in [16], where the Dirichlet–Laplace priors also yield larger SSEs when $A = 3, 4$.

Lastly, we conducted simulations where the signals of $\boldsymbol{\theta}$ are randomly drawn from a heavy-tailed Student's t-distribution with 3 degrees of freedom and scale parameter c, and then the observations $\mathbf{X}$ are generated from a normal distribution $\mathrm{N}_{100}(\boldsymbol{\theta}, \mathbf{I}_{100})$. Six configurations are investigated with various signal percentages ($q = 5\%, 10\%$ and 20%) and scale parameters ($c = 3, 6$). Table 4 summarizes the averaged squared error losses and their standard errors of the estimates in this study. The best methods (accounting for the standard errors) are highlighted in bold. It can be seen that the ANHIB and the horseshoe estimators

Table 4. Average SSEs and their standard errors of the MLEs and Bayesian estimates under the ANHIB, HS, SB, NEG(1/4, 1), NEG(3/4, 1) and DE priors with various sparsity levels and signal sizes.

Scale parameter	3			6		
Signal percentage	5%	10%	20%	5%	10%	20%
ANHIB	**13.72**	**23.90**	**39.55**	**15.55**	**27.22**	**46.22**
	0.67	0.81	0.91	0.71	0.80	1.18
HS	**13.62**	**23.83**	**39.85**	**15.76**	**28.28**	49.74
	0.66	0.81	0.94	0.73	0.81	1.25
SB	65.22	66.18	66.37	66.40	66.10	66.31
	1.02	0.98	1.17	1.05	1.04	1.07
NEG(1/4,1)	83.26	83.80	83.69	85.08	84.64	84.61
	1.25	1.16	1.37	1.32	1.28	1.27
NEG(3/4,1)	51.42	52.85	54.10	52.34	52.28	54.24
	0.85	0.83	1.02	0.86	0.85	0.94
DE	54.63	68.43	79.66	68.56	83.73	91.65
	1.88	1.34	1.36	1.81	1.33	1.31
MLE	98.90	99.44	99.12	101.10	100.78	102.39
	1.47	1.34	1.55	1.53	1.44	1.50

Notes: The data are generated from $N_{100}(\boldsymbol{\theta}, \mathbf{I}_{100})$, where all the non-zero signals are randomly drawn from a Student's t-distribution with 3 degrees of freedom. For each estimator, we use 100 replicates of samples. The upper row reports the averaged SSEs and the lower row reports their standard errors. The lowest SSEs (accounting for the standard errors) for each scenario are highlighted in bold.

provide substantially smaller risks than the other estimators in all considered scenarios. Furthermore, the ANHIB estimator is better than the horseshoe estimator when the scale parameter $c = 6$ and signal percentage $q = 20\%$, because there are a considerable number of large signals and the ANHIB estimator is more robust to large signals due to the heavier tails in the prior. These results are consistent with those in Tables 2 and 3.

In summary, the ANHIB estimator consistently provides accurate estimation across various sparsity levels and signal sizes. In particular, it substantially improves many common Bayesian shrinkage estimators when the sparsity level is high (e.g. 5% and 10% non-zero signals), or the signal size is reasonably large (e.g. ≥ 6).

5. Applications

In this section, we demonstrate the usage of the ANHIB prior in wavelet de-noising as well as in linear regression. The essential model setup for the wavelet de-noising case is equivalent to the normal model (1), while the inference of the regularized linear regression model is an extension of the normal mean estimation problem.

5.1. Wavelet de-noising

We first demonstrate the usage of the ANHIB prior in wavelet de-noising through the analysis of an electrocardiogram data set, which is available in the `R` package `wavelets`. It is a time series object of electrocardiogram measurements for an individual with arhythmia. Observations were made over an 11.37 second interval at a rate of 180 samples per second. There are in total 2, 048 millivolt readings. For more details of the data, see [26]. Polson and Scott [22,27] analysed the first 256 readings, which have been re-scaled to have a mean of zero. The standard deviation of the readings is approximately 0.2. Figure 3 shows the 256 millivolt readings against time. Only few absolute values of the data exceed 0.4 (two standard deviations of the data), which indicates that the data is highly sparse.

Following the framework in [22,27], we regard these data points as the 'true' function f sampled at equi-spaced intervals, and simulate noisy observations of f via $y_i = f_i + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$ for $i = 1, \ldots, 256$. We construct 100 simulated data sets each for three different noise levels: $\sigma = 0.1$, $\sigma = 0.2$ and $\sigma = 0.4$. These three different σ's reflect low, medium and high noise level, respectively. Then both the true values f_i's and the observations y_i's are re-scaled as in [27]. In our case, the scaling factor is 3. For convenience, we keep the notation f_i and y_i for the re-scaled data in the following.

The discrete wavelet transform (DWT) of the ECG data f_i's and the noisy observations y_i's yields β_{jk}'s and d_{jk}'s, where β_{jk} and d_{jk} represents the kth coefficient of the DWT at resolution level j of the true vales f_i's and noisy observations y_i's, respectively. Following the practice in [28], we assume the model for the data in the wavelet domain is $d_{jk} = \beta_{jk} + \nu_{jk}$, where ν_{jk}'s are independent and identically distributed normal noises. We place shrinkage priors on β_{jk}, and obtain the posterior mean estimate $\hat{\beta}_{jk}$ of β_{jk}. Then we can also get the estimates $\hat{f}_i$ of f_i through the inverse discrete wavelet transform (IDWT) of the estimated coefficients $\hat{\beta}_{jk}$'s.

Our goal is to assess the performance of the ANHIB estimator against the other shrinkage estimators and the DWT method. Performance is evaluated by the sum of squared error

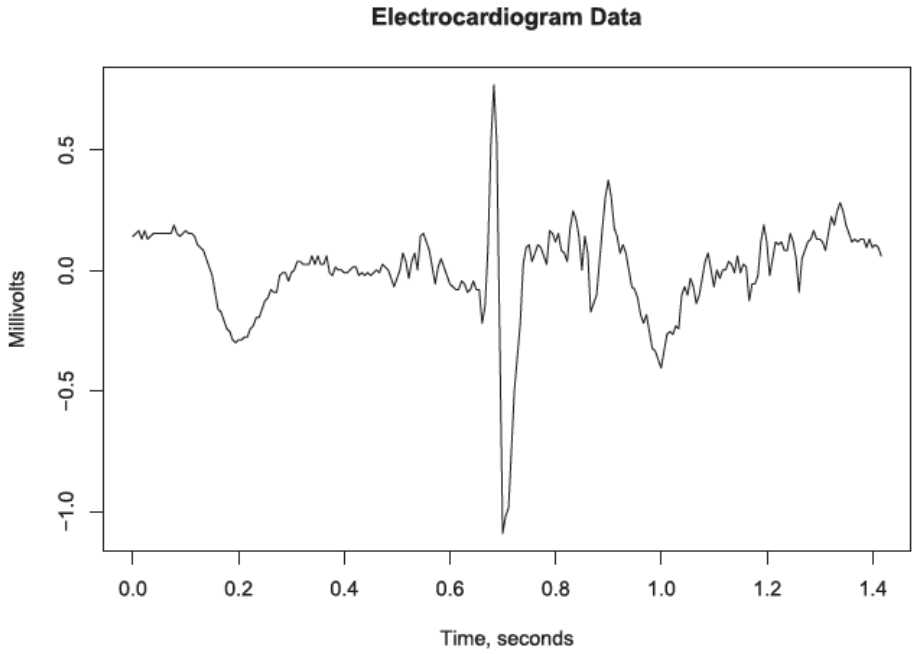

Figure 3. The 256 electrocardiogram readings against time.

Table 5. Average SSEs and their standard errors for the wavelet-denoising experiment under three different noise levels among different methods.

	ANHIB	HS	SB	NEG(1/4,1)	NEG(3/4,1)	DE	DWT
$\sigma = 0.1$	**7.78**	**7.64**	15.06	18.93	12.28	18.44	22.40
	0.11	0.11	0.14	0.17	0.12	0.20	0.19
$\sigma = 0.2$	**20.22**	**20.08**	59.83	75.53	48.03	52.59	89.03
	0.34	0.35	0.58	0.67	0.51	0.83	0.75
$\sigma = 0.4$	**48.79**	**49.08**	240.36	304.75	190.72	120.75	358.89
	0.95	0.98	2.20	2.78	1.77	3.14	3.30

Notes: The upper row reports the averaged SSEs and the lower row reports their standard errors. The lowest SSEs (accounting for the standard errors) for each scenario are highlighted in bold.

(SSE) in the data domain: $\text{SSE} = \sum_{i=1}^{256}(\hat{f}_i - f_i)^2$. Table 5 shows the averages and standard errors of the SSEs over the 100 data sets under three different noise levels among different methods. These results are consistent with those of the highly sparse cases in Section 4. We can see that the ANHIB estimator and the HS estimator outperform the others in all these settings. The SSE of the ANHIB estimator is slightly smaller than that of the HS estimator when the noise level is high. In the high noise level scenario, the observations tend to have large values, and the ANHIB estimator provides little shrinkage to large observations of the signals.

5.2. Linear regression

We further demonstrate the usage of the ANHIB prior in linear regression through the analysis of a prostate cancer data set, which is available in the `R` package `ElemStatLearn`. It has been investigated by Stamey et al. [29], Zou and Hastie [5] and Li and Lin [30]. The response variable of interest is lpsa (prostate specific antigen), and there

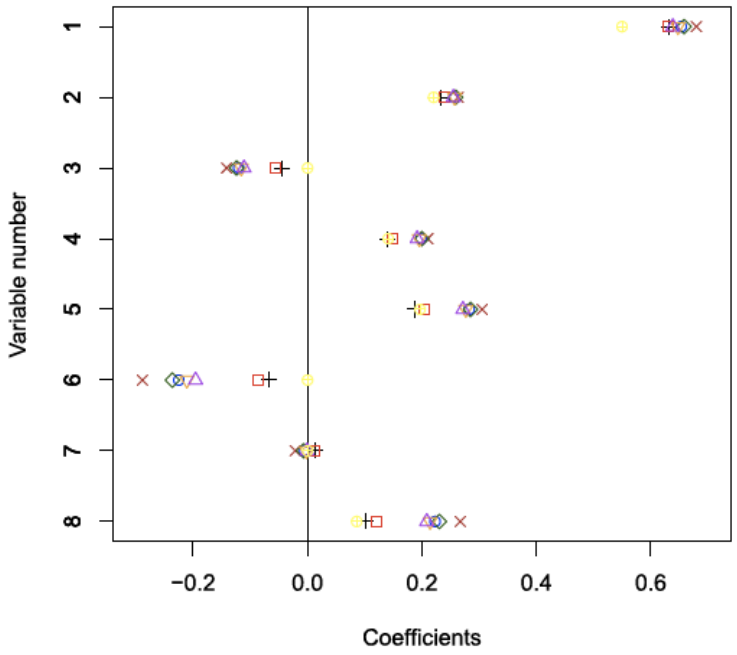

Figure 4. The comparisons of the ANHIB estimator (+), the HS estimator (□), the SB estimator (°), the NEG(1/4,1) estimator (◇), the NEG(3/4,1) estimator (▽), the DE estimator (△), the LASSO estimator (⊕), and the OLS estimator (×) for the prostate cancer data. The horizontal axis represents the coefficients and the vertical axis represents the variable number.

are 8 predictors of clinical measures – lcavol (log of the cancer volume), lweight (log of the prostate weight), age, lbph (log of the benign prostatic hyperplasia amount), svi (seminal vesicle invasion), lcp (log of the capsular penetration), gleason (Gleason score) and pgg45 (percentage Gleason scores 4 or 5). In total, we have 97 observations.

Following the practice in [5,30], we divide the data into a training set with 67 observations and a test set with 30 observations. Seven estimators are calculated based on the training set – the Bayesian estimates under the ANHIB, HS, SB, NEG(1/4,1), NEG(3/4,1) priors, as well as the LASSO and OLS estimates.

The coefficient estimates for the eight predictors are plotted in Figure 4. These estimates range between −0.3 and 0.7. The OLS estimates are always the furthest from 0, while the LASSO estimates are exactly 0 for three coefficients, and the Bayesian estimates have magnitudes between the OLS and LASSO ones. The ANHIB and the horseshoe estimates are very close for most of the coefficients.

Table 6 details the values of the coefficient estimates and uses $*$ to mark those significantly different from 0 based on the 95% posterior credible intervals. The results suggest that only two predictors, lcavol and lweight, are significant under the ANHIB and horseshoe priors, while more are significant under the other methods. Furthermore, the table reports the predictive mean squared errors (PMSEs) under these models over the 30 test observations. The Bayesian models under the ANHIB and horseshoe priors and the LASSO model provide substantially smaller PMSEs than the other methods, which is a consequence of their shrinkage of the coefficient estimates.

Table 6. The coefficient estimates and PMSEs under different methods for the prostate cancer data based on a linear regression model.

	ANHIB	HS	SB	NEG(1/4,1)	NEG(3/4,1)	DE	LASSO	OLS
intercept	2.470	2.470	2.467	2.465	2.468	2.467	2.464	2.465
lcavol	0.632*	0.630*	0.653*	0.658*	0.647*	0.638*	0.550*	0.680
lweight	0.234*	0.238*	0.256*	0.259*	0.257*	0.255*	0.220*	0.263
age	−0.046	−0.056	−0.120	−0.124	−0.116	−0.111	0.000	−0.141
lbph	0.140	0.149	0.198*	0.200*	0.195*	0.192	0.141*	0.210
svi	0.188	0.204	0.282*	0.285*	0.277*	0.272*	0.197*	0.305
lcp	−0.066	−0.087	−0.225	−0.236	−0.211	−0.196	0.000	−0.288
gleason	0.012	0.013	−0.005	−0.007	−0.002	−0.001	0.000	−0.021
pgg45	0.102	0.120	0.223	0.230	0.214	0.208	0.086*	0.267
PMSE	0.452	0.457	0.499	0.503	0.495	0.491	0.459	0.521

6. Discussion

Bayesian regularization methods provide powerful tools for high dimensional estimation in sparsity settings. However, many shrinkage priors are based on certain assumptions of sparsity levels, which could be hard to verify in practice. Violation of these assumptions might lead to unsatisfactory estimation performances. The major thrust of this paper is to propose a class of adaptive normal-hypergeometric-inverted-Beta (ANHIB) priors, which include many common shrinkage priors as special cases and provide accurate and robust estimation under various sparsity levels and signal sizes.

We establish the super-efficiency and tail-robustness properties of the Bayes estimator under the ANHIB priors, and recommend using the ANHIB prior with $M \sim Gamma(20, 20)$ and $N \sim Beta(1, 1)$ as a default choice. Through extensive simulation studies, we show that this ANHIB estimator substantially improves many common Bayesian shrinkage estimators when the sparsity level is high or the signal size is reasonably large.

The implications of this work will extend beyond the normal mean estimation problem. For example, as shown in the real data analysis in Section 5, the ANHIB priors can be readily applied to linear regression models, or even nonparametric regression models in the future. Datta and Ghosh [18] and Ghosh et al. [31] proposed asymptotically optimal hypothesis testing rules under sparsity for multivariate normal means with the horseshoe prior or the generalized double Pareto priors. Similar thresholding rules can be developed for the ANHIB priors to shrink small coefficient estimates to 0, and will be particularly useful for variable selection in regression models to produce robust and parsimonious models.

Acknowledgments

The authors thank the associate editor and the reviewers for their constructive suggestions. Hanjun is supported by the 2018 Scientific Research Foundation Project for Novice Teachers of Capital University of Economics and Business and Science Challenge Project under Award Number TZ2018001. Xinyi is supported by the National Science Foundation under Award Number 1613110.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Hanjun is supported by the 2018 Scientific Research Foundation Project for Novice Teachers of Capital University of Economics and Business and Science Challenge Project under Award Number TZ2018001. Xinyi is supported by the National Science Foundation under Award Number DMS 1613110 and DMS 2015552.

References

[1] Stein CM. Estimation of the mean of a multivariate normal distribution. Ann Stat. 1981;9(6):1135–1151.
[2] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970a;12(1):55–67.
[3] Hoerl AE, Kennard RW. Ridge regression: applications to nonorthogonal problems. Technometrics. 1970b;12(1):69–82.
[4] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol. 1996;58(1):267–288.
[5] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol. 2005;67(2):301–320.
[6] Zou H. The adaptive lasso and its oracle properties. J Amer Statist Assoc. 2006;101(476): 1418–1429.
[7] Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression. J Amer Statist Assoc. 1988;83(404):1023–1032.
[8] Berger JO. A robust generalized Bayes estimator and confidence region for a multivariate normal mean. Ann Statist. 1980;8(4):716–761.
[9] Strawderman WE. Proper Bayes minimax estimators of the multivariate normal mean. Ann Math Stat. 1971;42(1):385–388.
[10] Figueiredo MA. Adaptive sparseness for supervised learning. IEEE Trans Pattern Anal Mach Intell. 2003;25(9):1150–1159.
[11] Griffin JE, Brown PJ. Alternative prior distributions for variable selection with very many more variables than observations. Tech. Rep. University of Warwick; 2005.
[12] Griffin JE, Brown PJ. Inference with normal-gamma prior distributions in regression problems. Bayesian Anal. 2010;5(1):171–188.
[13] Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. Biometrika. 2010;97(2):465–480.
[14] Armagan A, Clyde M, Dunson DB. Generalized beta mixtures of Gaussians. In Advances in neural information processing systems; 2011. p. 523–531.
[15] Armagan A, Dunson DB, Lee J. Generalized double Pareto shrinkage. Statist Sinica. 2013;23(1):119–143.
[16] Bhattacharya A, Pati D, Pillai NS, et al. Dirichlet–Laplace priors for optimal shrinkage. J Amer Statist Assoc. 2015;110(512):1479–1490.
[17] Bhadra A, Datta J, Polson NG, et al. The horseshoe+ estimator of ultra-sparse signals the horseshoe+ estimator of ultra-sparse signals. Bayesian Analysis. 2017;12(4):1105–1131.
[18] Datta J, Ghosh JK. Asymptotic properties of Bayes risk for the horseshoe prior. Bayesian Anal. 2013;8(1):111–132.
[19] van der Pas S, Kleijn B, van der Vaart A. The horseshoe estimator: posterior concentration around nearly black vectors. Electron J Stat. 2014;8(2):2585–2618.
[20] Gradshteyn I, Ryzhik I. Table of integrals, series, and products. Cambridge: Academic Press; 1965.
[21] Gordy MB. A generalization of generalized beta distributions. Tech. Rep. Division of Research and Statistics, Division of Monetary Affairs, Federal Reserve Board; 1998.
[22] Polson NG, Scott JG. Alternative global–local shrinkage rules using hypergeometric–beta mixtures. Tech. Rep. No.Tech. Rep. 14. Department of Statistical Science, Duke University; 2009.

[23] Polson NG, Scott JG. Large-scale simultaneous testing with hypergeometric inverted-beta priors. arXiv:1010.5223; 2010a.
[24] Polson NG, Scott JG. On the half-cauchy prior for a global scale parameter. Bayesian Anal. 2012;7(4):887–902.
[25] Clarke BS, Barron AR. Information-theoretic asymptotics of Bayes methods. IEEE Trans Inf Theory. 1990;36(3):453–471.
[26] Percival DB, Walden AT. Wavelet methods for time series analysis. Cambridge, England: Cambridge University Press; 2000.
[27] Polson NG, Scott JG. Shrink globally, act locally: sparse Bayesian regularization and prediction. Bayesian Statist. 2010b;9:501–538.
[28] Clyde M, George EI. Flexible empirical Bayes estimation for wavelets. J R Stat Soc Ser B Stat Methodol. 2000;62(4):681–698.
[29] Stamey TA, Kabalin JN, McNeal JE, et al. Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. radical prostatectomy treated patients. J Urol. 1989;141(5):1076–1083.
[30] Li Q, Lin N. The Bayesian elastic net. Bayesian Anal. 2010;5(1):151–170.
[31] Ghosh P, Tang X, Ghosh M, et al. Asymptotic properties of Bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. Bayesian Anal. 2016;11(3):753–796.
[32] Abramowitz M, Stegun IA. Handbook of mathematical functions: with formulas, graphs, and mathematical tables. North Chelmsford: Courier Corporation; 1964.
[33] Slater LJ. Confluent hypergeometric functions. Cambridge: Cambridge University Press; 1960.

Appendix. Proofs of the main theorems

In this Appendix, we provide the proofs of the main theorems in the paper.

A.1 Proof of Theorem 1

The ANHIB prior is a scale mixture of normals. Conditional on the variance λ^2, the distribution of θ has the density

$$p(\theta \mid \lambda^2) = \frac{1}{(2\pi\lambda^2)^{1/2}} \exp\left(-\frac{\theta^2}{2\lambda^2}\right).$$

With $s = 0$ and $\tau^2 = 1$, the prior of λ can be represented by

$$p(\lambda^2 \mid a, b) = C(a, b)^{-1}(\lambda^2)^{b-1}(1+\lambda^2)^{-(a+b)},$$

where the normalizing constant $C(a, b) = \text{Beta}(a, b)\Phi_1(b, 1, a+b, 0, 0)$ and Φ_1 is the degenerate hypergeometric function of two variables [20, 9.261]. The hyper-parameters

$$a, b \sim \pi(a, b),$$

where $\pi(a, b)$ can be obtained from the representation (5) through re-parameterization. Therefore, the marginal density of θ is

$$\begin{aligned} p_{\text{ANHIB}}(\theta) &= \int_{a=0}^{\infty}\int_{b=0}^{\infty}\int_{\lambda^2=0}^{\infty} p(\theta \mid \lambda^2)p(\lambda^2 \mid a, b)\,\mathrm{d}\lambda^2\,\mathrm{d}\pi(a, b) \\ &= \frac{1}{\sqrt{2\pi}}\int_{a=0}^{\infty}\int_{b=0}^{\infty}\int_{\lambda^2=0}^{\infty} C(a, b)(\lambda^2)^{b-\frac{3}{2}}(1+\lambda^2)^{-(a+b)}\exp\left(-\frac{\theta^2}{2\lambda^2}\right)\mathrm{d}\lambda^2\,\mathrm{d}\pi(a, b), \end{aligned}$$

Letting $u = 1/\lambda^2$, we get

$$p_{\text{ANHIB}}(\theta) = \frac{1}{\sqrt{2\pi}}\int_{a=0}^{\infty}\int_{b=0}^{\infty}\int_{u=0}^{\infty} C(a, b)u^{a-\frac{1}{2}}(1+u)^{-(a+b)}\exp\left(-\frac{\theta^2 u}{2}\right)\mathrm{d}u\,\mathrm{d}\pi(a, b).$$

To derive the upper bound, we rewrite the density as

$$\begin{aligned} p_{\text{ANHIB}}(\theta) &= \frac{1}{\sqrt{2\pi}} \int_{a=0}^{\infty}\int_{b=0}^{\infty}\int_{u=0}^{\infty} C(a,b)u^{-\frac{1}{2}} \left(\frac{u}{1+u}\right)^{a}\left(\frac{1}{1+u}\right)^{b} \exp\left(-\frac{\theta^2 u}{2}\right) \mathrm{d}u\,\mathrm{d}\pi(a,b) \\ &< \frac{1}{\sqrt{2\pi}} \int_{a=0}^{\infty}\int_{b=0}^{\infty}\int_{u=0}^{\infty} C(a,b)u^{-\frac{1}{2}} \exp\left(-\frac{\theta^2 u}{2}\right) \mathrm{d}u\,\mathrm{d}\pi(a,b) \\ &= \frac{\Gamma(1/2)}{\sqrt{2\pi}\,(\theta^2/2)^{1/2}} \int_{a=0}^{\infty}\int_{b=0}^{\infty} C(a,b)\,\mathrm{d}\pi(a,b) = \frac{C_1}{|\theta|}, \end{aligned}$$

where $C_1 = \int_{a=0}^{\infty}\int_{b=0}^{\infty} C(a,b)\,\mathrm{d}\pi(a,b)$ and the inequality follows from the fact that $0 < u/(1+u) < 1$ for any $u > 0$.

For the lower bound, we restrict the values of a and b to the subregion $A = \{0 < a \le 1/2, 0 < b \le 1/2\}$, that is,

$$\begin{aligned} p_{\text{ANHIB}}(\theta) &= \frac{1}{\sqrt{2\pi}} \int_{a=0}^{\infty}\int_{b=0}^{\infty}\int_{u=0}^{\infty} C(a,b)u^{a-\frac{1}{2}}(1+u)^{-(a+b)} \exp\left(-\frac{\theta^2 u}{2}\right) \mathrm{d}u\,\mathrm{d}\pi(a,b) \\ &> \frac{1}{\sqrt{2\pi}} \int_{a=0}^{1/2}\int_{b=0}^{1/2}\int_{u=0}^{\infty} C(a,b)u^{a-\frac{1}{2}}(1+u)^{-(a+b)} \exp\left(-\frac{\theta^2 u}{2}\right) \mathrm{d}u\,\mathrm{d}\pi(a,b). \end{aligned}$$

Moreover, noting that $f(a,b) = u^{a-\frac{1}{2}}(1+u)^{-(a+b)}$ is decreasing in both a and b, we have $f(a,b) = u^{a-\frac{1}{2}}(1+u)^{-(a+b)} > f(1/2,1/2) = 1/(1+u)$ on this subregion A. Therefore,

$$\begin{aligned} p_{\text{ANHIB}}(\theta) &> \frac{1}{\sqrt{2\pi}} \int_{a=0}^{1/2}\int_{b=0}^{1/2}\int_{u=0}^{\infty} C(a,b)\frac{1}{1+u} \exp\left(-\frac{\theta^2 u}{2}\right) \mathrm{d}u\,\mathrm{d}\pi(a,b) \\ &> C_2 \log\left(1+\frac{4}{\theta^2}\right), \end{aligned}$$

where $C_2 = (1/2\sqrt{2\pi}) * \int_{a=0}^{1/2}\int_{b=0}^{1/2} C(a,b)\,\mathrm{d}\pi(a,b)$.

To study the behaviour of $p_{\text{ANHIB}}(\theta)$ as $\theta \to 0$, note that the above strategy can also be applied to the subregion $A(\infty,\delta_0) = \{0 < a < \infty, 0 < b \le \delta_0\}$ with any $\delta_0 > 0$, which leads to

$$\begin{aligned} p_{\text{ANHIB}}(\theta) &= \frac{1}{\sqrt{2\pi}} \int_{a=0}^{\infty}\int_{b=0}^{\infty}\int_{u=0}^{\infty} C(a,b)u^{a-\frac{1}{2}}(1+u)^{-(a+b)} \exp\left(-\frac{\theta^2 u}{2}\right) \mathrm{d}u\,\mathrm{d}\pi(a,b) \\ &> \frac{1}{\sqrt{2\pi}} \int_{a=0}^{\infty}\int_{b=0}^{\delta_0}\int_{u=0}^{\infty} C(a,b)u^{a-\frac{1}{2}}(1+u)^{-(a+b)} \exp\left(-\frac{\theta^2 u}{2}\right) \mathrm{d}u\,\mathrm{d}\pi(a,b) \\ &= \frac{1}{\sqrt{2\pi}} \int_{a=0}^{\infty}\int_{b=0}^{\delta_0} C(a,b)\Gamma\left(a+\frac{1}{2}\right) U\left(a+\frac{1}{2},\frac{3}{2}-b,\frac{\theta^2}{2}\right) \mathrm{d}\pi(a,b) \\ &> \frac{1}{\sqrt{2\pi}} \int_{a=0}^{\infty}\int_{b=0}^{\delta_0} C(a,b)\Gamma\left(a+\frac{1}{2}\right) U\left(a+\frac{1}{2},\frac{3}{2}-\delta_0,\frac{\theta^2}{2}\right) \mathrm{d}\pi(a,b), \end{aligned}$$

where $U(a,b,z) = \frac{1}{\Gamma(a)}\int_0^{\infty} e^{-zt}t^{a-1}(1+t)^{b-a-1}\,\mathrm{d}t$ is the confluent hypergeometric function. Using the following fact from formula 13.5.8 of [32]: as $|z| \to 0$, for any $1 < b < 2$

$$U(a,b,z) = \frac{\Gamma(b-1)}{\Gamma(a)} z^{1-b} + O(1),$$

and letting $0 < \delta_0 < 1/2$, we obtain that as $|\theta| \to 0$,

$$U\left(a+\frac{1}{2},\frac{3}{2}-\delta_0,\frac{\theta^2}{2}\right) = \frac{\Gamma(1/2-\delta_0)}{\Gamma(1/2+a)}\left(\frac{\theta^2}{2}\right)^{\delta_0-1/2} + O(1).$$

Then we have that the marginal density under the ANHIB prior satisfies

$$p_{\text{ANHIB}}(\theta) > \frac{1}{\sqrt{2\pi}} \int_{a=0}^{\infty} \int_{b=0}^{\delta_0} C(a,b)\Gamma\left(a+\frac{1}{2}\right)\left[\frac{\Gamma(1/2-\delta_0)}{\Gamma(1/2+a)}\left(\frac{\theta^2}{2}\right)^{\delta_0-1/2} + O(1)\right] \mathrm{d}\pi(a,b)$$
$$= C_3|\theta|^{2\delta_0-1} + O(1),$$

where $C_3 = 2^{-\delta_0}\Gamma(1/2-\delta_0)/\sqrt{\pi} * \int_{a=0}^{\infty}\int_{b=0}^{\delta_0} C(a,b)\,\mathrm{d}\pi(a,b)$.

A.2 Proof of Theorem 2

At any θ value where the ANHIB prior is bounded, it is easy to see that

$$\int_{\theta}^{\theta+\frac{1}{\sqrt{n}}} p_{\text{ANHIB}}(\theta)\,\mathrm{d}\theta \geq O\left(\frac{1}{\sqrt{n}}\right).$$

Setting $\varepsilon = 1/n$ in Lemma 1 yields that when the prior is bounded around θ_0,

$$R_{n,\,\text{ANHIB}}(\theta_0) \leq \frac{1}{n} - \frac{1}{n}\log\left(\nu(A_\varepsilon)\right) \leq \frac{1}{n} - \frac{1}{n}\left[-\frac{1}{2}\log n + \log\left(O(1)\right)\right]$$
$$= \frac{\log n}{2n} + O\left(\frac{1}{n}\right).$$

A.3 Proof of Theorem 3

Under the ANHIB prior, at $\theta_0 = 0$, by part (2) of Theorem 1,

$$\int_0^{\frac{1}{\sqrt{n}}} p_{\text{ANHIB}}(\theta)\,\mathrm{d}\theta \geq \int_0^{\frac{1}{\sqrt{n}}} \left[C_3|\theta|^{2\delta_0-1} + C_4\right]\mathrm{d}\theta$$
$$= \frac{C_3}{2\delta_0}\left(\frac{1}{\sqrt{n}}\right)^{2\delta_0} + C_4\frac{1}{\sqrt{n}}$$
$$= \left(\frac{1}{\sqrt{n}}\right)^{2\delta_0} O(1),$$

for any constant $0 < \delta_0 < 1/2$. Setting $\varepsilon = 1/n$ in Lemma 1 yields that when $\theta_0 = 0$,

$$R_{n,\,\text{ANHIB}}(0) \leq \frac{1}{n} - \frac{1}{n}\log\left(\nu(A_\varepsilon)\right) \leq \frac{1}{n} - \frac{1}{n}\left[-\delta_0\log n + \log\left(O(1)\right)\right]$$
$$= \frac{\delta_0\log n}{n} + O\left(\frac{1}{n}\right).$$

A.4 Proof of Theorem 4

Integrating out θ in the likelihood with respect to a hypergeometric inverted-Beta prior HIB$(a,b,\tau,s=0)$, we obtain

$$m(x\,|\,a,b,\tau) = \int \frac{C^{-1}}{\sqrt{2\pi}(1+\lambda^2)^{1/2}} \exp\left\{-\frac{x^2}{2(1+\lambda^2)}\right\}(\lambda^2)^{b-1}(1+\lambda^2)^{-(a+b)}$$
$$\left\{\frac{1}{\tau^2} + \left(1-\frac{1}{\tau^2}\right)\frac{1}{1+\lambda^2}\right\}^{-1} \mathrm{d}\lambda^2,$$

where $C = \text{Beta}(a,b)\Phi_1(b,1,a+b,0,1-1/\tau^2)$ is the normalizing constant of the hypergeometric inverted-Beta prior and Φ_1 is the degenerate hypergeometric function of two variables [20, 9.261].

Letting $k = 1/(1+\lambda^2)$, this distribution can be rewritten as

$$\begin{aligned} m(x \mid a,b,\tau) &= \frac{C^{-1}}{\sqrt{2\pi}} \int \exp\left\{-\frac{kx^2}{2}\right\} k^{a-1/2}(1-k)^{b-1} \left\{\frac{1}{\tau^2} + \left(1-\frac{1}{\tau^2}\right)k\right\}^{-1} \mathrm{d}k \\ &= \frac{C^{-1}}{\sqrt{2\pi}} \operatorname{Beta}\left(a+\frac{1}{2}, b\right) \exp\left(-\frac{x^2}{2}\right) \Phi_1\left(b, 1, a+b+\frac{1}{2}, \frac{x^2}{2}, 1-\frac{1}{\tau^2}\right). \end{aligned}$$

Taking the derivative with respect to x, we obtain

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x} m(x \mid a,b,\tau) &= -\frac{C^{-1}x}{\sqrt{2\pi}} \int \exp\left\{-\frac{kx^2}{2}\right\} k^{a+1/2}(1-k)^{b-1} \left\{\frac{1}{\tau^2} + \left(1-\frac{1}{\tau^2}\right)k\right\}^{-1} \mathrm{d}k \\ &= -\frac{C^{-1}x}{\sqrt{2\pi}} \operatorname{Beta}\left(a+\frac{3}{2}, b\right) \exp\left(-\frac{x^2}{2}\right) \Phi_1\left(b, 1, a+b+\frac{3}{2}, \frac{x^2}{2}, 1-\frac{1}{\tau^2}\right). \end{aligned}$$

Under the ANHIB prior, the hyper-parameters follow a prior $\pi(a,b,\tau)$. Therefore, the marginal density of x is

$$m_{\mathrm{ANHIB}}(x) = \int m(x \mid a,b,\tau)\, \mathrm{d}\pi(a,b,\tau)$$

and the score function can be represented by

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}x} \log m_{\mathrm{ANHIB}}(x) &= \frac{\displaystyle\int \frac{\mathrm{d}}{\mathrm{d}x} m(x \mid a,b,\tau)\, \mathrm{d}\pi(a,b,\tau)}{\displaystyle\int m(x \mid a,b,\tau)\, \mathrm{d}\pi(a,b,\tau)} \\ &= \int \frac{\dfrac{\mathrm{d}}{\mathrm{d}x} m(x \mid a,b,\tau)}{m(x \mid a,b,\tau)}\, \mathrm{d}\pi(a,b,\tau \mid x) \\ &= -\frac{(a+1/2)x}{a+b+1/2} \int \frac{\Phi_1\left(b, 1, a+b+\frac{3}{2}, \frac{x^2}{2}, 1-\frac{1}{\tau^2}\right)}{\Phi_1\left(b, 1, a+b+\frac{1}{2}, \frac{x^2}{2}, 1-\frac{1}{\tau^2}\right)} \mathrm{d}\pi(a,b,\tau \mid x), \end{aligned} \tag{A1}$$

where $\pi(a,b,\tau \mid x)$ is the posterior distribution of a, b and τ, and the second line follows from the Bayes theorem.

Gordy [21] showed

$$\Phi_1(\alpha,\beta,\gamma,x,y) = \begin{cases} e^x \displaystyle\sum_{n=0}^{\infty} \frac{(\alpha)_n(\beta)_n}{(\gamma)_n} \frac{y^n}{n!} {}_1F_1(\gamma-\alpha, \gamma+n, -x), & \text{for } 0 \le y < 1 \text{ and } 0 < \alpha < \gamma \\ e^x (1-y)^{-\beta} \Phi_1\left(\gamma-\alpha, \beta, \gamma, -x, \dfrac{y}{y-1}\right), & \text{for } y < 0 \text{ and } 0 < \alpha < \gamma, \end{cases}$$

where $(a)_n$ is the rising factorial and ${}_1F_1(\alpha,\beta,x)$ is Kummer's function of the first kind. Moreover, Chapter 4 of [33] showed that for any real number x,

$${}_1F_1(a,b,x) = \begin{cases} \dfrac{\Gamma(a)}{\Gamma(b)} e^x x^{a-b}\{1+O(x^{-1})\}, & x > 0 \\ \dfrac{\Gamma(a)}{\Gamma(b-a)} (-x)^{-a}\{1+O(x^{-1})\}, & x < 0. \end{cases}$$

Thus, for $\tau \geq 1$,

$$\int \frac{\Phi_1\left(b, 1, a+b+\frac{3}{2}, \frac{x^2}{2}, 1-\frac{1}{\tau^2}\right)}{\Phi_1\left(b, 1, a+b+\frac{1}{2}, \frac{x^2}{2}, 1-\frac{1}{\tau^2}\right)} \mathrm{d}\pi(a, b, \tau \mid x)$$

$$= \int \frac{\sum_{n=0}^{\infty} \frac{(b)_n (1)_n}{(a+b+3/2)_n} \frac{(1-1/\tau^2)^n}{n!} \frac{\Gamma(a+3/2)}{\Gamma(b+n)} \left(\frac{x^2}{2}\right)^{-(a+3/2)} \left(1+O(x^{-2})\right)}{\sum_{n=0}^{\infty} \frac{(b)_n (1)_n}{(a+b+1/2)_n} \frac{(1-1/\tau^2)^n}{n!} \frac{\Gamma(a+1/2)}{\Gamma(b+n)} \left(\frac{x^2}{2}\right)^{-(a+1/2)} \left(1+O(x^{-2})\right)} \mathrm{d}\pi(a, b, \tau \mid x)$$

$$\leq \frac{2}{x^2} \int (a+1/2)\, \mathrm{d}\pi(a, b, \tau \mid x) \tag{A2}$$

where the inequality comes from the fact that $1/(a+b+3/2)_n \leq 1/(a+b+1/2)_n$ for any positive a, b and n. On the other hand, for $0 < \tau < 1$,

$$\int \frac{\Phi_1\left(b, 1, a+b+\frac{3}{2}, \frac{x^2}{2}, 1-\frac{1}{\tau^2}\right)}{\Phi_1\left(b, 1, a+b+\frac{1}{2}, \frac{x^2}{2}, 1-\frac{1}{\tau^2}\right)} \mathrm{d}\pi(a, b, \tau \mid x)$$

$$= \int \frac{\sum_{n=0}^{\infty} \frac{(a+3/2)_n (1)_n}{(a+b+3/2)_n} \frac{(1-\tau^2)^n}{n!} \frac{\Gamma(a+1)}{\Gamma(a+b+n+3/2)} \left(\frac{x^2}{2}\right)^{-(a+n+3/2)} \left(1+O(x^{-2})\right)}{\sum_{n=0}^{\infty} \frac{(a+1/2)_n (1)_n}{(a+b+1/2)_n} \frac{(1-\tau^2)^n}{n!} \frac{\Gamma(a)}{\Gamma(a+b+n+1/2)} \left(\frac{x^2}{2}\right)^{-(a+n+1/2)} \left(1+O(x^{-2})\right)}$$

$$\times \mathrm{d}\pi(a, b, \tau \mid x)$$

$$\leq \frac{2}{x^2}, \tag{A3}$$

where the inequality follows from the relationship

$$\sum_{n=0}^{\infty} \frac{(a+3/2)_n (1)_n}{(a+b+3/2)_n} \frac{(1-\tau^2)^n}{n!} \frac{\Gamma(a+1)}{\Gamma(a+b+n+3/2)}$$

$$= \sum_{n=0}^{\infty} \frac{(a+1/2)_n (1)_n}{(a+b+1/2)_n} \frac{(1-\tau^2)^n}{n!} \frac{\Gamma(a)}{\Gamma(a+b+n+1/2)} \cdot \frac{a(a+n+1/2)(a+b+1/2)}{(a+1/2)(a+b+n+1/2)^2}$$

and that the last fraction on the right is obviously smaller than 1 for any positive a, b and n. Combining (A1), (A2) and (A3) yields that for all $\tau > 0$, the score function is a polynomial of the order x^{-1}, which converges to 0 as $x \to \infty$.