BAYESIAN ADAPTIVE ESTIMATION OF HIGH DIMENSIONAL VECTORS

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Di Cao, M.S. of Statistics

Graduate Program in Department of Statistics

The Ohio State University

2014

Dissertation Committee:

Xinyi Xu, Advisor

Steven N. MacEachern

Christopher M. Hans

Abstract

The normal mean estimation problem has a central role in statistical estimation. The maximum likelihood estimator (MLE) is a traditional estimator for this problem. Under the squared error loss, the MLE is unbiased, minimax, best invariant, and admissible when the parameter's dimension is one or two. However, when the parameter's dimension is three or higher, it is inadmissible and can be uniformly dominated by other estimators.

In the past half century, a great amount of research has been devoted to developing optimal estimators in high dimensional settings. The Bayesian approaches are highly appealing for this problem. However, selecting a prior becomes challenging when the parameter's dimension is high. Subjective elicitation of prior knowledge is almost infeasible. Therefore, we have to turn to some formal methods. In this dissertation, we develop a class of new prior distributions, namely, the adaptive inverted-Beta (AIB) priors, that lead to Bayesian estimators that often outperform many common estimators in the literature when the parameter is high dimensional.

In the first part, Chapter 2, we focus on the situations where the observations are independent from a high dimensional normal distribution. We incorporate both global and local parameters in our AIB priors, so that different dimensions of the parameter can be shrunk by different factors. Most existing priors in the literature assume that data have a certain sparsity level. Instead, we utilize a hierarchical structure which allows the shrinkage powers of the corresponding Bayesian estimators to be adaptive to the data sparsity levels. We establish theoretical properties of the Bayesian estimators under the AIB priors. We show that they provide strong shrinkage to noise close to 0, while providing essentially no shrinkage to large signals. We also demonstrate the estimation performances of these Bayesian estimators in many simulation scenarios with different sparsity levels and different signal sizes, and compare the performances with those of many common estimators in the literature.

Then in the second part of this dissertation, Chapter 3, we extend the AIB priors to the linear regression settings. We consider both $n \geq p$ (that is, the number of observations is larger than the number of parameters) and $n < p$ (that is, the number of observations is smaller than the number of parameters) situations. For both situations, we conduct simulation studies to investigate the performances of the Bayesian estimators under the AIB priors. We further demonstrate the use of the AIB priors using NIR spectroscopy data.

In the end, Chapter 4, we summarize the main findings of our work and discuss potential extensions. In particular, we generalize the AIB priors for estimating the mean from multivariate normal distributions with general covariance structures. We again investigate the performances of the corresponding Bayesian estimators through simulation studies. We also discuss the connection between the normal mean estimation problem and the portfolio choice problem. We apply the generalized AIB prior to portfolio choice based on the Fama–French 25 portfolios data, and show that the resulting portfolios have superior performances.

I dedicate my dissertation to my family, my friends and my advisor.

Acknowledgments

First, I would like to thank my advisor, Dr. Xinyi Xu, for inspiring this project and advising me with endless patience. She is so generous to share her brilliant and creative ideas with me and direct me when I was lost. I sincerely appreciate the time she spent thinking about the observations I found and checking every page of this thesis. Without her great help, this research would never be done. I also would love to thank The Ohio State University, Department of Statistics for funding me as a graduate associate assistance during my study.

I also wish to thank Dr. Steven N. MacEachern and Dr. Chrisphor Hans, for serving in my dissertation committee, my candidacy exam committee, and their more than helpful comments and suggestions. They helped me understand my research problem much deeper.

At the end, I would like to thank my parents and my wife, Tongyao Cheng, for their love all the time.

Working with everyone who made this thesis possible is a precious experience to me.

This work was supported in part by the US National Science Foundation under award number DMS-12-09194.

Vita

June 16, 1985	Born - Tianjin, China
2008	B.S. Statistics, Nankai University
2011	M.S. Statistics, The Ohio State University
2012	Intern, American Credit Acceptance
2009-present	Graduate Teaching Associate, The Ohio State University.

Fields of Study

Major Field: Statistics

Table of Contents

Page

List of Tables

Table **Page**

List of Figures

Chapter 1: Introduction

1.1 The Normal Mean Estimation Problem in High-Dimensional Spaces

The normal mean estimation problem holds a central place in statistical estimation. Let $\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma} \sim \mathrm{N}_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ be a p-dimensional multivariate normal vector with unknown mean $\boldsymbol{\theta}$ and unknown positive definite covariance matrix $\boldsymbol{\Sigma}$. Based on observations $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$, this problem aims at estimating $\boldsymbol{\theta}$ under the invariant squared error loss

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}). \tag{1.1}$$

An estimator $\hat{\boldsymbol{\theta}}$ is evaluated by its expected loss or risk function

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \mathrm{E}_\theta L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}). \tag{1.2}$$

When the covariance matrix $\boldsymbol{\Sigma}$ is known or can be estimated independently, through a sufficiency reduction and a variable transformation, the problem is equivalent to estimating the normal mean $\boldsymbol{\theta}$ where the components of $\boldsymbol{X}$ are independent and have a common unknown variance σ^2, that is

$$\boldsymbol{X} \mid \boldsymbol{\theta}, \sigma \sim \mathrm{N}(\boldsymbol{\theta}, \sigma^2 I_p), \tag{1.3}$$

where I_p is the $p \times p$ identity matrix. The loss function is then reduced to

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{1}{\sigma^2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}). \tag{1.4}$$

For the majority of this dissertation, we assume the normal distributions to have the independent structure (1.3), and at the end, we generalize to normal distributions with general covariance structures.

In statistical estimation, there are several principles for choosing an estimator. The first principle is admissibility. Given a risk function, an estimator δ is called inadmissible if there exists another estimator δ' such that $R(\theta, \delta') \leq R(\theta, \delta)$ for all values of θ in the parameter space and $R(\theta, \delta') < R(\theta, \delta)$ for some θ. In such case, the estimator δ' is said to dominate the estimator δ. In contrast, an estimator δ is called admissible if it cannot be dominated by any other estimators. For the normal mean estimation problem under the model (1.3) with $\sigma^2 = 1$, Brown (1971) showed that all admissible estimators are (generalized) Bayesian estimators. The second principle is minimaxity. An estimator δ^* is said to be minimax with respect to a risk function $R(\theta, \delta)$, if it minimizes $\sup_\theta R(\theta, \delta)$, that is,

$$\sup_\theta R(\theta, \delta^*) = \inf_\delta \sup_\theta R(\theta, \delta). \tag{1.5}$$

Given one observation $\boldsymbol{X}$, a traditional estimator of the normal mean estimation problem is the maximum likelihood estimator (MLE) $\hat{\boldsymbol{\theta}}_{MLE} = \boldsymbol{X}$. It has many attractive properties. Under the likelihood (1.3) with $\sigma^2 = 1$, $\hat{\boldsymbol{\theta}}_{MLE}$ is unbiased, minimax, best invariant (Lehmann and Casella 1998) and admissible when $p = 1$ or 2 (Berger 1985). Its risk is a constant

$$E_\theta L(\boldsymbol{\theta}, \boldsymbol{X}) = E(\boldsymbol{\theta} - \boldsymbol{X})^T(\boldsymbol{\theta} - \boldsymbol{X}) = p\sigma^2. \tag{1.6}$$

Moreover, it is the Bayesian estimator under the improper uniform prior distribution $\pi(\boldsymbol{\theta}) = 1$. However, when p is 3 or higher, $\hat{\boldsymbol{\theta}}_{MLE}$ is inadmissible. In fact, Stein (1956) showed that for large p,

$$\|\boldsymbol{X}\|_2^2 = p + \|\boldsymbol{\theta}\|_2^2 + O_p(\sqrt{p + \|\boldsymbol{\theta}\|_2^2}), \tag{1.7}$$

where $X_n = O_p(a_n)$ means that the set of values X_n/a_n is stochastically bounded, i.e., for any ϵ, there exists A_ϵ such that for all n,

$$P(|X_n| \leq A_\epsilon a_n) \geq 1 - \epsilon.$$

Therefore, in high dimensional spaces, the L_2-norm of $\boldsymbol{X}$ is greater than the L_2-norm of the true parameter $\boldsymbol{\theta}$. In other words, the MLE $\boldsymbol{X}$ is outside the sphere at which $\boldsymbol{\theta}$ is located. Therefore, intuitively, appropriate shrinkages can yield better estimators. The estimation accuracy can be improved by trading a small increase of bias for a large reduction of variance. Also, the overall performance of the estimator vector can be improved by borrowing information from all components even if observations from different dimensions are assumed to be independent from each other. When the true parameter value is close to the shrinkage target, the potential risk reduction can be large. Most of the common improved estimators for the normal mean problem in the literature can be viewed as shrinkage estimators of the MLE to some constant values or some subspaces. In the following a few sections, we summarize them into a few classes.

1.2 Empirical Bayes Approaches: The James-Stein Estimator and Its Extensions

In the seminal work of James and Stein (1961), they proposed the famous James–Stein estimator for the normal mean problem: Assuming the likelihood function (1.3) with $\sigma^2 = 1$,

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{JS}} = \left(1 - \frac{p-2}{\|\boldsymbol{X}\|_2^2}\right)\boldsymbol{X}. \tag{1.8}$$

When $p \geq 3$, the risk of the James–Stein estimator is

$$E_\theta L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{\boldsymbol{JS}}) = E\|\hat{\boldsymbol{\theta}}_{\boldsymbol{JS}} - \boldsymbol{\theta}\|_2^2 = p - E\frac{(p-2)^2}{p-2+2K} < p = E_\theta L(\boldsymbol{\theta}, \boldsymbol{X}), \quad \forall \boldsymbol{\theta}, \tag{1.9}$$

where K follows a Poisson distribution with parameter $\|\boldsymbol{\theta}\|_2^2/2$. Therefore, the James–Stein estimator dominates the MLE $\boldsymbol{X}$ as $p \geq 3$.

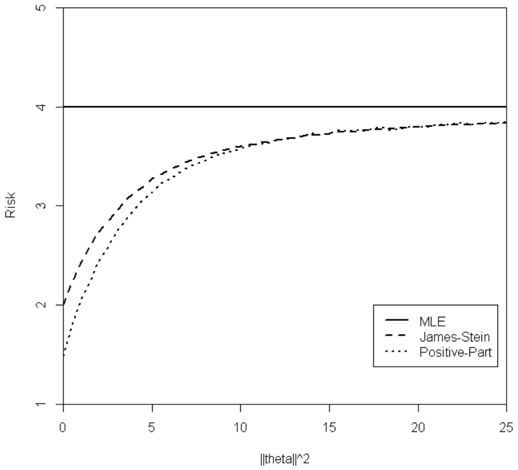

Figure 1.1: Risks of the James–Stein estimator, the James–Stein Positive-Part estimator and the MLE against the squared L_2-norm of $\boldsymbol{\theta}$ when $p = 4$.

Figure 1.1 plots the squared L_2-norm of $\boldsymbol{\theta}$ versus the risks of the MLE, the James–Stein estimator and the James–Stein Positive-Part estimator (introduced later) when $p = 4$ under squared error loss. It shows that the risk reduction can be large when $\boldsymbol{\theta}$ is close to the shrinking target $\mathbf{0}$. When the true value of $\boldsymbol{\theta}$ moves away from the shrinkage target, the risk of the James–Stein estimator approaches the risk of the MLE, but is always smaller. From the expression (1.8) we can see that each individual element's estimate borrows information from other dimensions, and the overall performance is improved.

The James–Stein estimator can be interpreted as an empirical Bayes estimator. Again assume that the likelihood function is (1.3) with $\sigma^2 = 1$, and put a p-dimensional multivariate normal prior distribution on $\boldsymbol{\theta}$: $\boldsymbol{\theta} \mid \tau \sim \mathrm{N}(\mathbf{0}, \tau^2 I)$. When τ is known, the posterior mean of $\boldsymbol{\theta}$ is:

$$E(\boldsymbol{\theta} \mid \boldsymbol{X}) = \frac{\tau^2}{\tau^2 + 1}\boldsymbol{X} = \left(1 - \frac{1}{\tau^2 + 1}\right)\boldsymbol{X}. \tag{1.10}$$

When τ is unknown, Judge and Bock (1978) showed that $(p-2)/\|\boldsymbol{X}\|_2^2$ is an unbiased estimator of $1/(\tau^2 + 1)$. Plugging it into (1.10) yields the James–Stein estimator.

The shrinkage target, or the prior mean in the empirical Bayes approach for the James–Stein estimator, does not have to be 0. Instead, it can be an arbitrary constant or a subspace. Suppose that our prior information suggests that $\boldsymbol{\theta}$ is close to a constant $\boldsymbol{\nu} \in \mathbb{R}^p$, then the James–Stein estimator shrinking towards $\boldsymbol{\nu}$ can be represented by

$$\hat{\boldsymbol{\theta}}_{JS}^{\nu} = \boldsymbol{\nu} + \left(1 - \frac{p-2}{\|\boldsymbol{X} - \boldsymbol{\nu}\|_2^2}\right)(\boldsymbol{X} - \boldsymbol{\nu}). \tag{1.11}$$

Same as above, the risk reduction is large when the true $\boldsymbol{\theta}$ is close to the shrinkage target $\boldsymbol{\nu}$, and diminishes as $\boldsymbol{\theta}$ moves away. Similarly, if our prior information suggests that the true $\boldsymbol{\theta}$ lies in a subspace $\boldsymbol{V} \subset \mathbb{R}^p$, then the James–Stein estimator can be modified as

$$\hat{\boldsymbol{\theta}}_{JS}^{V} = P_V\boldsymbol{X} + \left(1 - \frac{q-2}{(\boldsymbol{X} - P_V\boldsymbol{X})^T(\boldsymbol{X} - P_V\boldsymbol{X})}\right)(\boldsymbol{X} - P_V\boldsymbol{X}), \tag{1.12}$$

where $\boldsymbol{V}$ has dimension $p-q$ with $q \geq 2$, and P_V is a projection operator that projects from $\mathbb{R}^p$ into $\boldsymbol{V}$. For example, if the elements of $\boldsymbol{\theta}$ are believed to be the same, then the subspace $\boldsymbol{V}$ is $[\mathbf{1}]_p$. The James–Stein estimator shrinking towards this subspace can be written as:

$$\hat{\boldsymbol{\theta}}_{JS}^{V} = \bar{\boldsymbol{X}} + \left(1 - \frac{p-3}{(\boldsymbol{X}-\bar{\boldsymbol{X}})^T(\boldsymbol{X}-\bar{\boldsymbol{X}})}\right)(\boldsymbol{X}-\bar{\boldsymbol{X}}), \tag{1.13}$$

where $\bar{\boldsymbol{X}} = \bar{X}\mathbf{1}_p$ and $\bar{X}$ is the sample average of $\boldsymbol{X}$.

More generally, when vague or conflicting prior information suggests multiple shrinkage targets, that is, any one of the subspaces $\boldsymbol{V_1}, \cdots, \boldsymbol{V_K} \subset \mathbb{R}^p$ might be an appropriate shrinkage target, George (1986) developed a class of multiple shrinkage Stein estimators. Suppose that the K shrinkage targets are constants and denote the James–Stein positive-part estimator (introduced later) for the target k by δ_k, $k = 1, \cdots, K$. A multiple shrinkage Stein estimator can be written as

$$\delta_*(\boldsymbol{X}) = \sum_{k=1}^{K} \rho_k(\boldsymbol{X})\delta_k(\boldsymbol{X}), \tag{1.14}$$

where $\rho_1, \cdots, \rho_K$ satisfy $\sum_{k=1}^{K} \rho_k(\boldsymbol{X}) = 1$ and are adaptive weights of the shrinkage estimators $\delta_1, \cdots, \delta_K$. The author also showed that δ_* is minimax and able to offer substantial risk reduction at each target.

In the above discussion, we have assumed that σ^2 is known, without loss of generality, as 1. However, in practice, the variance of the normal likelihood is usually unknown. James and Stein (1961) provided the modified James–Stein estimators for such situations. Assume that $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\theta}, \sigma^2 I)$ where σ is unknown and S is another random variable, independent of $\boldsymbol{X}$, such that $S \sim \sigma^2\chi_n^2$. The James–Stein estimator can be modified as

$$\hat{\boldsymbol{\theta}}_{JS}^{\sigma} = \left(1 - \frac{\frac{p-2}{n+2}S}{\|\boldsymbol{X}\|_2^2}\right)\boldsymbol{X}. \tag{1.15}$$

The corresponding risk is

$$E_\theta L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS}^{\sigma}) = \sigma^2\left(p - \frac{n}{n+2}(p-2)^2 E\frac{1}{p-2+2K}\right), \tag{1.16}$$

where K follows a Poisson distribution with parameter $\|\boldsymbol{\theta}\|_2^2/2\sigma^2$. Thus, as $p \geq 3$, it has a lower risk than the risk of $\boldsymbol{X}$ which is $p\sigma^2$. Similarly, when $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is unknown, assume that S is another random variable, independent of $\boldsymbol{X}$, having a $p \times p$ Wishart distribution with n degrees of freedom and expectation $n\boldsymbol{\Sigma}$. Then the James–Stein estimator has the following form

$$\hat{\boldsymbol{\theta}}_{JS}^{\Sigma} = \left(1 - \frac{\frac{p-2}{n-p+3}}{\boldsymbol{X}^T S^{-1} \boldsymbol{X}}\right) \boldsymbol{X}. \tag{1.17}$$

Under the loss function (1.1), the risk is

$$E_\theta L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS}) = p - \frac{n-p+1}{n-p+3}(p-2)^2 E \frac{1}{p-2+2K}, \tag{1.18}$$

where K has a Poisson distribution with parameter $\frac{1}{2}\boldsymbol{\theta}^T \Sigma^{-1} \boldsymbol{\theta}$. Again it has a lower risk than the MLE as $p \geq 3$ (James and Stein 1961).

Although the James–Stein estimator and its extensions are good for the normal mean problem, they sometimes shrink too much and lead to an opposite sign as $\boldsymbol{X}$. To address this issue, James and Stein (1961) proposed using the following James–Stein Positive-Part estimator

$$\hat{\boldsymbol{\theta}}_{JS+} = \left(1 - \frac{p-2}{\|\boldsymbol{X}\|_2^2}\right)^+ \boldsymbol{X}, \tag{1.19}$$

where $(\bullet)^+ = \max(0, \bullet)$. Baranchik (1964) showed that this positive part estimator dominates the original James–Stein estimator when $p \geq 3$. However, it is still inadmissible as it is not smooth enough to be a generalized Bayesian estimator. Shao and Strawderman (1994) provided several estimators that further dominate the James–Stein Positive-Part estimator.

1.3 Full Bayesian Approaches Under Shrinkage Priors

1.3.1 Introduction

Bayesian estimators are studied for several reasons among which an important one is that any proper Bayes rule is admissible and so could not be uniformly improved upon. For the normal mean estimation problem, a specially successful strand of Bayesian approaches are through using scale mixture of normals priors in the form of

$$\pi(\boldsymbol{\theta}) = \int \mathrm{N}(\boldsymbol{\theta} \mid 0, \boldsymbol{\Lambda}) G(d\boldsymbol{\lambda}), \tag{1.20}$$

where $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_p)$, $\boldsymbol{\Lambda}$ is the $p \times p$ diagonal matrix $\mathrm{Diag}(\boldsymbol{\lambda})$ and $G(d\boldsymbol{\lambda})$ is the mixing distribution. When $\lambda_1 = \cdots = \lambda_p = \lambda$, this prior is called a global shrinkage prior and λ is called a global parameter. When λ_i's are not equal, this prior is called a local shrinkage prior and λ_i's are called local parameters. The Bayesian estimator under this prior and the squared error loss is the posterior mean of θ_i given κ_i

$$\mathrm{E}(\theta_i \mid \kappa_i, X_i) = (1 - \kappa_i) X_i, \tag{1.21}$$

where $\kappa_i = \sigma^2/(\sigma^2 + \lambda_i)$. As shown in (1.21), this Bayesian estimator can be viewed as a shrinkage estimator of $\boldsymbol{X}$. Under a global shrinkage prior, all components of $\boldsymbol{X}$ receive the same shrinkage degree, while under a local shrinkage prior, different components of $\boldsymbol{X}$ receive different degrees of shrinkage.

1.3.2 Global Shrinkage Priors

A simple example of global shrinkage priors is the prior $\pi(\boldsymbol{\theta}) = \mathrm{N}(\mathbf{0}, \tau^2 I)$ used in the construction of the James–Stein estimator. The global shrinkage parameter τ is estimated by an empirical Bayes method, and then plugged into equation (1.10) to yield the James–Stein estimator.

Using a full Bayesian approach, Strawderman (1971) slightly extended the class of minimax estimators in Baranchik (1964), and provided a class of Bayesian estimators under the following global shrinkage priors

$$\begin{aligned} \boldsymbol{\theta} \mid \tau &\sim \mathrm{N}_p\left(\mathbf{0}, \frac{1-\tau}{\tau} I\right), \\ \tau &\sim \mathrm{Beta}(1-a, 1). \end{aligned} \tag{1.22}$$

He showed that the Bayesian estimator, $\hat{\boldsymbol{\theta}}_a$, is minimax when $\frac{1}{2} \leq a < 1$ and $p = 5$, or when $0 \leq a < 1$ and $p \geq 6$. The commonly used Strawderman–Berger prior (Strawderman 1971, Berger 1980), is a special case of (1.22) with $a = 1/2$. Stein (1981) and Fourdrinier, Strawderman and Wells (1998) further investigated the construction of Bayesian minimax estimators, under a few classes of global shrinkage priors, including the Strawderman–type priors

$$\begin{aligned} \boldsymbol{\theta} \mid \tau &\sim \mathrm{N}(\mathbf{0}, \tau I), \\ h(\tau) &= c(\tau+1)^{1-(b+p/2)}, \end{aligned} \tag{1.23}$$

where c is the normalizing constant and $b \leq 0$. The corresponding Bayesian estimator is minimax when $2 - p/2 \leq b \leq 0$. Also they considered the shifted inverse gamma priors:

$$\begin{aligned} \boldsymbol{\theta} \mid \tau &\sim \mathrm{N}(\mathbf{0}, \tau I), \\ h(\tau) &= c \exp\left(-\frac{a}{\tau+1}\right)(\tau+1)^{1-(b+p/2)}, \end{aligned} \tag{1.24}$$

where c is the normalizing constant, $a > 0$ and $b \leq 0$. The corresponding Bayesian estimator is minimax when $2 - p/2 \leq b \leq 0$ as well.

In addition, Polson and Scott (2010) emphasized the appropriateness of using a half-Cauchy prior on the scale parameter τ instead of the usual conjugate choice of an inverse-gamma prior in the normal prior $\boldsymbol{\theta} \mid \tau \sim \mathrm{N}(\mathbf{0}, \tau^2 I)$. Their argument is that, consider the marginal likelihood of $\boldsymbol{X}$ as a function of τ, this marginal likelihood does not vanish as $\tau = 0$, therefore neither should the prior on τ. In contrast, the inverse-gamma prior density vanishes when $\tau = 0$. Furthermore, the authors extended the half-Cauchy distribution on τ to the class of hypergeometric inverted beta priors on τ^2:

$$p(\tau^2) = C^{-1}(\tau^2)^{b-1}(\tau^2+1)^{-(a+b)}\exp\left(-\frac{s}{\tau^2+1}\right)\left(\delta^2 + \frac{1-\delta^2}{\tau^2+1}\right)^{-1}, \tag{1.25}$$

where C is the normalizing constant, $a > 0$, $b > 0$, $\delta^2 > 0$, and s is a real number. Important summary statistics, for example, the posterior moments, marginal densities and frequentist risks, were provided under this family of priors.

1.3.3 Local Shrinkage Priors

When the true parameter $\boldsymbol{\theta}$ is high dimensional and the signal in it is sparse, global shrinkage priors are not appropriate because they provide the same degree of shrinkage to all coordinates. On the other hand, as seen in (1.21), local shrinkage priors allow different shrinkage degrees for different coordinates. Therefore, they have been widely used in high dimensional estimation and in variable selection for regression models. A common local shrinkage prior is the "spike-and-slab" prior provided by Mitchell and Beauchamp (1988), which is a mixture of a point mass at 0 and a uniform distribution:

$$\pi(\theta_i) = h_{0i} I_{\theta_i=0} + \frac{1-h_{0i}}{2f_i} I_{-f_i \le \theta_i \le f_i}. \tag{1.26}$$

The bounds of the uniform distribution, $\pm f_i$, are assumed to be large to express prior uncertainty. The weight of the point mass at 0 can be determined by the

Bayesian cross-validation method or the goodness-of-fit plot. George and McCulloch (1997) proposed a similar prior by replacing the uniform "slab" by a normal "slab". Their work was further extended to multivariate regressions in Brown and Vannucci (1998). Johnstone and Silverman (2004) considered other densities as the "slab", for example, a double exponential distribution or a distribution with tails that decay at polynomial rate. The probability mass at zero of the i-th component, h_{0i}, is estimated by $\hat{h}_{0i}$, which is the maximizer of the density of X_i conditional on h_{0i}. Then $\hat{\boldsymbol{h}}_0 = (\hat{h}_{01}, \cdots, \hat{h}_{0p})$ is plugged back into the prior of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}$ is estimated using the posterior median or the posterior mean.

It is interesting to note that some widely used estimators in the literature, although not constructed as Bayesian estimators, also have Bayesian interpretations. For example, it is well known that Tibshirani's Lasso (Tibshirani 1996) can be interpreted as the posterior mode under a double exponential (Laplacian) prior distribution on $\boldsymbol{\theta}$, that is,

$$\begin{aligned} \boldsymbol{X} \mid \boldsymbol{\theta}, \sigma^2 &\sim \mathrm{N}(\boldsymbol{\theta}, \sigma^2 I), \\ \theta_i \mid \tau &\overset{iid}{\sim} \mathrm{DE}(\tau). \end{aligned} \tag{1.27}$$

This prior is in fact a scale mixture of normals with local parameters, where the local parameters follow independent exponential distributions, that is, the prior can be represented by the following hierarchical form

$$\begin{aligned} \boldsymbol{\theta} \mid \boldsymbol{\lambda} &\sim \mathrm{N}(\boldsymbol{0}, \mathrm{Diag}(\boldsymbol{\lambda})), \\ \lambda_i \mid \tau &\overset{iid}{\sim} \mathrm{Exp}\left(\frac{\tau}{2}\right). \end{aligned} \tag{1.28}$$

In the models (1.27) and (1.28), the hyperparameter τ controls the shrinkage degree of the Bayesian estimator. In contrast to fixing this hyperparameter at a pre-specified constant, Figueiredo (2003) proposed the Normal-Jeffreys prior which removes τ by placing the Jeffreys prior on λ_i in (1.28). Integrating out the hyperparameter τ yields

$$\begin{aligned} \boldsymbol{\theta} \mid \boldsymbol{\lambda} &\sim \mathrm{N}(0, \mathrm{Diag}(\boldsymbol{\lambda})), \\ p(\lambda_i) &\propto \frac{1}{\lambda_i}, \end{aligned} \tag{1.29}$$

and if we further integrate out λ_i, the prior on θ_i can be written as

$$p(\theta_i) \propto \frac{1}{\mid \theta_i \mid}. \tag{1.30}$$

From (1.30) we can see that the Normal-Jeffreys prior has an infinite peak at 0 and also thick tails. This shape property offers strong shrinkage around 0 and little shrinkage for large signals. We will see the details in Chapter 2.

As extensions of the double-exponential prior or the normal-exponential prior, Griffin and Brown (2005) proposed another two priors which are members in the family of scale mixture of normals. The first one is called the normal-gamma (NG) prior, which generalizes the exponential hyperprior in (1.28) to a gamma distribution:

$$\begin{aligned} \boldsymbol{\theta} \mid \boldsymbol{\lambda} &\sim \mathrm{N}(\mathbf{0}, \mathrm{Diag}(\boldsymbol{\lambda})), \\ \lambda_i \mid \gamma, \tau &\overset{iid}{\sim} \mathrm{Gamma}(\gamma, \tau). \end{aligned} \tag{1.31}$$

The exponential density can be regained from the gamma distribution with the shape parameter γ fixed at 1. An important motivation for this generalization is that the double exponential prior has only one hyperparameter to control the shrinkage power, which is restrictive, for example, a known issue of the double exponential prior is the over shrinkage of large signals. By using the gamma distribution with both shape and scale parameters, the normal gamma prior can provide more flexible shrinkage patterns under different parameter specifications (more details are given in Chapter 2). The second generalized family of priors is known as the normal-exponential-gamma (NEG) prior which puts a gamma hyperprior on the scale parameter of the exponential mixing distribution:

$$\begin{aligned} \boldsymbol{\theta} \mid \boldsymbol{\lambda} &\sim \mathrm{N}(\mathbf{0}, \mathrm{Diag}(\boldsymbol{\lambda})), \\ \lambda_i \mid \tau &\overset{iid}{\sim} \mathrm{Exp}(\tau), \\ \tau \mid \gamma, \delta &\sim \mathrm{Gamma}(\gamma, \delta). \end{aligned} \tag{1.32}$$

Interestingly, there are another two ways to rewrite (1.32). In the first way, integrating out τ leads to the prior of λ_i conditional on γ and δ:

$$\begin{aligned} \boldsymbol{\theta} \mid \boldsymbol{\lambda} &\sim \mathrm{N}(\mathbf{0}, \mathrm{Diag}(\boldsymbol{\lambda})), \\ \lambda_i \mid \gamma, \delta &\sim \frac{\gamma}{\delta}(1 + \frac{\lambda_i}{\delta})^{-(\gamma+1)}. \end{aligned} \tag{1.33}$$

In Chapter 2 we will see that the prior (1.33) is a special case of the inverted-Beta prior which is defined later. The second way comes from the fact that the parameter τ in (1.32) is a global scale parameter. Therefore, we can also represent the NEG prior in (1.32) as

$$\begin{aligned} \boldsymbol{\theta} \mid \tau, \boldsymbol{\lambda}^* &\sim \mathrm{N}(\mathbf{0}, \tau\mathrm{Diag}(\boldsymbol{\lambda}^*)), \\ \lambda_i^* &\overset{iid}{\sim} \mathrm{Exp}(1), \\ \tau \mid \gamma, \delta &\sim \mathrm{Gamma}(\gamma, \delta). \end{aligned} \tag{1.34}$$

Scheipl and Kneib (2008) applied the NEG model on the parameters under the locally adaptive Bayesian P-splines to estimate the nonlinear dependence between the response and the predictor variables.

Recently, Armagan, Dunson and Lee (2013) introduced the generalized double Pareto (GDP) prior which is similar the NEG prior. The differences between the GDP and the NEG priors are that λ_i's in the GDP prior are assumed to follow an exponential distribution with parameter $\tau^2/2$ instead of τ in (1.32), and a gamma hyperprior is again placed on τ:

$$\begin{aligned}\boldsymbol{\theta} \mid \boldsymbol{\lambda} &\sim \mathrm{N}(\mathbf{0}, \mathrm{Diag}(\boldsymbol{\lambda})), \\ \lambda_i \mid \tau &\overset{iid}{\sim} \mathrm{Exp}\left(\frac{\tau^2}{2}\right), \\ \tau \mid \gamma, \delta &\sim \mathrm{Gamma}(\gamma, \delta). \end{aligned} \tag{1.35}$$

Although the differences are subtle, the GDP prior leads to an analytic form of the marginal density of θ_i's:

$$\theta_i \mid \xi = \frac{\delta}{\gamma}, \gamma \overset{iid}{\sim} \mathrm{GDP}(\xi, \gamma), \tag{1.36}$$

with distribution density

$$\frac{1}{2\xi}\left(1 + \frac{|\theta_i|}{\gamma\xi}\right)^{-(\gamma+1)}.$$

Moreover, model (1.35) degenerates to the double exponential prior as $\gamma \to \infty$ and $0 < 1/\xi < 1$, and to the Norma–Jeffreys prior as $\delta = \gamma = 0$.

Moreover, Carvalho, Polson and Scott (2010) developed another prior in the form of the scale mixture of normals class, the horseshoe (HS) prior, which can be represented by

$$\begin{aligned}\boldsymbol{\theta} \mid \boldsymbol{\lambda} &\sim \mathrm{N}(\mathbf{0}, \mathrm{Diag}(\boldsymbol{\lambda})), \\ \sqrt{\lambda_i} \mid \tau &\overset{iid}{\sim} \mathrm{Cauchy}^+(0, \tau), \\ \tau \mid \sigma &\sim \mathrm{Cauchy}^+(0, \sigma), \end{aligned} \tag{1.37}$$

where σ is the standard deviation of X_i. When $\sigma^2 = \tau^2 = 1$, the positive Cauchy prior on $\sqrt{\lambda_i}$ implies a Beta$\,(1/2, 1/2)$ distribution on $\kappa_i = 1/(1+\lambda_i)$, which goes to infinity when $\kappa_i \to 0$ or $\kappa_i \to 1$. Similar to the NEG prior, the hyperparamter τ is a global parameter, so we can write the model as

$$\begin{aligned} \boldsymbol{\theta} \mid \boldsymbol{\lambda}^*, \tau &\sim \mathrm{N}(\mathbf{0}, \tau^2 \mathrm{Diag}(\boldsymbol{\lambda}^*)), \\ \sqrt{\lambda_i^*} &\overset{iid}{\sim} \mathrm{Cauchy}^+(0,1), \\ \tau \mid \sigma &\sim \mathrm{Cauchy}^+(0,\sigma). \end{aligned} \tag{1.38}$$

Alternatively, using $\boldsymbol{\kappa}$ as the parameter, we can rewrite the model as

$$\begin{aligned} \boldsymbol{\theta} \mid \boldsymbol{\kappa}, \tau &\sim \mathrm{N}\left(\mathbf{0}, \tau^2 \mathrm{Diag}\left(\frac{1-\boldsymbol{\kappa}}{\boldsymbol{\kappa}}\right)\right), \\ \kappa_i &\overset{iid}{\sim} \mathrm{Beta}\left(\frac{1}{2}, \frac{1}{2}\right), \\ \tau \mid \sigma &\sim \mathrm{Cauchy}^+(0,\sigma). \end{aligned} \tag{1.39}$$

Furthermore, Polson and Scott (2009) generalized the horseshoe prior to a bigger class of scale mixture of normals priors using hypergeometric-beta as the mixture distribution, that is

$$\begin{aligned} \boldsymbol{\theta} \mid \boldsymbol{\kappa} &\sim \mathrm{N}\left(\mathbf{0}, \mathrm{diag}\left(\frac{1-\boldsymbol{\kappa}}{\boldsymbol{\kappa}}\right)\right), \\ \kappa_i &\overset{iid}{\sim} \mathrm{HB}(a,b,\tau,s) \\ &= C^{-1}\kappa_i^{a-1}(1-\kappa_i)^{b-1}\left\{\frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right)\kappa_i\right\}^{-1} \exp(-s\kappa_i), \end{aligned} \tag{1.40}$$

where C is the normalizing constant

$$C = \exp(-s)\mathrm{Beta}(a,b)\Phi_1(b,1,a+b,s,1-1/\tau^2), \tag{1.41}$$

where $\mathrm{Beta}(a,b)$ is the beta function and Φ_1 is the degenerate hypergeometric function of two variables. The author studied the roles of the four hyperparameters and found that a and b mainly control the shape of the distribution analogous to the two parameters in a beta distribution, and τ and s are two global scaling factors. However, the effects of the two scale parameters τ and s are not clearly separated. Similar shrinkage performances can be attained from different combinations of τ and s.

1.4 Penalized Least Squares Approaches

Another strand of shrinkage estimators that have had great successes are penalized least squares estimators. Consider the general multivariate normal likelihood with unknown mean $\boldsymbol{\theta}$ and unknown covariance matrix $\boldsymbol{\Sigma}$. A penalized least squares estimator of $\boldsymbol{\theta}$ can be expressed as the minimizer of $(\boldsymbol{X}-\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\theta})$ subject to a constrained function $P(\boldsymbol{\theta}) \leq t$, that is, the minimizer of

$$L(\boldsymbol{\theta}) = (\boldsymbol{X}-\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\theta}) + \lambda P(\boldsymbol{\theta}), \tag{1.42}$$

where λ is a positive tuning parameter and can be selected via cross validation, and $P(\boldsymbol{\theta})$ is the penalty function. Differentq shrinkage estimators have been derived using different penalty functions. The followings are a few examples.

Under the L_0 norm penalty, the penalty function $P(\boldsymbol{\theta}) = \sum_j I(\theta_j \neq 0)$ is the number of non-zero components. Assume the normal likelihood (1.3) with $\sigma^2 = 1$, the loss function (1.42) can be rewritten as:

$$L(\boldsymbol{\theta}) = \|\boldsymbol{X}-\boldsymbol{\theta}\|_2^2 + \lambda \sum_j I(\theta_j \neq 0).$$

For any fixed $\lambda > 0$, the solution to this minimization problem is

$$\hat{\theta}_i^{L_0} = \begin{cases} X_i, & \text{when } \; X_i \geq \sqrt{\lambda} \\ 0, & \text{when } \; |X_i| < \sqrt{\lambda} \\ X_i, & \text{when } \; X_i \leq -\sqrt{\lambda} \end{cases}. \tag{1.43}$$

It shrinks X_i to zero if X_i falls in the range $(-\sqrt{\lambda}, \sqrt{\lambda})$ and leaves it as it is if X_i is outside of the range. The estimator (1.43) has a simple form and is easy to be calculated. However, it is discontinuous and may be very sensitive to small changes in the data. As shown in Figure 1.2, when X_i is close to the cutoff points, which are $\pm\sqrt{2}$ here, a small change in X_i may change the estimate of θ_i from 0 to $\pm\sqrt{2}$ or more extreme.

The L_1 norm penalty function uses the penalty function $P(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_1$. The corresponding estimator is known as the Lasso (Tibshirani 1996), which can be represented by

$$\hat{\theta}_i^{L_1} = \text{sign}(X_i)\left(|X_i| - \frac{1}{2}\lambda\right)^+ . \tag{1.44}$$

Compared to the L_0 norm penalty, the L_1 norm penalty estimator or the Lasso estimator is continuous with a constant shrinkage amount $\lambda/2$ beyond the range $(-\lambda/2,\ \lambda/2)$. As discussed in Section 1.3.3, the Lasso estimator has a Bayesian interpretation as the posterior mode under the double exponential prior.

In addition, the L_2 norm penalty function uses the function $P(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|_2^2$. The corresponding estimator is known as the ridge estimator, which can be represented by:

$$\hat{\boldsymbol{\theta}}^{L_2} = \hat{\boldsymbol{\theta}}_{Ridge} = \frac{\boldsymbol{X}}{1+\lambda}. \tag{1.45}$$

Similar to the Lasso estimator, the ridge estimator also has a Bayesian interpretation. Given a normal likelihood with mean θ, the ridge estimator can be obtained as the posterior mean (median and mode) of θ under a normal prior.

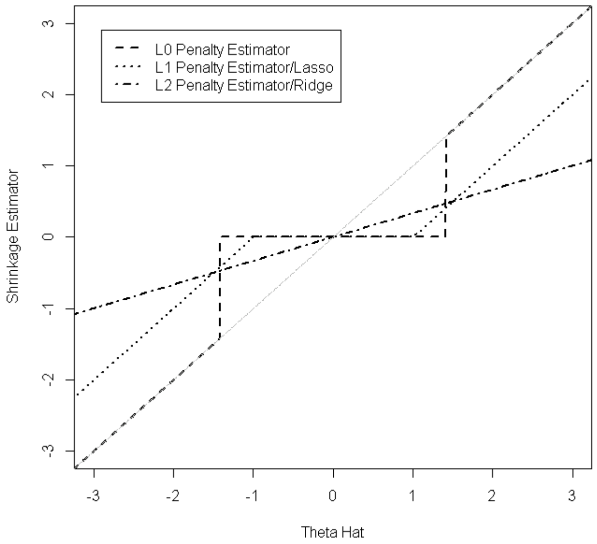

Figure 1.2: Shrinkage estimator comparison when $p = 1$ and $\lambda = 2$. The horizontal axis is the MLE.

Figure 1.2 plots the MLE versus the penalized least squares estimators under the L_0, L_1 and L_2 penalty functions when $p = 1$ and $\lambda = 2$ for all three methods. We can see that the ridge estimator does not shrink the MLE to exactly zero but instead shrinks by a constant proportion to the origin. However the other two methods both shrink the MLE to exactly zero in some ranges. People sometimes prefer the Lasso to the ridge estimator due to this issue for the purpose of variable selection and model simplicity.

The generalized double Pareto prior in Armagan et al. (2013) induces another penalty function. Suppose that

$$\boldsymbol{Y} \mid \boldsymbol{\beta}, \sigma \sim \mathrm{N}(\boldsymbol{X\beta}, \sigma^2 I), \tag{1.46}$$

then the penalized least square estimator can be written as:

$$\tilde{\boldsymbol{\beta}}^{GDP} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2\sigma^2} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{i=1}^{p} p(|\beta_i|) \right\}, \tag{1.47}$$

with the penalty function

$$p(|\beta_i|) = (\gamma + 1)\log(\sigma\delta + |\beta_i|).$$

Armagan et al. (2013) showed that this penalty function can lead to estimators with the good properties from Fan and Li (2001): (i) nearly unbiased when the signal is large, (ii) able to set small estimated coefficients to zero and (iii) continuous in data to avoid instability.

Another penalized least squares estimator, the elastic net estimator, is raised by Zou and Hastie (2005). Their motivation comes from the following limitations of the Lasso in variable selection problems. First, when the parameter dimension p is larger than the number of observations n, the Lasso can give at most n estimated components. Second, when a group of variables are highly correlated, the Lasso may select only one variable from the group but not care which one is selected. To improve the Lasso in such situations, Zou and Hastie (2005) provided two estimators: the naive elastic net estimator and the elastic net estimator. The first one is defined as the minimizer of

$$L(\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2, \tag{1.48}$$

where $\boldsymbol{Y}$ is the response vector, $\boldsymbol{X}$ is the design matrix and λ_1 and λ_2 are two tuning parameters. Equation (1.48) points out that the naive elastic net can be treated as a combination of the Lasso and the ridge estimators. To find the solution when $p > n$, one augmented dataset need to be defined first:

$$\boldsymbol{X}^*_{(n+p)\times p} = \frac{1}{\sqrt{1+\lambda_2}}\begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda_2} I \end{pmatrix}, \quad \boldsymbol{Y}^*_{n+p} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{0} \end{pmatrix}.$$

Furthermore, define $\gamma = \lambda_1/\sqrt{1+\lambda_2}$ and $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$. Then the target function (1.48) can be expressed as

$$L(\boldsymbol{\beta}) = L(\boldsymbol{\beta}^*) = \|\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*\|_2^2 + \gamma\|\boldsymbol{\beta}^*\|_1. \tag{1.49}$$

When the design matrix is orthogonal, the solution is

$$\hat{\beta}_i^{NEN} = \frac{\text{sign}(\hat{\beta}_i^O)}{1+\lambda_2}\left(|\hat{\beta}_i^O| - \frac{\lambda_1}{2}\right)^+, \tag{1.50}$$

where $\hat{\boldsymbol{\beta}}^O = \boldsymbol{X}^T\boldsymbol{Y}$. Now the problem is transformed to finding a Lasso type solution with the values of λ_1 and λ_2 selected via cross-validation. The authors showed that the naive elastic net estimator can solve the issues of the Lasso in the above situations. However, this estimator tends to give over shrinkage to large observations. To cure this problem and make it comparable to the Lasso and the ridge estimators, the elastic net estimator is constructed:

$$\hat{\beta}^{EN} = \sqrt{1+\lambda_2}\hat{\beta}^{NEN}. \tag{1.51}$$

It is easy to see that when the design matrix is orthogonal, $\hat{\beta}^{EN}$ has a similar form as the lasso.

Most above estimators own the global shrinkage property, which is that all nonzero estimates have the same shrinkage amount or proportion. For example, the Lasso has a constant shrinkage amount $\lambda/2$ beyond the range $(-\lambda/2,\ \lambda/2)$. However, this shrinkage amount may be too much when the observation $\boldsymbol{X}_i$ is large. The lack of the adaptive shrinkage property and moderate shrinkage strength motivated the research on frequentist shrinkage estimators with local shrinkage property. There is a large literature in this area, among which one important approach is the adaptive Lasso (Zou 2006). The main idea of this method is to assign different weights to different components of $\boldsymbol{\theta}$ in the L_1 penalized function. Suppose that an initial estimator $\tilde{\boldsymbol{\theta}}$ is available. For any given γ, define

$$w_i = \frac{1}{\tilde{\theta}_i^{\gamma}}. \tag{1.52}$$

Then the adaptive lasso estimator is the minimizer of

$$L(\boldsymbol{\theta}) = \|\boldsymbol{X} - \boldsymbol{\theta}\|_2^2 + \lambda \sum_i w_i |\theta_i|. \tag{1.53}$$

Zou (2006) suggested using two-dimensional cross-validation to select the values of γ and λ.

1.5 Comparison Between Penalized Likelihood Approaches And Bayesian Approaches

In the previous sections we see many examples of connections between classical methods and Bayesian works. Besides those, there is a noticeable connection between the penalty functions and the Bayesian priors. Given a prior distribution, a penalized least squares estimator can be derived with the penalty function being the negative logarithm of the corresponding prior. In other words, the penalized least squares estimator can be viewed as the Maximum A Posteriori (MAP) estimate given that prior density. For example, the Lasso can be expressed as the posterior mode given a double exponential prior and the ridge estimator is equivalent to the posterior mode (which equals to the posterior mean) given a normal prior.

Compared to the penalized likelihood methods, Bayesian methods have a few important advantages. First of all, Bayesian methods allow us to apply prior knowledge about the parameters and to use observed data to update this prior information. A good prior should yield an estimator with superior performances.

Secondly, as introduced at the beginning of section 1.3, proper Bayesian rules are admissible. This can be easily verified. Let θ be the parameter. A decision rule δ is called a Bayes rule if it minimizes the Bayes risk $r(\pi, \delta) = E^{\pi} R(\theta, \delta)$. A Bayes rule δ is admissible. If it is not, then there exists another rule δ' with smaller risk $R(\theta, \delta')$ for all values of θ. Then automatically δ' will have a smaller Bayes risk which contradicts with the definition of Bayes rules. It is noticeable that under the squared error loss, the posterior mean is the Bayes rule.

In this dissertation we use squared error loss as the loss function and the posterior mean as the estimator of $\boldsymbol{\theta}$. When the conditions $(B1) - (B5)$ in Lehmann and Casella (1998) hold, we further have

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{L} \mathrm{N}\left(0, \frac{1}{I(\theta_0)}\right), \tag{1.54}$$

where $\tilde{\theta}_n$ is the posterior mean given sample size n and θ_0 is the true value of θ. That is, the posterior mean is consistent and asymptotically efficient.

1.6 Outline of the Dissertation

The major thrust of this dissertation is to develop a class of new prior distributions: the adaptive inverted-Beta priors (AIB). In addition to the above Bayesian advantages, our method has several other nice properties. First of all, the class of AIB priors have the adaptive shrinkage property and the flexibility to handle different types of data with a large variety of sparsity levels and signal sizes. Secondly, the class of AIB priors include several well known Bayesian shrinkage priors as special cases. Our method behaves similarly as these members in some situations and provides substantial improvements in other situations. Thirdly, focusing on scale mixture of normals priors of the form (1.20), we investigate the substantial impact from the global shrinkage parameter to the shrinkage performance.

The remaining chapters of this thesis are organized as the following. Chapter 2 proposes the adaptive inverted-Beta priors, investigates their theoretical properties, and demonstrates their performances for IID observations through simulation studies. Chapter 3 extends the AIB priors to linear regressions and demonstrates the superior performances using simulation studies and one real data example. Chapter 4 summarizes our main findings and extends the AIB priors for the normal mean problem with unknown covariance structures.

Chapter 2: Adaptive Inverted-Beta Priors For IID Observations

The shape of the marginal distributions on $\boldsymbol{\theta}$ has a big impact on the shrinkage properties of the Bayesian estimators. In this chapter, we first study a few important shrinkage priors in the literature. Next we propose a new class of adaptive priors, the adaptive inverted-Beta (AIB) priors, which allows the data to decide the shrinkage degree and provides strong shrinkage to noises while little shrinkage to signals. We conduct simulation studies where the observations are IID. The performances of the Bayesian estimators under the AIB priors are compared with those under the horseshoe prior, the Strawderman-Berger prior, the normal-exponential-gamma prior and the double-exponential prior. A few important features are worth noting: 1) the proposed priors have the adaptive shrinkage property and the flexibility to handle different sparsity levels and signal sizes; 2) the global shrinkage parameter introduces extra variability but improves the shrinkage performances.

2.1 Existing Local Shrinkage Priors

In high dimensional spaces, the following properties are desirable for shrinkage priors. 1) Signal detectability: the prior distribution should be able to distinguish the noises from the signals and give strong shrinkage for small noises and little shrinkage for signals. 2) Adaptivity: the prior density should be flexible enough to handle different data sparsity levels. That is, the corresponding Bayesian estimator should perform well under a large range of sparsity levels. 3) Invariance: the prior distribution should be invariant to the measurement units. Next we investigate the connection between the shapes of shrinkage priors and shrinkage properties.

Recall that under the likelihood function (1.3) with $\sigma^2 = 1$, and a scale mixture of normals prior (1.20) on $\boldsymbol{\theta}$, the posterior mean of θ_i is:

$$\mathrm{E}(\theta_i \mid X_i) = \left(1 - \mathrm{E}\left(\frac{1}{1+\lambda_i} \mid X_i\right)\right) X_i = (1 - \mathrm{E}(\kappa_i \mid X_i))X_i, \tag{2.1}$$

where $\kappa_i = 1/(1+\lambda_i)$ is the shrinkage factor. Strong shrinkage is given when κ_i approaches 1, or equivalently when λ_i approaches 0, and little shrinkage is given when κ_i approaches zero or λ_i approaches infinity. Therefore, the density of $\pi(\kappa_i)$ near 0 controls the tail robustness of the prior, and the density near 1 controls the shrinkage strength to noises. Integrating out $\boldsymbol{\kappa}$ yields the marginal prior of $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$. The density of $\pi(\boldsymbol{\theta})$ around 0 controls the shrinkage power to noises and the tail thickness controls the shrinkage power to signals. To provide additional insight of how $\pi(\kappa_i)$ and $\pi(\theta_i)$ affect the shrinkage performance, we choose a few shrinkage priors mentioned in Chapter 1 as examples to construct a direct comparison. Since some priors do not have an analytic form of $\pi(\theta)$, only kernel densities of λ and κ in the one dimensional case are summarized in Table 2.1. The densities of $\pi(\kappa)$ and $\pi(\theta)$ are given in Figure 2.1.

The comparison is among the horseshoe prior (1.38) with τ fixed at 1, the Strawderman–Berger prior (1.22) with $a = 1/2$, the NEG prior (1.32) with $\gamma = 2$ and $\delta = 1$, the NG prior (1.31) with $\gamma = 2$ and $\delta = 1$ and the standard double exponential prior. We

Prior for θ	Prior for λ	Prior for κ
Horseshoe	$\lambda^{-\frac{1}{2}}(1+\lambda)^{-1}$	$\kappa^{-\frac{1}{2}}(1-\kappa)^{-\frac{1}{2}}$
Strawderman–Berger	$(1+\lambda)^{-\frac{3}{2}}$	$\kappa^{-\frac{1}{2}}$
NEG(2,1)	$(1+\lambda)^{-3}$	κ
NG(2,1)	$\lambda e^{-\lambda}$	$\frac{1-\kappa}{\kappa^3}e^{-(1-\kappa)/\kappa}$
Double Exponential	$e^{-\lambda/2}$	$\frac{1}{\kappa^2}e^{-(1-\kappa)/2\kappa}$

Table 2.1: Priors on λ and κ of some common shrinkage priors in one dimensional case. The two parameters in both the normal-exponential-gamma and the normal-gamma priors are fixed at 2 and 1. The priors are given up to constants.

use model (1.32) rather than (1.34) to represent the NEG prior because that if the global shrinkage parameter τ in (1.34) is fixed at 1, then the NEG prior degenerates to the double exponential prior. We choose these priors as they are typical examples that work well when the data have some certain sparsity properties.

The left graph in Figure 2.1 compares the densities of κ, the middle one compares the peak behaviors of $\pi(\theta)$ and the right one compares the tail behaviors of $\pi(\theta)$. Explicitly, when the closed form of $\pi(\theta)$ is not available, we approximate this density by taking the average of the normal priors on θ over a set of κ values, which are sampled from their corresponding distribution respectively. The horseshoe prior has unbounded $\pi(\kappa)$ near 1, reflecting the infinite peak of $\pi(\theta)$ and strong shrinkage to small noises. In contrast, all other priors have $\pi(\kappa)$ bounded at 1, reflecting the lower peaks of $\pi(\theta)$. Although $\pi(\theta)$ of the NEG$(2,1)$ prior has a finite peak, the high densities around zero, for example in the range $(-1,\ 1)$, also offers the advantage in handling small noises. The unbounded $\pi(\kappa)$ of the Strawderman–Berger prior and the horseshoe prior near 0 correspond to the thick tails of $\pi(\theta)$ and little shrinkage to

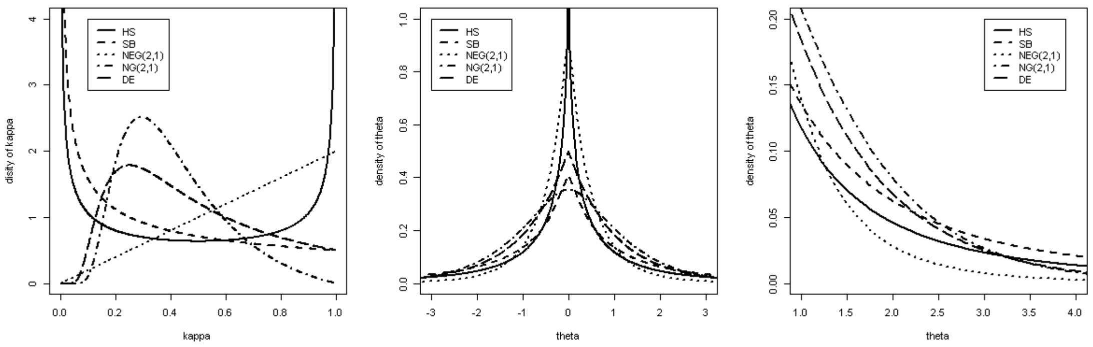

Figure 2.1: Density comparison of $\pi(\kappa)$ and $\pi(\theta)$ in the one dimensional case among the horseshoe prior, the Strawderman–Berger prior, the NEG(2,1) prior, the NG(2,1) prior and the standard double exponential prior. The left graph compares the densities of κ, the middle graph compares the densities of θ and the right one compares the tail behaviors of $\pi(\theta)$.

large signals. However, the other three priors have $\pi(\kappa)$ vanish at 0 leading to thin tails of $\pi(\theta)$ and over shrinkage issue to large signals.

In fact, the NEG priors can have very different shapes by varying the two hyperparameter values. The shape parameter γ controls the tail thickness and the rate parameter δ controls the scale. When the ratio between γ and δ is small, the peak is low and tails are thick; when the ratio is large, the peak is high and tails are thin. Similar to Figure 2.1, Figure 2.2 shows the shapes and behaviors of $\pi(\kappa)$ and $\pi(\theta)$ of a few NEG priors. We use the NEG$(2, 1)$ prior as a base level and compare it with the NEG(1,1), NEG(5,1) and NEG(0.1,1) priors.

Similarly, the NG priors can have very different shapes, even an infinite peak of $\pi(\theta)$ by changing the two hyperparameters. More details are given in Griffin and

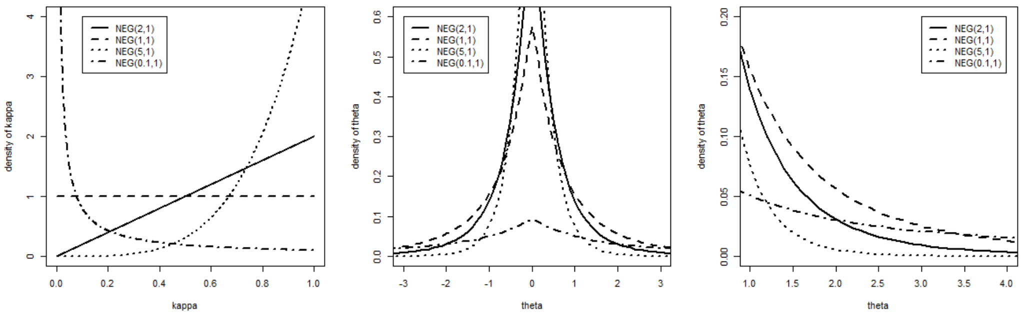

Figure 2.2: Density comparison of $\pi(\kappa)$ and $\pi(\theta)$ in the one dimensional case among NEG priors with different hyperparameter values. The left graph compares the densities of κ, the middle graph compares the densities of θ and the right one compares the tail behaviors of $\pi(\theta)$.

Brown (2005) and Griffin and Brown (2010). Instead of unfixing the two hyperparameters, the priors can gain shape flexibility by introducing another global shrinkage parameter. The over shrinkage issue of the NEG$(2, 1)$ prior is cured by doing so. More details are given in Section 2.4.

As we seen, these priors have pre-specified shrinkage properties, so work well only when the data has certain sparsity levels. In the next section, we generalize them and provide a new class of adaptive priors.

2.2 The Class of Adaptive Inverted-Beta Priors

Motivated by the series of paper mentioned, we find that satisfying the following conditions would yield estimators with the good properties described at the beginning of section 2.1.

First, put a scale mixture of normals prior on $\boldsymbol{\theta}$. The choice of hyperprior on the local variances of $\boldsymbol{\theta}$ can adjust the peak height and the tail thickness of $\pi(\boldsymbol{\theta})$

and further yields adaptive estimators. In addition, Stein (1981) stated that under likelihood (1.3) with $\sigma^2 = 1$ and the squared error loss, the posterior mean of $\boldsymbol{\theta}$ can be expressed as

$$E(\boldsymbol{\theta} \mid \boldsymbol{X}) = \boldsymbol{X} + \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{X}} \log m(\boldsymbol{X}), \tag{2.2}$$

where $m(\boldsymbol{X})$ is the marginal likelihood of $\boldsymbol{X}$. Equation 2.2 indicates that a prior distribution with shrinkage adaptivity should satisfy $\lim_{X_i \to \infty} \frac{\mathrm{d}}{\mathrm{d}\mathrm{X}_i} \log m(\boldsymbol{X}) = 0$.

Second, to make the prior distribution flexible, more than one hyperparameter or an extra global shrinkage parameter is needed.

Third, as in model (1.38), putting σ, the standard deviation of X_i given θ_i, inside the prior of $\boldsymbol{\theta}$ as a scale parameter can easily link the prior density with the data variance, and therefore, satisfy the invariance property.

Following these conditions, we develop a large new class of prior distributions that have flexible shapes and can give adaptive shrinkages to observations.

2.2.1 Introduction of The Adaptive Inverted Beta Prior

As shown in Table 2.1, under the horseshoe, the Strawderman–Berger and the NEG(2, 1) priors, the distributions of λ_i have the form $\lambda_i^s(1 + \lambda_i)^t$ for some constants s and t. They can be viewed as special cases of the inverted-Beta (IB) distribution, which is also called the beta prime distribution, with the density function

$$\lambda \mid a, b \sim \mathrm{IB}(a, b) = \frac{1}{\mathrm{Beta}(a, b)} \frac{\lambda^{b-1}}{(1 + \lambda)^{a+b}}, \tag{2.3}$$

where Beta(a, b) is the beta function. In fact, many other priors can be viewed as special cases of IB priors. For example, the IB prior degenerates to the Jeffreys prior when both a and b approach 0. When $\delta = 1$, the prior (1.33) simplifies to

$$p(\lambda \mid \gamma, \delta = 1) = \gamma(1 + \lambda)^{-(\gamma+1)}, \tag{2.4}$$

which is $\mathrm{IB}(\gamma, 1)$.

It is easy to check that $\kappa = 1/(1+\lambda)$ follows a beta distribution

$$\kappa \sim \mathrm{Be}(a, b) = \frac{\kappa^{a-1}(1-\kappa)^{b-1}}{\mathrm{Beta}(a, b)}. \tag{2.5}$$

The beta distribution behaves like κ^{a-1} near the origin and like $(1-\kappa)^{b-1}$ near $\kappa = 1$. Therefore the values of the constants a and b decide the shape of the distribution of κ, and thus, the shrinkage properties. For example, $a = b = 1/2$ results in the U-shape of $\pi(\kappa)$ in the horseshoe prior and furthermore the adaptive shrinkage property. For another example, when $a > 1$, $\pi(\kappa)$ vanishes at the origin which causes the over shrinkage issue of the $\mathrm{NEG}(2, 1)$ prior.

This observation motivated us to let data decide the values of the hyperparameters, a and b, and thus decide the shrinkage degrees. Instead of putting diffuse priors on a and b directly, we consider putting priors on the transformed parameters $M = a + b$ and $N = a/(a+b)$. The main reason of doing so is that, N is the mean of the shrinkage factors $\kappa_i, i = 1, \cdots, p$. Putting a diffuse prior on N will allow the data to determine its posterior distribution which reflects the overall shrinkage strength from local shrinkage parameters. For example, the posterior mean of N is expected to be greater than 0.5 for sparse data. When the signal size is fixed, a larger posterior mean of N is expected, indicating strong shrinkages, when the sparsity level increases. For a fixed sparsity level, a smaller posterior mean of N is expected when the signal size increases. In order to trace this parameter directly, we put priors $\mathrm{Gamma}(2, 1)$ and

Beta(1, 1) on M and N respectively, and propose the AIB prior

$$\begin{aligned} \boldsymbol{\theta} \mid \boldsymbol{\lambda}, \tau, \sigma &\sim \mathrm{N}(\mathbf{0}, \tau^2\sigma^2\mathrm{Diag}(\boldsymbol{\lambda})), \\ \tau &\sim \mathrm{Cauchy}^+(0,1), \\ p(\sigma^2) &\propto \frac{1}{\sigma^2}, \\ \lambda_i \mid M, N &\overset{iid}{\sim} \mathrm{IB}(MN, M(1-N)), \\ M &\sim \mathrm{Gamma}(2,1), \\ N &\sim \mathrm{Beta}(1,1). \end{aligned} \tag{2.6}$$

Here we follow the suggestion from Polson and Scott (2010) and put the half Cauchy distribution on the global shrinkage parameter τ. In Section 2.4, we will see that unfixing N and excluding τ give the AIB prior flexibility to handle different types of data. Unfixing both parameters involves more variability, but improves the flexibility and the shrinkage performances substantially in some scenarios.

Furthermore, when $\tau = 1$, the AIB prior (2.6) is equivalent to the hypergeometric inverted beta prior (1.25) with $s = 0$ and $\delta = 1$. But differently, the AIB prior has only one global shrinkage parameter τ and thus the impact from the global shrinkage parameter to shrinkage performances is much clearer.

2.2.2 Properties of AIB prior

In this section, we focus on the theoretical properties of the AIB priors. Essentially, Theorem 1 shows the tail robustness of the AIB priors in one dimensional case and Corollary 1 shows that the posterior mean under an AIB prior has a bounded risk.

Theorem 1. *Suppose $X \sim N(\theta, 1)$ and the AIB prior (2.6) on θ. Let $m(X)$ denote the marginal likelihood of X and let $E(\theta \mid X = x)$ denote the posterior mean of θ*

given one observation x. Then

$$| X - E(\theta \mid X) | \leq B, \tag{2.7}$$

where B is a constant and

$$\lim_{|X| \to +\infty} \frac{\mathrm{d}}{\mathrm{d}X} \log m(X) = 0. \tag{2.8}$$

Proof. It is easy to see that

$$\lim_{|X| \to +\infty} \frac{\mathrm{d}}{\mathrm{d}X} \log m(X) = \lim_{|X| \to +\infty} \frac{1}{m(X)} \frac{\mathrm{d}}{\mathrm{d}X} m(X).$$

In model (2.6), when $a = MN$, $b = M(1 - N)$ and τ are fixed, the joint distribution of X and λ is

$$\begin{aligned} p(X, \lambda \mid a, b, \tau) &= \int_{-\infty}^{+\infty} p(X \mid \theta) \pi(\theta \mid \lambda, \tau) \pi(\lambda \mid a, b) \mathrm{d}\theta \\ &= \frac{1}{\sqrt{2\pi(1 + \lambda\tau^2)}} \frac{1}{\mathrm{B}(a, b)} \frac{\lambda^{b-1}}{(1 + \lambda)^{a+b}} \exp\left(-\frac{X^2}{2(1 + \lambda\tau^2)}\right), \end{aligned}$$

where $\mathrm{B}(a, b)$ is the beta function. Let $z = 1/(1 + \lambda\tau^2)$, then $\lambda = (1 - z)/(z\tau^2)$, and $\mathrm{d}\lambda = \mathrm{d}z/(\tau^2 z^2)$. Therefore, the distribution of X conditional on a, b and τ is

$$\begin{aligned} m(X \mid a, b, \tau) &= \int_0^{+\infty} p(X, \lambda \mid a, b, \tau) \mathrm{d}\lambda \\ &= \int_0^{+\infty} \frac{1}{\sqrt{2\pi(1 + \lambda\tau^2)}} \frac{1}{\mathrm{B}(a, b)} \frac{\lambda^{b-1}}{(1 + \lambda)^{a+b}} \exp\left(-\frac{X^2}{2(1 + \lambda\tau^2)}\right) \mathrm{d}\lambda \\ &= \int_0^1 \frac{\sqrt{z}}{\sqrt{2\pi}} \frac{1}{\mathrm{B}(a, b)} \left(\frac{z\tau^2}{1 - z + z\tau^2}\right)^{a+b} \left(\frac{1 - z}{z\tau^2}\right)^{b-1} e^{-\frac{X^2 z}{2}} \frac{1}{\tau^2 z^2} \mathrm{d}z \\ &= \int_0^1 C(a, b, \tau) z^{a - \frac{1}{2}} (1 - z)^{b-1} e^{-\frac{X^2 z}{2}} \left(\frac{1}{\tau^2} + \left(1 - \frac{1}{\tau^2}\right) z\right)^{-(a+b)} \mathrm{d}z, \end{aligned}$$

where $C(a, b, \tau) = \frac{1}{\sqrt{2\pi}} \frac{1}{\mathrm{B}(a,b)} \frac{1}{\tau^{2b}}$.

Applying the Taylor expansion of the exponential function $e^x = \sum_{m=0}^{\infty} \frac{x^m}{m!}$ and letting $\eta = 1 - z$ yields

$$\begin{aligned}
& m(X \mid a, b, \tau) \\
= \; & \int_0^1 C(a,b,\tau)(1-\eta)^{a-\frac{1}{2}}\eta^{b-1}e^{-\frac{X^2}{2}(1-\eta)}\left(\frac{1}{\tau^2}+\left(1-\frac{1}{\tau^2}\right)(1-\eta)\right)^{-(a+b)}\mathrm{d}\eta \\
= \; & C(a,b,\tau)e^{-\frac{X^2}{2}}\int_0^1\left(1-\left(1-\frac{1}{\tau^2}\right)\eta\right)^{-(a+b)}e^{\frac{X^2\eta}{2}}(1-\eta)^{a-\frac{1}{2}}\eta^{b-1}\mathrm{d}\eta \\
= \; & C(a,b,\tau)e^{-\frac{X^2}{2}}\sum_{m=0}^{\infty}\frac{(X^2/2)^m}{m!}\int_0^1\eta^{b+m-1}(1-\eta)^{a-\frac{1}{2}}\left(1-\left(1-\frac{1}{\tau^2}\right)\eta\right)^{-(a+b)}\mathrm{d}\eta.
\end{aligned}$$

Implementing formulas (15.1.1) and (15.3.1) in Abramowitz and Stegun (1964), we obtain

$$\begin{aligned}
& m(X \mid a, b, \tau) \\
= \; & C(a,b,\tau)e^{-\frac{X^2}{2}}\sum_{m=0}^{\infty}\frac{(X^2/2)^m}{m!}\mathrm{F}\left(a+b, b+m, a+b+m+\frac{1}{2}, 1-\frac{1}{\tau^2}\right) \\
& \mathrm{B}\left(b+m, a+\frac{1}{2}\right) \\
= \; & C(a,b,\tau)e^{-\frac{X^2}{2}}\sum_{m=0}^{\infty}\frac{(X^2/2)^m}{m!}\sum_{n=0}^{\infty}\frac{(a+b)_n(b+m)_n}{\left(a+b+m+\frac{1}{2}\right)_n}\frac{\left(1-1/\tau^2\right)^n}{n!} \\
& \mathrm{B}\left(b+m, a+\frac{1}{2}\right) \\
= \; & C(a,b,\tau)e^{-\frac{X^2}{2}}\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\frac{(a+b)_n(b)_{m+n}}{\left(a+b+\frac{1}{2}\right)_{m+n}}\mathrm{B}\left(b, a+\frac{1}{2}\right)\frac{(X^2/2)^m\left(1-1/\tau^2\right)^n}{m!n!} \\
= \; & C(a,b,\tau)e^{-\frac{X^2}{2}}\Phi_1\left(b, a+b, a+b+\frac{1}{2}, \frac{X^2}{2}, 1-\frac{1}{\tau^2}\right), \qquad (2.9)
\end{aligned}$$

where $\mathrm{F}(a,b,c,d)$ denotes the hypergeometric function, $(q)_n$ is the rising factorial, defined by

$$(q)_n = \begin{cases} 1, & n = 0 \\ q(q+1)\cdots(q+n-1), & n > 0 \end{cases}, \qquad (2.10)$$

and Φ_1 is the degenerate hypergeometric function of two variables.

Gordy (1998) showed that

$$\Phi_1(a,b,c,x,y) = \begin{cases} e^x \sum_{n=0}^{\infty} \frac{(a)_n(b)_n}{(c)_n} \frac{y^n}{n!} \mathrm{F}_1(c-a, c+n, -x), & \text{for} \quad 0 \leq y < 1, \;\; 0 < a < c \\ e^x(1-y)^{-b}\Phi_1(c-a, b, c, -x, \frac{y}{y-1}), & \text{for} \quad y < 0, \;\; 0 < a < c \end{cases} \tag{2.11}$$

where F_1 is the Kummer's function of the first kind. In addition, it is shown in Chapter 4 of Slater (1960) that

$$\mathrm{F}_1(a,b,x) = \begin{cases} \frac{\Gamma(a)}{\Gamma(b)} e^x x^{a-b}\{1+O(x^{-1})\}, & x > 0 \\ \frac{\Gamma(a)}{\Gamma(b-a)}(-x)^{-a}\{1+O(x^{-1})\}, & x < 0 \end{cases}. \tag{2.12}$$

When the parameters a, b and τ are random, the marginal distribution of X can be represented by

$$\begin{aligned} & m(X) \\ = \; & \int_0^1 \int_0^1 \int_0^\infty m(X \mid a,b,\tau)\pi(M)\pi(N)\pi(\tau)\mathrm{d}M\mathrm{d}N\mathrm{d}\tau \\ & + \int_1^\infty \int_0^1 \int_0^\infty m(X \mid a,b,\tau)\pi(M)\pi(N)\pi(\tau)\mathrm{d}M\mathrm{d}N\mathrm{d}\tau \\ = \; & m_{\tau<1}(X) + m_{\tau>1}(X) \\ = \; & \int_0^1 \int_0^1 \int_0^\infty C(M,N,\tau)\tau^{2M} \sum_{n=0}^{\infty} \frac{(MN+1/2)_n(M)_n}{(M+1/2)_n} \frac{(1-\tau^2)^n}{n!} \frac{\Gamma(M-MN)}{\Gamma(M+1/2+n)} \\ & \left(\frac{X^2}{2}\right)^{-(MN+1/2+n)} \left\{1+O\left(\frac{1}{X^2}\right)\right\} \pi(M)\pi(N)\pi(\tau)\mathrm{d}M\mathrm{d}N\mathrm{d}\tau \\ & + \int_1^\infty \int_0^1 \int_0^\infty C(M,N,\tau) \sum_{n=0}^{\infty} \frac{(M-MN)_n(M)_n}{(M+1/2)_n} \frac{(1-1/\tau^2)^n}{n!} \frac{\Gamma(MN+1/2)}{\Gamma(M-MN+n)} \\ & \left(\frac{X^2}{2}\right)^{-(MN+1/2)} \left\{1+O\left(\frac{1}{X^2}\right)\right\} \pi(M)\pi(N)\pi(\tau)\mathrm{d}M\mathrm{d}N\mathrm{d}\tau. \end{aligned} \tag{2.13}$$

Recall that a and b are functions of M and N, we change the notations to M and N in the last equation.

Taking the derivative with respect to X, we obtain

$$\begin{aligned}
&\frac{\mathrm{d}m_{\tau<1}(X)}{\mathrm{d}X} \\
= \quad & \int_0^1 \int_0^1 \int_0^\infty \frac{\mathrm{d}}{\mathrm{d}X} m(X \mid a, b, \tau)\pi(M)\pi(N)\pi(\tau)\mathrm{d}M\mathrm{d}N\mathrm{d}\tau \\
= \quad & \int_0^1 \int_0^1 \int_0^\infty -C(a,b,\tau)Xe^{-\frac{X^2}{2}}\Phi_1\left(b, a+b, a+b+\frac{3}{2}, \frac{X^2}{2}, 1-\frac{1}{\tau^2}\right) \\
& \pi(M)\pi(N)\pi(\tau)\mathrm{d}M\mathrm{d}N\mathrm{d}\tau \\
= \quad & -\frac{2}{X}\int_0^1 \int_0^\infty \int_0^1 C(M,N,\tau)\tau^{2M}\sum_{n=0}^\infty \frac{(MN+3/2)_n(M)_n}{(M+3/2)_n}\frac{(1-\tau^2)^n}{n!} \\
& \frac{\Gamma(M-MN)}{\Gamma(M+3/2+n)}\left(\frac{X^2}{2}\right)^{-(MN+1/2+n)}\left\{1+O\left(\frac{1}{X^2}\right)\right\} \\
& \pi(M)\pi(N)\pi(\tau)\mathrm{d}N\mathrm{d}M\mathrm{d}\tau \\
= \quad & -\frac{2}{X}\int_0^1 \int_0^1 \int_0^\infty C(M,N,\tau)\tau^{2M}\sum_{n=0}^\infty \frac{\left(MN+\frac{1}{2}\right)_n(M)_n}{\left(M+\frac{1}{2}\right)_n}\frac{(1-\tau^2)^n}{n!} \\
& \frac{\Gamma(M-MN)}{\Gamma\left(M+\frac{1}{2}+n\right)}\left(\frac{X^2}{2}\right)^{-(MN+1/2+n)}\left\{1+O\left(\frac{1}{X^2}\right)\right\} \\
& \frac{MN+\frac{1}{2}+n}{\left(M+\frac{1}{2}+n\right)^2}\pi(M)\pi(N)\pi(\tau)\mathrm{d}M\mathrm{d}N\mathrm{d}\tau. \qquad (2.14)
\end{aligned}$$

Since $0 \le N \le 1$,

$$0 \le \frac{MN+1/2+n}{(M+1/2+n)^2} \le \frac{M+1/2+n}{(M+1/2+n)^2} \le \frac{1}{1/2} = 2.$$

Therefore

$$0 \le -\frac{X}{2}\frac{\mathrm{d}m_{\tau<1}(X)}{\mathrm{d}X} \le 2m_{\tau<1}(X).$$

Similarly,

$$\begin{aligned}
& \frac{\mathrm{d}m_{\tau>1}(X)}{\mathrm{d}X} \\
= \;& -\frac{2}{X}\int_1^\infty\int_0^1\int_0^\infty C(M,N,\tau)\sum_{n=0}^\infty \frac{(M-MN)_n(M)_n}{(M+3/2)_n}\frac{(1-1/\tau^2)^n}{n!} \\
& \frac{\Gamma(MN+3/2)}{\Gamma(M-MN+n)}\left(\frac{x^2}{2}\right)^{-(MN+1/2)}\left\{1+O\left(\frac{1}{X^2}\right)\right\} \\
& \pi(M)\pi(N)\pi(\tau)\mathrm{d}M\mathrm{d}N\mathrm{d}\tau \\
= \;& -\frac{2}{X}\int_1^\infty\int_0^1\int_0^\infty C(M,N,\tau)\sum_{n=0}^\infty \frac{(M-MN)_n(M)_n}{(M+1/2)_n}\frac{(1-1/\tau^2)^n}{n!} \\
& \frac{\Gamma(MN+1/2)}{\Gamma(M-MN+n)}\left(\frac{x^2}{2}\right)^{-(MN+1/2)}\left\{1+O\left(\frac{1}{X^2}\right)\right\}\frac{MN+1/2}{M+1/2+n} \\
& \pi(M)\pi(N)\pi(\tau)\mathrm{d}N\mathrm{d}M\mathrm{d}\tau, \qquad (2.15)
\end{aligned}$$

and

$$0 \le -\frac{X}{2}\frac{\mathrm{d}m_{\tau>1}(X)}{\mathrm{d}X} \le m_{\tau>1}(X),$$

since

$$0 \le \frac{MN+1/2}{M+1/2+n} \le \frac{M+1/2}{M+1/2+n} \le 1.$$

Combining equations (2.13), (2.14) and (2.15) together, we have

$$\frac{\mathrm{d}}{\mathrm{d}X}\log m(X) = \frac{\frac{\mathrm{d}m_{\tau<1}(X)}{\mathrm{d}X}+\frac{\mathrm{d}m_{\tau>1}(X)}{\mathrm{d}X}}{m_{\tau<1}(X)+m_{\tau>1}(X)}, \qquad (2.16)$$

and (2.16) belongs to $(-4/X,0)$ when $X>0$ or $(0,-4/X)$ when $X<0$. Therefore, equation (2.8) holds as $X\to\infty$.

The inequality (2.7) comes from the continuity of $\mathrm{d}\log m(X)/\mathrm{d}X$. Plugging equation (2.2) into (2.7) gives

$$\mid X - E(\theta \mid X)\mid = \left|\frac{d}{dX}\log m(X)\right|. \qquad (2.17)$$

The continuity of $\mathrm{d}\log m(X)/\mathrm{d}X$ along with the facts that $\mathrm{d}\log m(X)/\mathrm{d}X = 0$ when $X=0$ and $\lim_{|X|\to+\infty}\mathrm{d}\log m(X)/\mathrm{d}X = 0$ shows that B exists. □

Notice that in Theorem 1, the dimension is one and σ is assumed fixed. Under the independence assumption, it is very easy to extend it to multi-dimensional case.

As a direct consequence from Theorem 1, Corollary 1 proves that the risk of the posterior mean with the AIB prior is bounded.

Corollary 1. *Suppose $\boldsymbol{X} \sim N(\boldsymbol{\theta}, I)$ and the AIB prior (2.6) on $\boldsymbol{\theta}$, then $E(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2)$ is bounded for all $\boldsymbol{\theta}$ where $\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta} \mid \boldsymbol{X})$.*

Proof.

$$\begin{aligned} E(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|^2) &= E(\sum_{i=1}^{p}(\theta_i - \hat{\theta}_i)^2) \\ &= E(\sum_{i=1}^{p}(\theta_i - X_i + X_i - \hat{\theta}_i)^2) \\ &\leq \sum_{i=1}^{p} E((\mid \theta_i - X_i \mid + \mid X_i - \hat{\theta}_i \mid)^2) \\ &\leq \sum_{i=1}^{p} E((\mid \theta_i - X_i \mid + B)^2) \\ &= p + 2\sqrt{\frac{2}{\pi}} pB + pB^2 \end{aligned}$$

□

2.3 Computational Algorithms

In the Bayesian framework, a posterior distribution

$$\pi(\boldsymbol{\theta} \mid \boldsymbol{X}) = \frac{f(\boldsymbol{X} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int f(\boldsymbol{X} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \tag{2.18}$$

is of great interest. We can draw inferences of $\boldsymbol{\theta}$ from a posterior distribution or use it as new prior information for future study. When a posterior distribution has a complicated form, even no closed-form, the Markov Chain Monte Carlo (MCMC) methods offer a class of tools to get samples from the posterior distribution. An MCMC algorithm is an iterative sampling process whose stationary distribution, in our case, is the posterior distribution. Initializing the process with arbitrary values in the domain, we obtain an ergodic process and the effect from the initial values will be "washed out". In the process of updating parameters, we apply a hybrid of Gibbs sampler and Metropolis–Hastings algorithms.

2.3.1 Monte Carlo Markov Chain Algorithm With Gibbs Sampler

Given a set of observations $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$ with likelihood (1.3) and the AIB prior in (2.6), we use the first iteration as an example to illustrate the Gibbs sampler procedure: Starting with initial values M_0, N_0, $\boldsymbol{\lambda}_0$, τ_0, σ_0^2 and $\boldsymbol{\theta}_0$,

(1) Update M with $M^{(1)}$ from $p(M \mid N_0, \boldsymbol{\lambda}_0)$.

(2) Update N with $N^{(1)}$ from $p(N \mid M^{(1)}, \boldsymbol{\lambda}_0)$.

(3) Update $\boldsymbol{\lambda}$ with $\boldsymbol{\lambda}^{(1)}$ from $p(\boldsymbol{\lambda} \mid M^{(1)}, N^{(1)}, \boldsymbol{\theta}_0, \tau_0, \sigma_0^2)$.

(4) Update τ with $\tau^{(1)}$ from $p(\tau \mid \boldsymbol{\theta}_0, \boldsymbol{\lambda}^{(1)}, \sigma_0^2)$.

(5) Update σ^2 with $\sigma^{2(1)}$ from $p(\sigma^2 \mid \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \boldsymbol{\theta}_0, \boldsymbol{\lambda}^{(1)}, \tau^{(1)})$.

(6) Update $\boldsymbol{\theta}$ with $\boldsymbol{\theta}^{(1)}$ from $p(\boldsymbol{\theta} \mid \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \sigma^{2(1)}, \boldsymbol{\lambda}^{(1)}, \tau^{(1)})$.

Here $p(\bullet \mid \bullet)$ stands for a full conditional distribution. Starting from the second iteration, the initial values are replaced by the current states to update the parameters.

When it is not analytically or computationally feasible to get a sample from the full conditional distribution, the Metropolis–Hastings algorithm is applied to generate a future state which is accepted according to an acceptance ratio. Next, we show the derivation of conditional distributions and acceptance ratios in detail.

(1) For updating M, the prior is not conjugate for the IB likelihood function and the Metropolis–Hastings algorithm is applied. Let M denote the current state and let M^* denote the future state. Considering the domain of M, a normal distribution centered at $\log(M)$ is used as the proposal distribution to generate $\log(M^*)$, which is then transformed back to get M^*. The acceptance ratio of M^* is

$$\text{Ratio}_M = \frac{\frac{1}{M}\prod_i \frac{1}{\text{Beta}(M^*N, M^*(1-N))}\frac{\lambda_i^{M^*(1-N)-1}}{(1+\lambda_i)^{M^*}} f(M^*)}{\frac{1}{M^*}\prod_i \frac{1}{\text{Beta}(MN, M(1-N))}\frac{\lambda_i^{M(1-N)-1}}{(1+\lambda_i)^{M}} f(M)}, \tag{2.19}$$

where $1/M$ and $1/M^*$ are the Jacobian determinants from the logarithm transformation, and $f(M)$ is the Gamma(2,1) density for the current value M.

(2) For updating N, the prior is not conjugate. Considering the domain of N, a normal proposal distribution centered at $\text{logit}(N)$ is used to generate $\text{logit}(N^*)$, which is transformed back to get N^*. The acceptance ratio is:

$$\text{Ratio}_N = \frac{\frac{1}{N(1-N)}\prod_i \frac{1}{\text{Beta}(MN^*, M(1-N^*))}\lambda_i^{M(1-N^*)-1}}{\frac{1}{N^*(1-N^*)}\prod_i \frac{1}{\text{Beta}(MN^*, M(1-N^*))}\lambda_i^{M(1-N)-1}}, \tag{2.20}$$

where the first terms in both the numerator and the denominator are the Jacobian determinants from the logit transformation. Note that the $(1+\lambda)^M$ part inside the IB prior does not involve N and that the $\text{Beta}(1,1)$ prior on N is equivalent to the standard uniform distribution, so they are both omitted.

(3) For updating $\boldsymbol{\lambda}$, since the elements of $\boldsymbol{\lambda}$ are conditionally independent, we use one dimension as an example to show the computation details and the acceptance ratio. Considering the domain of λ_i and the prior on it, a normal proposal

distribution centred at $\log(\lambda_i)$ is used to generate $\log(\lambda_i^*)$. Then $\log(\lambda_i^*)$ is transformed back to get λ_i^*. The acceptance ratio is:

$$\text{Ratio}_{\lambda_i} = \frac{\frac{1}{\lambda_i} \frac{1}{\sqrt{2\pi\lambda_i^*}\tau\sigma} e^{-\frac{1}{2\lambda_i^*\tau^2\sigma^2}\theta_i^2} \frac{\lambda_i^{*M(1-N)-1}}{(1+\lambda_i^*)^M}}{\frac{1}{\lambda_i^*} \frac{1}{\sqrt{2\pi\lambda_i}\tau\sigma} e^{-\frac{1}{2\lambda_i\tau^2\sigma^2}\theta_i^2} \frac{\lambda_i^{M(1-N)-1}}{(1+\lambda_i)^M}}, \tag{2.21}$$

with the constant part inside the IB prior cancelled.

(4) For updating τ, a normal proposal distribution centred at $\log(\tau)$ is used to generate $\log(\tau^*)$, which is transformed back to get τ^*. The acceptance ratio is:

$$\text{Ratio}_{\tau} = \frac{\frac{1}{\tau} \prod_i \frac{1}{\sqrt{2\pi\lambda_i}\tau^*\sigma} e^{-\frac{1}{2\lambda_i\tau^{*2}\sigma^2}\theta_i^2} \frac{2}{\pi(1+\tau^{*2})}}{\frac{1}{\tau^*} \prod_i \frac{1}{\sqrt{2\pi\lambda_i}\tau\sigma} e^{-\frac{1}{2\lambda_i\tau^2\sigma^2}\theta_i^2} \frac{2}{\pi(1+\tau^2)}}. \tag{2.22}$$

(5) For updating σ^2, the inverse gamma prior is a conjugate prior for normal distributions and the Jeffreys prior is a special case of the inverse gamma family. Therefore, the full conditional distribution is

$$\sigma^2 \mid \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \tau \sim \text{IG}\left(\frac{pq+p}{2}, \sum_i \frac{\theta_i^2}{2\tau\lambda_i} + \sum_{i,j} \frac{(Y_{i,j}-\theta_i)^2}{2}\right), \tag{2.23}$$

where IG stands for the inverse gamma distribution.

(6) For updating $\boldsymbol{\theta}$, since the prior is conjugate and the elements of $\boldsymbol{\theta}$ are conditionally independent, the full conditional distribution of θ_i is:

$$\theta_i \mid \bar{X}_i, \sigma, \lambda_i, \tau \sim \text{N}\left(\frac{n\lambda_i\tau^2\bar{X}_i}{n\lambda_i\tau^2+1}, \frac{\lambda_i\tau^2\sigma^2}{n\lambda_i\tau^2+1}\right), \tag{2.24}$$

where $\bar{\boldsymbol{X}}$ is the sample average of $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n\}$ and $\bar{X}_i$ is the i-th element of $\bar{\boldsymbol{X}}$.

2.3.2 Monte Carlo Markov Chain Algorithm With Partially Collapsed Gibbs Sampler

Although the Gibbs sampler is very popular for its simplicity, it is often criticized for its slow convergence speed, especially when the model structure is complex (Park and Dyk 2009). Instead of using the Gibbs sampler, we can use the partially collapsed Gibbs sampler to accelerate the convergence speed by drawing $\boldsymbol{\lambda}$, τ and σ from the full conditional distributions in which $\boldsymbol{\theta}$ is integrated out. Similar to Section 2.3.1, we use the first iteration as an example to illustrate the partially collapsed Gibbs sampler procedure. Starting with initial values M_0, N_0, $\boldsymbol{\lambda}_0$, τ_0, σ_0^2 and $\boldsymbol{\theta}_0$,

(1) Update M with $M^{(1)}$ from $p(M \mid N_0, \boldsymbol{\lambda}_0)$.

(2) Update N with $N^{(1)}$ from $p(N \mid M^{(1)}, \boldsymbol{\lambda}_0)$.

(3) Update $\boldsymbol{\lambda}$ with $\boldsymbol{\lambda}^{(1)}$ from $p(\boldsymbol{\lambda} \mid \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, M^{(1)}, N^{(1)}, \tau_0, \sigma_0^2))$.

(4) Update τ with $\tau^{(1)}$ from $p(\tau \mid \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \boldsymbol{\lambda}^{(1)}, \sigma_0^2)$.

(5) Update σ^2 with $\sigma^{2(1)}$ from $p(\sigma^2 \mid \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \boldsymbol{\lambda}^{(1)}, \tau^{(1)})$.

(6) Update $\boldsymbol{\theta}$ with $\boldsymbol{\theta}^{(1)}$ from $p(\boldsymbol{\theta} \mid \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n, \sigma^{2(1)}, \boldsymbol{\lambda}^{(1)}, \tau^{(1)})$.

Again, starting from the second iteration, we use the updated values instead of the initial values to get future states of the parameters in the above procedure.

Since the elements of $\boldsymbol{X}_i$ are mutually independent given $\boldsymbol{\theta}$ and σ^2 and the elements of $\boldsymbol{\theta}$ are mutually independent given τ, σ^2 and $\boldsymbol{\lambda}$, the elements of $\boldsymbol{X}_i$ are still mutually independent after integrating $\boldsymbol{\theta}$ out. By some straightforward algebra, we get the joint distribution of $\{X_{ij},\ j = 1, \cdots, n\}$ given σ^2, τ and λ_i:

$$p(X_{i1}, \cdots, X_{in} \mid \sigma^2, \tau, \lambda_i) \propto (\sigma^2)^{n/2}(n\tau^2\lambda_i{+}1)^{-1/2} \exp\left\{ -\frac{\sum_j X_{ij}^2}{2\sigma^2} + \frac{\lambda_i \tau^2 (\sum_j X_{ij})^2}{2\sigma^2(n\tau^2\lambda_i + 1)} \right\}.$$

Similar to section 2.3.1, we get the acceptance ratios and conditional distributions as follows:

(1) For updating M

$$\mathrm{Ratio}_M = \frac{\frac{1}{M}\prod_i \frac{1}{\mathrm{Beta}(M^*N, M^*(1-N))}\frac{\lambda_i^{M^*(1-N)-1}}{(1+\lambda_i)^{M^*}} f(M^* \mid 2,1)}{\frac{1}{M^*}\prod_i \frac{1}{\mathrm{Beta}(MN, M(1-N))}\frac{\lambda_i^{M(1-N)-1}}{(1+\lambda_i)^{M}} f(M \mid 2,1)}. \tag{2.25}$$

(2) For updating N

$$\mathrm{Ratio}_N = \frac{\frac{1}{N(1-N)}\prod_i \frac{1}{\mathrm{Beta}(MN^*, M(1-N^*))}\lambda_i^{M(1-N^*)-1}}{\frac{1}{N^*(1-N^*)}\prod_i \frac{1}{\mathrm{Beta}(MN^*, M(1-N^*))}\lambda_i^{M(1-N)-1}}. \tag{2.26}$$

(3) For updating λ_i

$$\mathrm{Ratio}_{\lambda_i} = \frac{\frac{1}{\lambda_i}p(X_{i1}, \cdots, X_{in} \mid \sigma^2, \tau, \lambda_i^*)\frac{\lambda_i^{*M(1-N)-1}}{(1+\lambda_i^*)^M}}{\frac{1}{\lambda_i^*}p(X_{i1}, \cdots, X_{in} \mid \sigma^2, \tau, \lambda_i)\frac{\lambda_i^{M(1-N)-1}}{(1+\lambda_i)^M}}. \tag{2.27}$$

(4) For updating τ

$$\mathrm{Ratio}_\tau = \frac{\frac{1}{\tau}\prod_i p(X_{i1}, \cdots, X_{in} \mid \sigma^2, \tau^*, \lambda_i)\frac{2}{\pi(1+\tau^{*2})}}{\frac{1}{\tau^*}\prod_i p(X_{i1}, \cdots, X_{in} \mid \sigma^2, \tau, \lambda_i)\frac{2}{\pi(1+\tau^2)}}. \tag{2.28}$$

(5) For updating σ^2, the distribution of σ^2 conditional on $X_{i1}, \cdots, X_{in}, \tau, \boldsymbol{\lambda}$ is $\mathrm{IG}(\alpha, \beta)$, where

$$\begin{aligned} \alpha &= \frac{np}{2}, \\ \beta &= \sum_{i,j}\frac{X_{ij}^2}{2} - \sum_i \frac{\lambda_i\tau^2(\sum_j X_{ij})^2}{2(n\tau^2\lambda_i + 1)}. \end{aligned}$$

(6) For updating θ_i, the full conditional distribution is

$$\theta_i \mid \bar{X}_i, \sigma, \lambda_i, \tau \sim \mathrm{N}\left(\frac{n\lambda_i\tau^2\bar{X}_i}{n\lambda_i\tau^2+1}, \frac{\lambda_i\tau^2\sigma^2}{n\lambda_i\tau^2+1}\right). \tag{2.29}$$

2.4 Simulation Studies

2.4.1 Simulation 1 – Comparison Between The AIB Prior And Other Shrinkage Priors

In this section, we compare the estimation performances of the AIB prior with those of the horseshoe prior (HS), the Strawderman–Berger prior (SB), the NEG(2, 1) prior (NEG) and the double-exponential prior (DE) under different scenarios. For the purpose of fair comparisons, we let these priors have the common structure

$$\begin{aligned}\boldsymbol{\theta} \mid \boldsymbol{\lambda}, \tau, \sigma &\sim \mathrm{N}(\mathbf{0}, \tau^2\sigma^2\mathrm{Diag}(\boldsymbol{\lambda})),\\ p(\sigma^2) &\propto \frac{1}{\sigma^2},\\ \tau &\sim \mathrm{C}^+(0,1), \end{aligned} \tag{2.30}$$

with the priors on λ_i being IB$(MN, M(1-N))$, IB$(1/2, 1/2)$, IB$(1/2, 1)$, IB$(2, 1)$ and the standard double-exponential distribution, respectively.

Considering different sparsity levels and signal sizes, we select the combinations of three nonzero percentages 5%, 20% and 50%, and three nonzero values 1, 4, 10. For each combination, 100 data sets are generated with the parameter dimension $p = 100$. To generate the true $\boldsymbol{\theta}$ and $\boldsymbol{Y}$, we let all nonzero elements of $\boldsymbol{\theta}$ have the same value. In order to increase the stability, we fix the number of nonzero θ_i's and randomly select their locations. For example, for the 20% nonzero value at 4 scenario, we first randomly select 20 dimensions of $\boldsymbol{\theta}$ setting to 4 and then set the rest to 0. Without loss of generality, two observations $\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$, are generated using the likelihood function (1.3). The sufficient statistics in this case are $(\bar{\boldsymbol{Y}}, S^2)$, where

$$\begin{aligned}\bar{\boldsymbol{Y}} &= \frac{1}{2}(\boldsymbol{Y}_1 + \boldsymbol{Y}_2) \sim \mathrm{N}\left(\boldsymbol{\theta}, \frac{\sigma^2}{2}I\right),\\ S^2 &= (\boldsymbol{Y}_1 - \boldsymbol{Y}_2)^T(\boldsymbol{Y}_1 - \boldsymbol{Y}_2) \sim \sigma^2\chi_{(100)}.\end{aligned}$$

We set $\sigma^2 = 2$, so that $\bar{\boldsymbol{Y}}$ has variance 1.

To fit each method, 40000 iterations are run with a burn-in of the first 20000 iterations. The sum of squared error loss is used as the comparison criterion, that is,

$$\mathrm{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}), \tag{2.31}$$

where $\hat{\boldsymbol{\theta}}$ is a posterior mean of $\boldsymbol{\theta}$.

When the nonzero value is fixed at 1, the signals are hard to distinguish from the noises, over shrinkage may result in less losses, so the NEG model and the DE model may work well in such scenarios. When the nonzero value is 4, the noises and signals are partially mixed. Incorrect discriminations of noises and signals will lead to severely large losses. Therefore, all methods are expected to have larger losses in these scenarios than those when the signal is 1. When the nonzero value is 10, noises and signals are much separated and incorrect discriminations may hardly happen. The AIB, the HS and the SB models are expected to have losses greater than those when the signal is 1, but less than those when the signal is 4. But the NEG and the DE models may suffer from the over shrinkage issue and have greater losses than those when the signal size is 1 or 4.

Table 2.2 reports the averaged square error losses across 100 datasets in upper rows and the standard deviations of the average losses in lower rows in each cell. The risks of the MLE are given as benchmarks. Similarly, Table 2.3 and Table 2.4 report the parameter estimates in the upper rows and the average posterior standard deviations in the lower rows. Figure 2.3 plots the mean values $\bar{Y}_i$ against the posterior means of θ_i for each models in different scenarios. Figure 2.4 and Figure 2.5 plot the posterior means of the global shrinkage parameter τ and the local shrinkage parameters λ_i for all models, offering a closer look at underlying signal/noise discrimination. Note that all dimensions of $\boldsymbol{\lambda}$ are plotted in one box in Figure 2.5 for each model in each scenario.

Scenario		AIB	HS	SB	NEG	DE	MLE
1	5%	5.42	5.35	5.34	5.67	5.94	101.67
		0.18	0.16	0.18	0.22	0.25	1.48
	20%	17.67	18.08	18.24	17.74	17.88	100.82
		0.20	0.20	0.21	0.22	0.25	1.34
	50%	37.49	42.20	41.90	37.24	36.40	99.86
		0.42	0.36	0.39	0.41	0.43	1.60
4	5%	21.07	18.55	18.76	26.50	31.87	98.55
		0.93	0.88	0.90	0.83	0.79	1.43
	20%	55.21	52.54	54.01	61.00	63.78	99.46
		1.31	1.25	1.23	1.12	1.09	1.55
	50%	90.44	90.79	91.67	90.91	90.76	98.63
		1.24	1.67	1.53	1.33	1.28	1.30
10	5%	12.73	13.24	13.46	27.96	46.11	100.08
		0.67	0.66	0.68	0.88	0.99	1.52
	20%	37.53	44.56	45.50	68.70	76.45	100.61
		1.09	1.10	1.11	1.29	1.36	1.57
	50%	80.03	80.36	86.83	92.41	93.72	101.73
		1.42	1.23	1.28	1.32	1.33	1.38

Table 2.2: Comparison of losses among the AIB, the HS, the SB, the NEG and the DE models under the IID assumptions in different scenarios. The losses of the MLE estimate are given as benchmarks. The averaged squared error losses across 100 datasets are reported in the upper rows and the standard deviations of the average losses are reported in the lower rows. For all datasets, $p = 100$ and 2 observations are generated for each dimension. The signal sizes are listed in the first column and the nonzero percentages are in the second column.

Generally speaking, all methods have smaller losses than those of the MLE in all scenarios. For any fixed signal size, all models have greater losses when the sparsity decreases. For a fixed sparsity level, as we expected, the AIB, the HS and the SB models have the largest losses when the signal is 4. However, the NEG prior and the

DE prior always have greater losses under bigger signal sizes. This shows that the over shrinkage issue is very severe to both models.

When the signal is 1, all models work well for sparse datasets. The AIB, the NEG and the DE models work a little better than the other two methods when the nonzero percentage is 50. When the signal is 4 and the nonzero percentage is 5 or 20, the NEG and the DE models have bigger losses than other models. The AIB model has slightly bigger losses than the HS and the SB models but is still better than the NEG and the DE models. When the nonzero percentage is high, all models have similar performances. When the signal size is 10, the AIB prior consistently gives the smallest losses, especially when the nonzero percentage is 20. This shows that the AIB prior has the ability to yield better performances than other models when they work normally or poorly. The SB, the NEG and the DE models have larger losses than the other two models.

Figure 2.3 provides explanations to the patterns found in Table 2.2. When the signal is 1 (first row in Figure 2.3), all methods have similar shrinkage patterns near 0, but the HS and the SB model have sharper slope changes, especially when the nonzero percentage is 50. However, this shrinkage pattern is not favored by the data which leads to bigger losses. When the signal is 4, especially when the nonzero percentage is 5 or 20, the over shrinkage issues of the NEG and the DE model are obvious. The AIB prior has similar but lighter issues in these scenarios. Carefully looking at the last row in Figure 2.3, we can see that the AIB model has stronger shrinkage for noises which is desired for these scenarios and this explains why the AIB model has smaller losses.

Scenario		M	N	τ(AIB)	τ(HS)	τ(SB)	τ(NEG)	τ(DE)
1	5%	2.75	0.76	0.35	0.04	0.03	0.25	0.24
		1.325	0.148	0.198	0.026	0.017	0.090	0.141
	20%	2.79	0.75	0.43	0.06	0.04	0.32	0.31
		1.582	0.172	0.328	0.072	0.038	0.193	0.155
	50%	3.08	0.71	0.54	0.09	0.06	0.47	0.46
		1.467	0.147	0.393	0.093	0.066	0.195	0.168
4	5%	1.90	0.73	0.97	0.17	0.10	0.70	0.69
		0.859	0.128	0.616	0.069	0.044	0.160	0.141
	20%	2.59	0.74	2.50	0.57	0.35	1.47	1.33
		1.002	0.087	1.473	0.132	0.078	0.218	0.193
	50%	6.23	0.53	2.10	1.43	0.90	2.65	2.19
		2.064	.132	0.820	0.306	0.180	0.406	0.301
10	5%	0.91	0.67	0.67	0.19	0.12	0.95	1.25
		0.329	0.107	0.366	0.069	0.039	0.172	0.166
	20%	1.62	0.90	12.24	0.80	0.46	2.78	2.79
		0.814	0.067	5.197	0.197	0.105	0.373	0.342
	50%	3.05	0.84	15.17	3.60	2.09	6.78	5.60
		1.262	0.060	5.395	0.697	0.391	0.907	0.690

Table 2.3: Comparison of hyperparameters among the AIB, the HS, the SB, the NEG and the DE models under the IID assumptions in different scenarios. The averaged posterior means across 100 datasets are reported in the upper rows. The averaged posterior standard deviations are reported in the lower rows. For all datasets, $p = 100$ and 2 observations are generated for each dimension. The signal sizes are listed in the first column and the nonzero percentages are in the second column.

Next we look at the estimates of hyperparameters. The estimates of τ of the HS, the SB, the NEG and the DE models increase as the sparsity decreases or the signal size increases. The NEG and the DE models have larger estimates of τ than the HS model, and the HS model has larger estimates than the SB model across all scenarios. The reason for this pattern is that, the NEG model has greater value of N. As described in section 2.2.1, N is a measurement of the overall shrinkage strength from the local shrinkage parameters. A larger N tends to give stronger shrinkage and vice versa. The N value for the NEG model is $2/3$, while the N values for the HS and the SB models are 1/2 and 1/3 respectively. Therefore, the NEG prior tends to provide smaller estimates of λ_i's which leads to bigger estimates of τ than the other two models. Although the DE prior does not follow the IB form in (2.3), the density of κ_i's vanishes at 0 which avoids λ_i's to be very large and results in bigger estimates of τ.

The estimates of τ and N in the AIB prior do not have consistent changing patterns individually. The estimates of τ have clearly larger posterior standard deviations than the other four methods. The posterior standard deviations of N are also big. This is because that both N and τ are closely associated with the local shrinkage parameters λ_i's. The N value governs the shape of the prior on λ_i's and τ corporates with λ_i's to determine the shrinkage degrees. When both parameters are unfixed, more variability exists. Although the estimates of τ in the AIB model are not as stable as those in other priors, and do not have a clear changing pattern, it does consistently increase or decrease as the estimate of N decreases or increases.

Figure 2.4 clearly shows the observations found in Table 2.3. The NEG and the DE models have larger estimates of τ than the HS model and the HS model has larger estimates than the SB model across all scenarios. When the sparsity decreases, all methods have larger estimates of τ, which indicates that less shrinkage degrees are given from the global shrinkage parameter. The changes in the AIB prior are more than those of other priors. In every scenario, the estimates of τ in the AIB model have a clearly wider range than the other four methods. The range of the AIB prior becomes larger as the signal size increases, but those of other priors are relatively stable. Again this shows the increased variability when both N and τ are unfixed. Interestingly, when there are 20% or 50% nonzero values at 10, the estimates of τ in the AIB prior are much larger than those from the NEG and the DE priors. This indicates that the local shrinkage parameters in the AIB prior give more shrinkage to noises than those in the NEG and the DE priors. In addition, this also indicates that when both N and τ are unfixed, the global shrinkage parameter τ is more sensitive to the large signals than the local shrinkage parameters. In other words, when the percentage of large signals increases, the change of shrinkage power in the AIB prior is mainly expressed through the global shrinkage parameter τ.

Figure 2.5 supports the indications from Figure 2.4 and shows some new patterns. The NEG and the DE models have small estimates of λ_i's in all scenarios, which indicates strong shrinkage powers from the local shrinkage parameters. The AIB prior has similar performances in most scenarios but not in the combination of 5% nonzero value at 10. The boxes of the other four priors vary little in different scenarios, but the boxes of the AIB prior vary quite obviously. This is consistent with the pattern in Figure 2.4 and provides more evidence to support the great variability in the AIB prior.

Other than the above observations, a few new phenomena need to be noticed. First of all, fixing the nonzero percentage at 5, the estimates of $\boldsymbol{\lambda}$ in the AIB prior have a much wider range when the signal size is 10. Associating this graph with the corresponding graph in Figure 2.4, we can see that most estimates of τ in this scenario are less than 2. Considering the large signal size, more large estimates of λ_i are on demand to reserve the shrinkage adaptivity. Secondly, when there are 50% nonzero values at 4 or 10, the boxes of $\boldsymbol{\lambda}$ from the AIB prior barely have tails. When the signal size is 4, most estimates of τ are in the range $(1.5, 3)$. Considering the signal size and the range of τ, λ_i's do not need to be extremely large. Similarly, when the signal is 10, most estimates of τ are in the range $(10, 25)$. Given large estimates of τ, λ_i's do not need to be large as well. On the other hand, in the 20% nonzero values at 10 scenario, we see a few small estimates of τ less than 5 in Figure 2.4. For those τ, large estimates of λ_i's are needed to provide little shrinkage. These findings support the claim we made previously, that is the global shrinkage parameter τ in the AIB prior is more sensitive to the increased amount of large signals.

Scenario		σ(AIB)	σ(HS)	σ(SB)	σ(NEG)	σ(DE)
1	5%	1.40	1.40	1.40	1.39	1.38
		0.073	0.072	0.073	0.072	0.081
	20%	1.42	1.44	1.44	1.41	1.41
		0.087	0.085	0.084	0.091	0.081
	50%	1.46	1.50	1.50	1.45	1.43
		0.098	0.098	0.104	0.095	0.098
4	5%	1.37	1.37	1.37	1.36	1.36
		0.087	0.086	0.085	0.090	0.090
	20%	1.37	1.37	1.37	1.37	1.37
		0.096	0.098	0.099	0.098	0.094
	50%	1.46	1.48	1.48	1.46	1.46
		0.109	0.115	0.114	0.109	0.107
10	5%	1.35	1.35	1.35	1.32	1.33
		0.074	0.073	0.073	0.076	0.082
	20%	1.35	1.35	1.35	1.35	1.36
		0.081	0.082	0.082	0.091	0.091
	50%	1.36	1.35	1.37	1.38	1.39
		0.091	0.091	0.094	0.095	0.097

Table 2.4: Comparison of estimates of σ among the AIB, the HS, the SB, the NEG and the DE models under the IID assumptions in different scenarios. The averaged posterior means across 100 datasets are reported in the upper rows. The averaged posterior standard deviations are reported in the lower rows. For all datasets, $p = 100$ and 2 observations are generated for each dimension. The signal sizes are listed in the first column and the nonzero percentages are in the second column.

All methods have estimates of σ close to the true value $\sqrt{2}$. Fixing the signal size, a slightly increasing pattern can be detected in the estimates and in the averaged posterior standard deviations.

2.4.2 Simulation 2 – Study Of The Global Shrinkage Parameter τ In The AIB prior

In this section, we study the effect of the global shrinkage parameter τ in the AIB prior, by comparing the AIB model in (2.6) with the AIB model when τ is fixed at 1. We use the same datasets in Simulation 1 to fit the AIB model with $\tau = 1$. The averaged square error losses and the average posterior means of parameters across the 100 datasets are reported in the upper rows of Table 2.5. The standard deviations of the average losses and the averaged posterior standard deviations are reported in the lower rows. The columns with (τ) stand for the AIB model with τ, and others represent the one without τ. Similar to Simulation 1, Figure 2.6 plots the posterior means of θ_i's against $\bar{Y}_i$ using the 100 datasets for each model. Figures 2.7 and 2.8 compare the boxplots of N and λ_i's in each model.

Table 2.5 shows that adding τ improves the performance a little in most scenarios, and remarkably in the last scenario. For the estimates of N, the decreasing pattern in the AIB without τ model is clear when the sparsity level decreases or the signal size increases. This reflects that the AIB prior is able to detect the changes in sparsity levels and signal sizes, and adjust the shrinkage strength via N. This also illustrates the flexibility of the AIB prior in handling different types of data with variety sparsity levels and signal sizes. However, in contrast, this pattern is not clear in the with τ model. Furthermore, the posterior standard deviations of N become larger after adding τ into the model in most scenarios. As shown in Figure 2.7, when τ is unfixed, the posterior mean of N becomes less stable. Figure 2.8 shows that, without the help of τ, the estimates of λ_i are clearly larger when the signal size is large and the sparsity is not high. All these observations again support the arguments we made previously: (1) When both N and τ are unfixed, more variability exists; (2) When the data appear to have more large signals, the global shrinkage parameter is more sensitive to this change and the global shrinkage power increases outstandingly. Although the AIB with τ model has the potential unstable issue, Figure 2.6 clearly shows that adding τ provides more shrinkage for small observations and less shrinkage for large observations. Thus, we recommend to use the AIB with τ model as a conclusion.

2.5 Discussion

In this chapter, we focus on the normal mean estimation problem for IID observations. The maximum likelihood estimator (MLE), a standard estimator for this problem, is inadmissible under squared error loss when the parameter's dimension is high. We propose the adaptive inverted-Beta prior (AIB), which can yield adaptive Bayesian estimators for the normal mean parameter and improve the MLE outstandingly and consistently.

By putting a general inverted-Beta prior (IB) on the local shrinkage parameters λ_i, many common shrinkage priors are included as special cases, for example, the horseshoe prior, the Strawderman-Berger prior and the normal-exponential-gamma prior with the second parameter fixed at 1. These priors are proposed to handle data with certain sparsity properties. We demonstrate through simulation studies that the AIB prior can yield similar performances as these priors when they work well, and can substantially improve them when they work normally or poorly. By putting a beta(1,1) prior (or the standard uniform prior) on N, we allow the data to determine the overall shrinkage degrees from the local shrinkage parameters. This idea offers the AIB prior the flexibility to handle different types of data. Furthermore, we study the global shrinkage parameter τ in the AIB prior. We illustrate the advantages and disadvantages of including this parameter through simulation studies. Although including τ involves more variability, it offers more improvement in shrinkage performances.

In our simulations, we assume two observations available per dimension. In application, when there are more data available, we can split the dataset into a training dataset and a testing dataset, and use the training data to fit the AIB model to draw inferences about the parameters, especially about N. Then when we analyse the testing data, we can modify the prior on N by putting more densities around the inferred value which can effectively stabilize the posterior distributions and may have better performances. This is another unique advantage by unfixing N compared to the priors with fixed N.

At last, the AIB prior can be applied in other settings, for example, in normal linear regressions or when there is a general covariance matrix in the likelihood. Chapter 3 focuses on the application of the AIB prior in normal linear regressions.

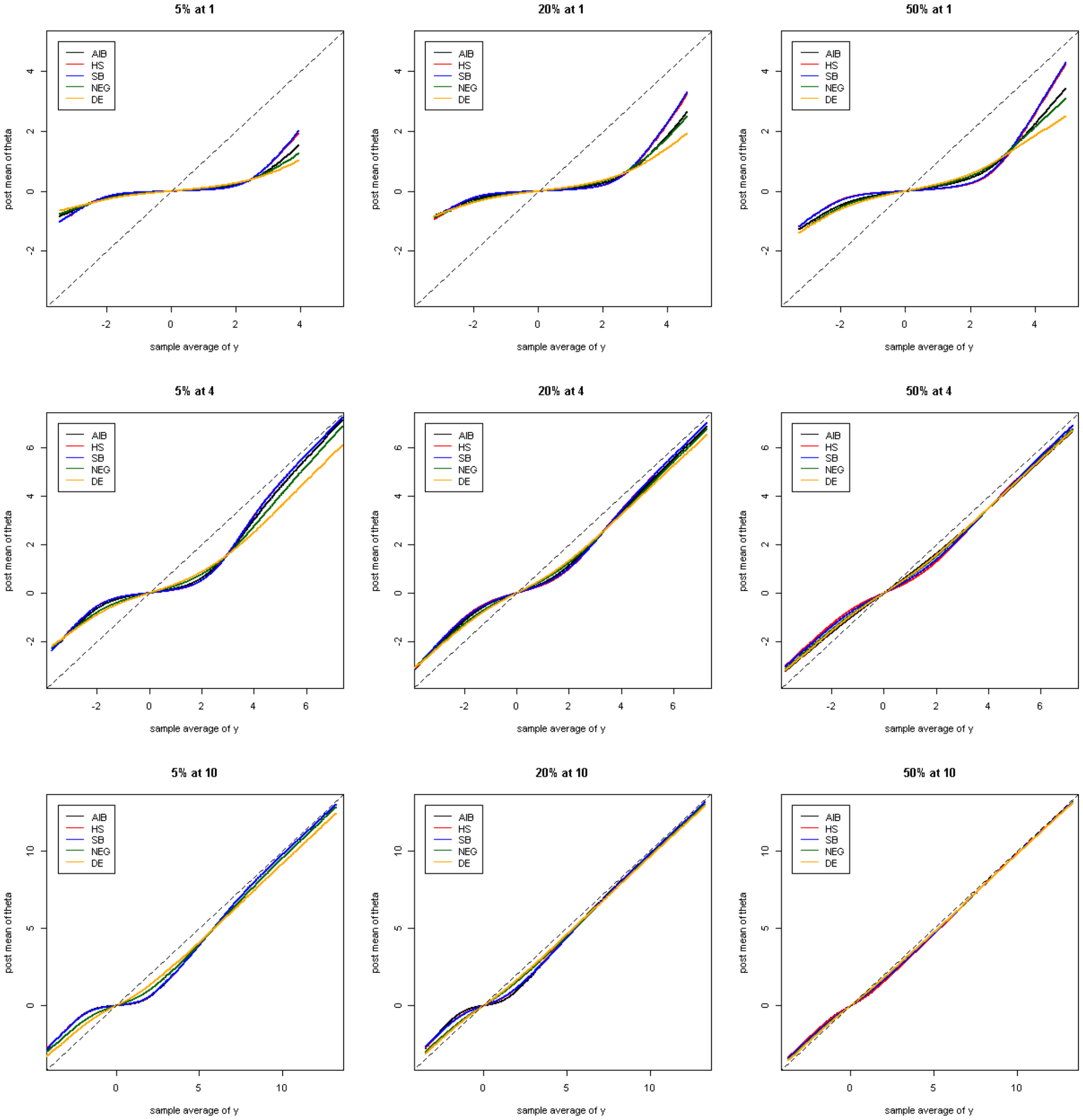

Figure 2.3: Shrinkage performance comparison among the AIB, the HS, the SB, the NEG and the DE models under the IID assumptions in different scenarios. All graphs plot the posterior means of θ_i's against $\bar{Y}_i$ using 100 datasets. For all datasets, $p = 100$ and 2 observations are generated for each dimension. The signal size is fixed at 1 in the upper row, 4 in the middle row and 10 in the lower row. The nonzero percentage is fixed at 5 in the left column, 20 in the middle column and 50 in the right column.

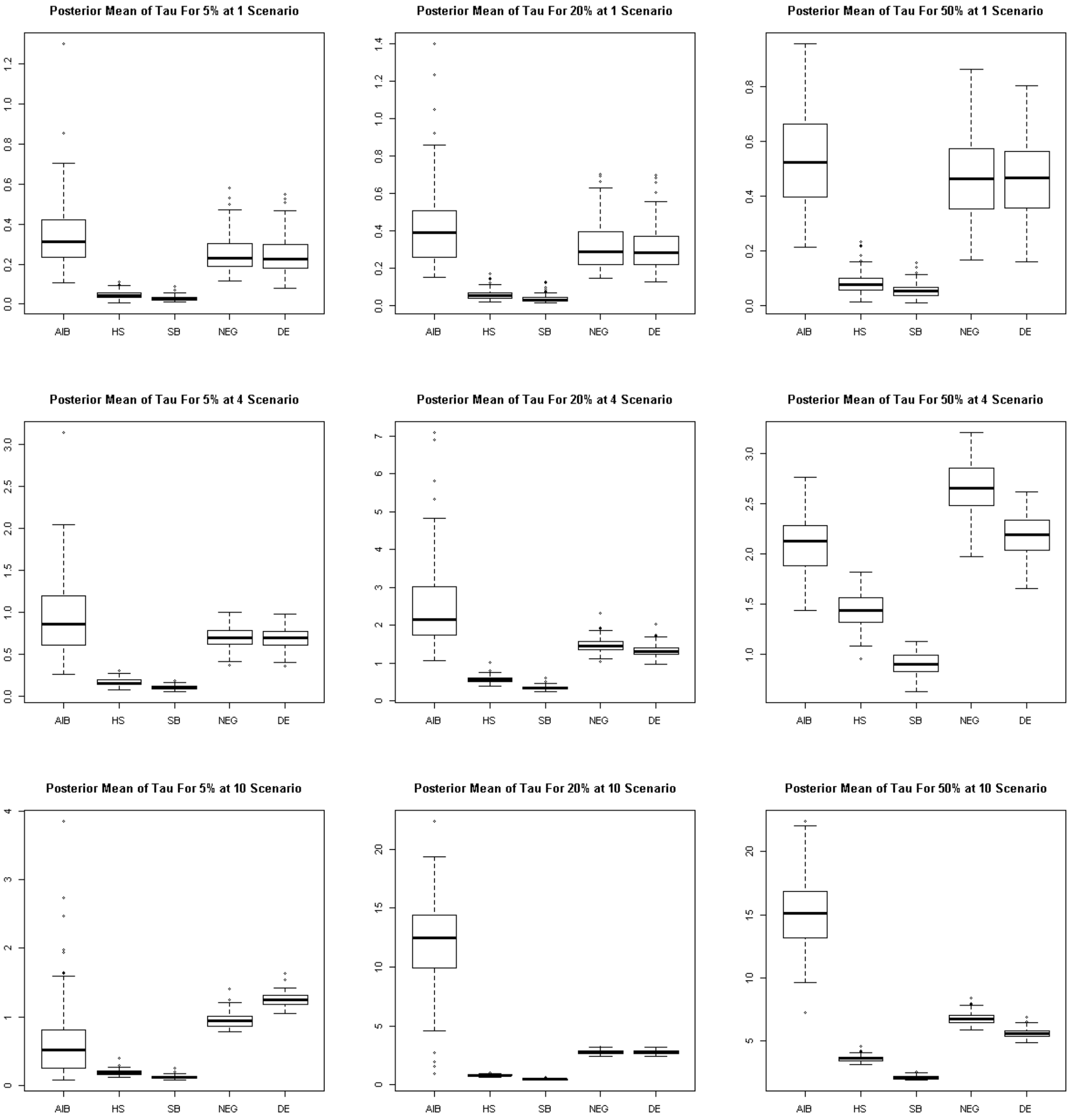

Figure 2.4: Boxplots of posterior means of τ in the AIB, the HS, the SB, the NEG and the DE models under the IID assumptions. Each box contains 100 posterior means. For all datasets, $p = 100$ and 2 observations are generated for each dimension. The signal size is fixed at 1 in the upper row, 4 in the middle row and 10 in the lower row. The nonzero percentage is fixed at 5 in the left column, 20 in the middle column and 50 in the right column.

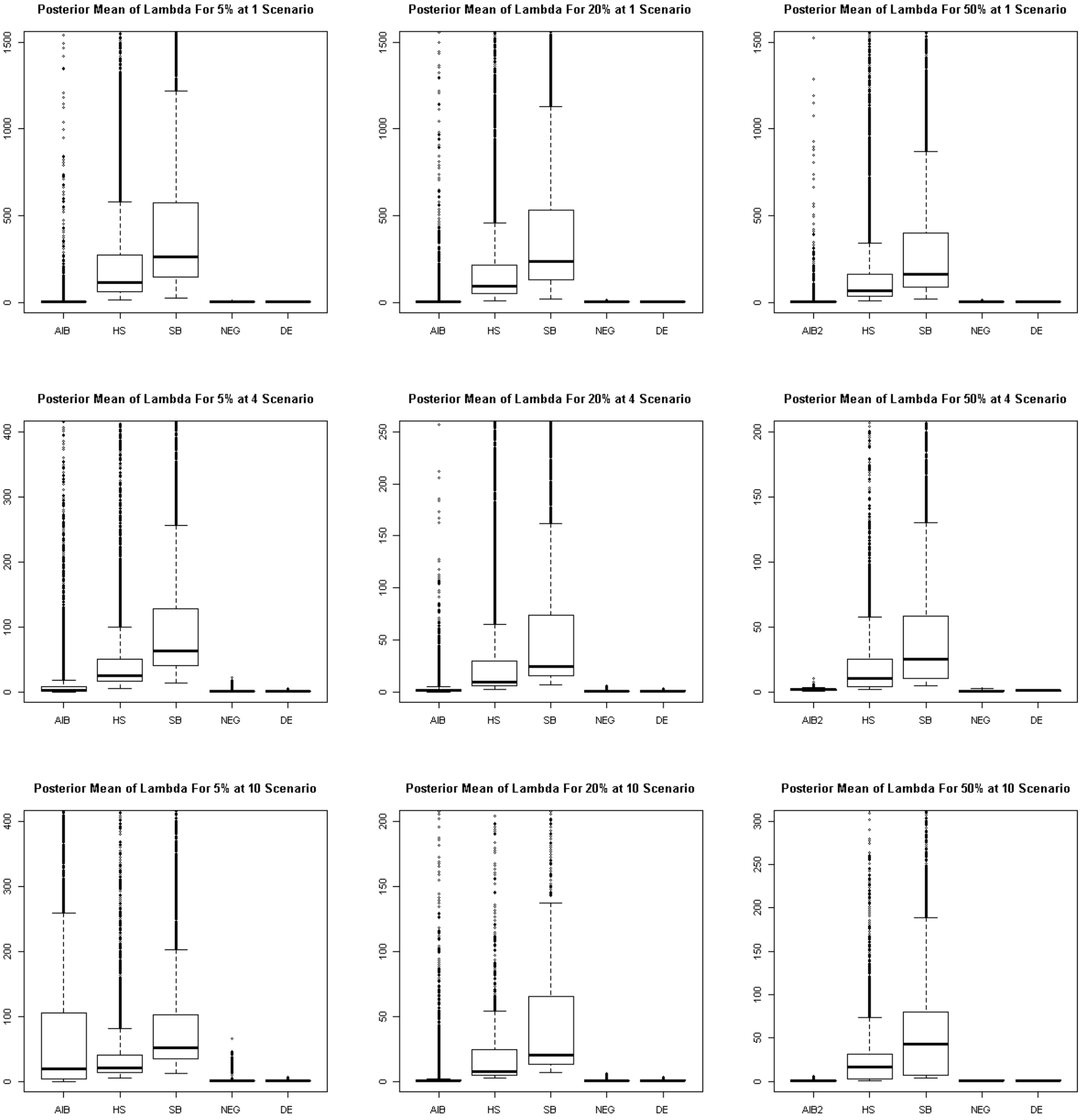

Figure 2.5: Boxplots of posterior means of λ_i's in the AIB, the HS, the SB, the NEG and the DE models under the IID assumptions. All dimensions of $\boldsymbol{\lambda}$ are plotted in one box using 100 datasets for each model. In each dataset, $p = 100$ with 2 observations per dimension. The signal size is fixed at 1 in the upper row, 4 in the middle row and 10 in the lower row. The nonzero percentage is fixed at 5 in the left column, 20 in the middle column and 50 in the right column.

Scenario		AIB	AIB(τ)	M	M(τ)	N	N(τ)	σ	$\sigma(\tau)$
1	5%	7.82	5.42	3.26	2.75	0.91	0.76	1.36	1.40
		0.29	0.18	1.569	1.325	0.025	0.148	0.072	0.081
	20%	18.74	17.67	3.05	2.79	0.90	0.75	1.39	1.42
		0.32	0.20	1.311	1.582	0.050	0.172	0.077	0.087
	50%	37.72	37.49	3.38	3.08	0.84	0.71	1.43	1.46
		0.45	0.42	1.371	1.467	0.064	0.147	0.084	0.098
4	5%	21.34	21.07	1.79	1.90	0.83	0.73	1.36	1.37
		0.90	0.93	0.667	0.859	0.041	0.128	0.082	0.087
	20%	57.42	55.21	2.07	2.59	0.71	0.74	1.39	1.37
		1.29	1.31	0.369	1.002	0.044	0.087	0.098	0.096
	50%	92.27	90.44	6.77	6.23	0.30	0.53	1.51	1.46
		1.26	1.24	2.188	2.064	0.046	0.132	0.114	0.109
10	5%	16.17	12.73	0.86	0.91	0.81	0.67	1.34	1.35
		0.68	0.67	0.169	0.329	0.032	0.107	0.072	0.074
	20%	37.82	37.53	0.60	1.62	0.64	0.90	1.35	1.35
		1.04	1.09	0.198	0.814	0.068	0.067	0.085	0.081
	50%	92.02	80.03	4.91	3.05	0.15	0.84	1.43	1.36
		1.40	1.42	2.611	1.262	0.034	0.060	0.101	0.091

Table 2.5: Comparison of losses and estimates of parameters between the AIB model with $\tau = 1$ and the AIB model in (2.6) under the IID assumptions in different scenarios. The signal sizes are listed in the first column and the nonzero percentages are in the second column. The averaged square error losses across 100 datasets are reported in column 3 and 4. The averaged posterior means of parameters are listed in columns 5 to 10. The columns with (τ) represent the model in (2.6) and those without stand for the AIB model with $\tau = 1$. The averaged losses or posterior means are in the upper rows and the standard deviations of the averaged losses and the averaged posterior standard deviations are reported in the lower rows. In all datasets, $p = 100$ with 2 observations per dimension.

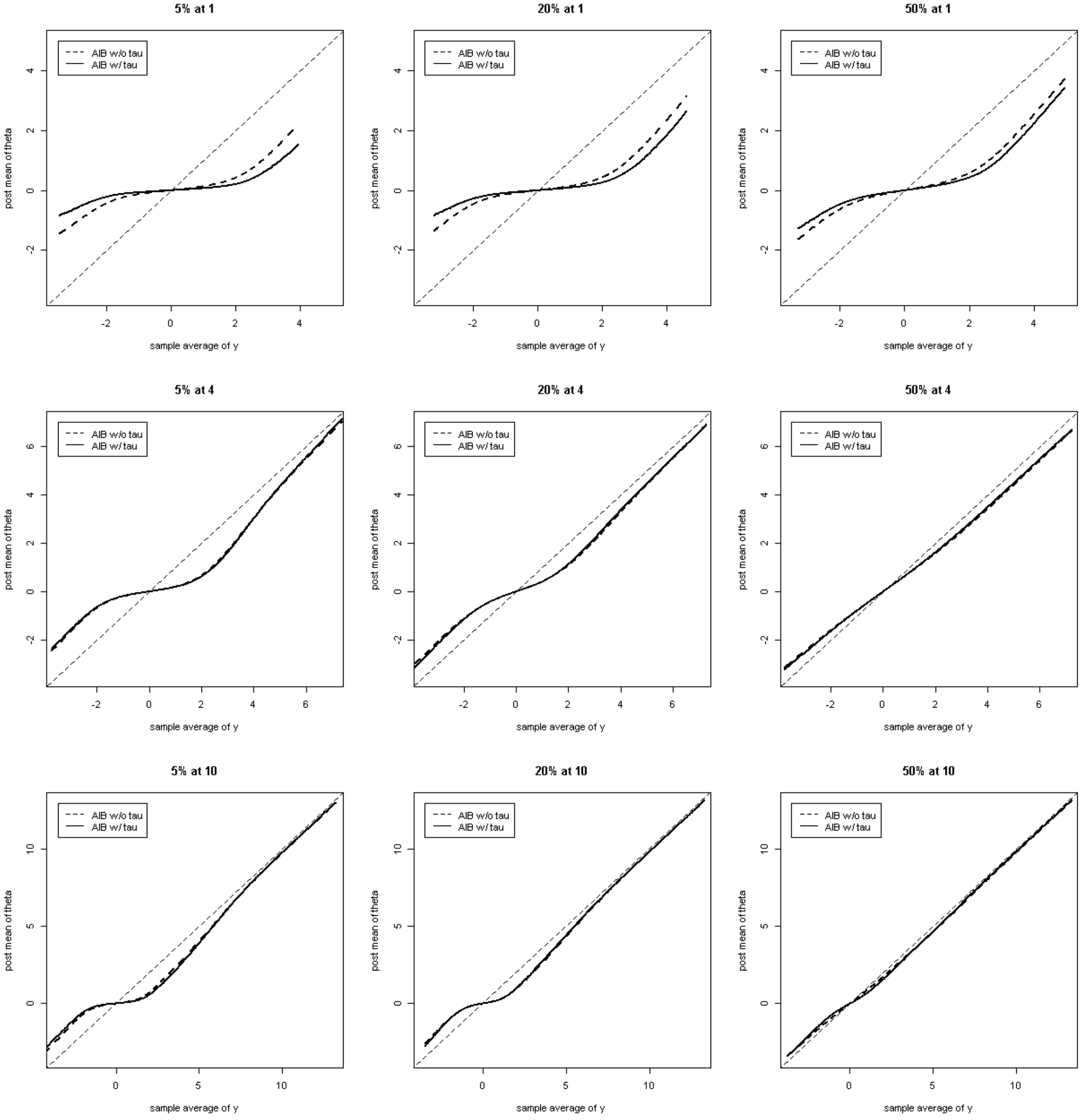

Figure 2.6: Shrinkage performance comparison between the AIB without τ model and the AIB with τ model under the IID assumptions in different scenarios. All graphs plot the posterior means of θ_i's against $\hat{Y}_i$'s using 100 datasets. For all datasets, $p = 100$ and 2 observations are generated for each dimension. The signal size is fixed at 1 in the first row, at 4 in the second row and at 10 in the third row. The nonzero percentage is fixed at 5 in the left column, at 20 in the middle column and at 50 in the right column.

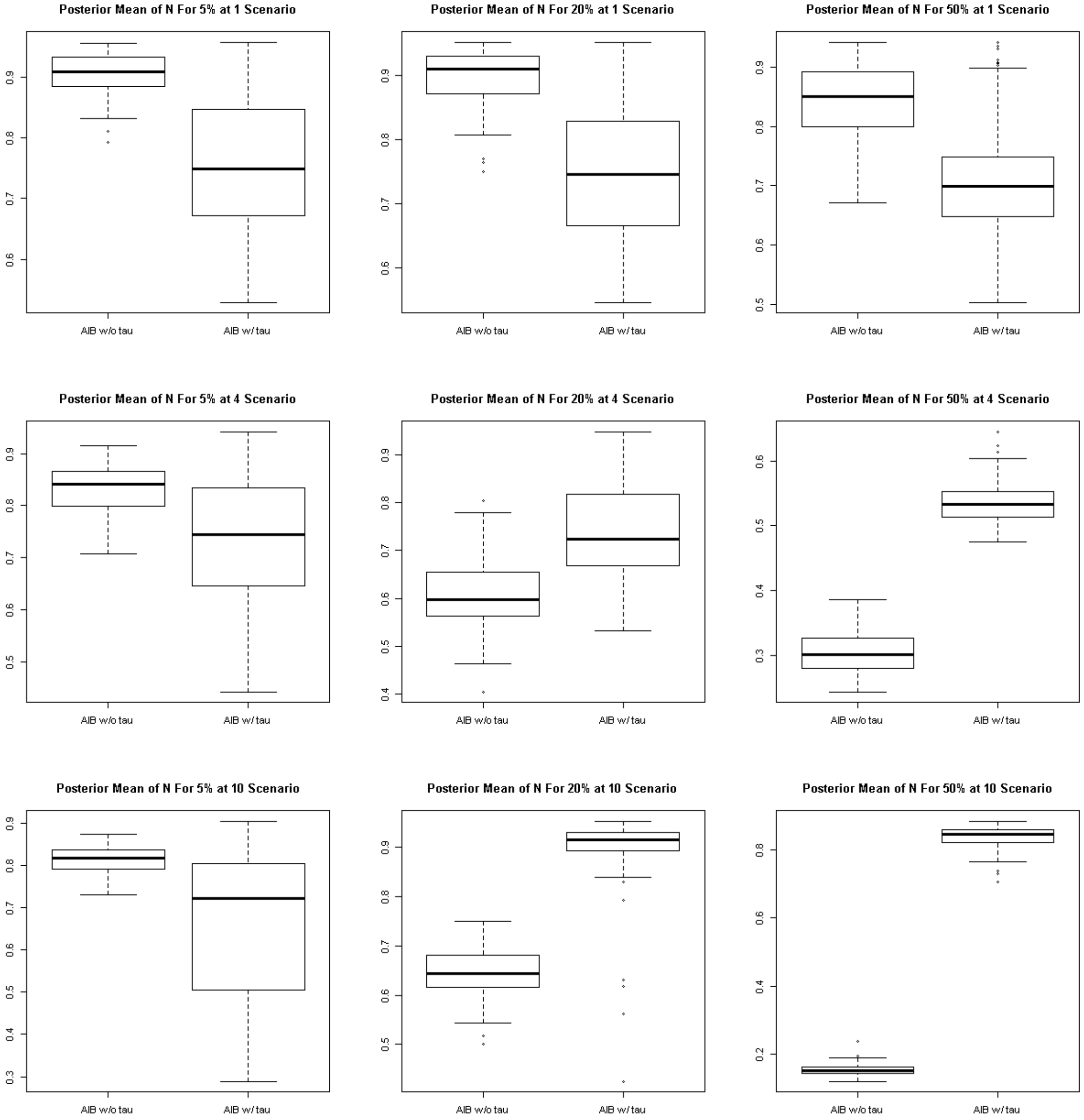

Figure 2.7: Boxplots of posterior means of N in the AIB without τ model and the AIB with τ model under the IID assumptions in different scenarios. All boxes are created with 100 datasets. In each dataset, $p = 100$ and 2 observations are generated for each dimension. The signal size is fixed at 1 in the upper row, 4 in the middle row and 10 in the lower row. The nonzero percentage is fixed at 5 in the left column, 20 in the middle column and 50 in the right column.

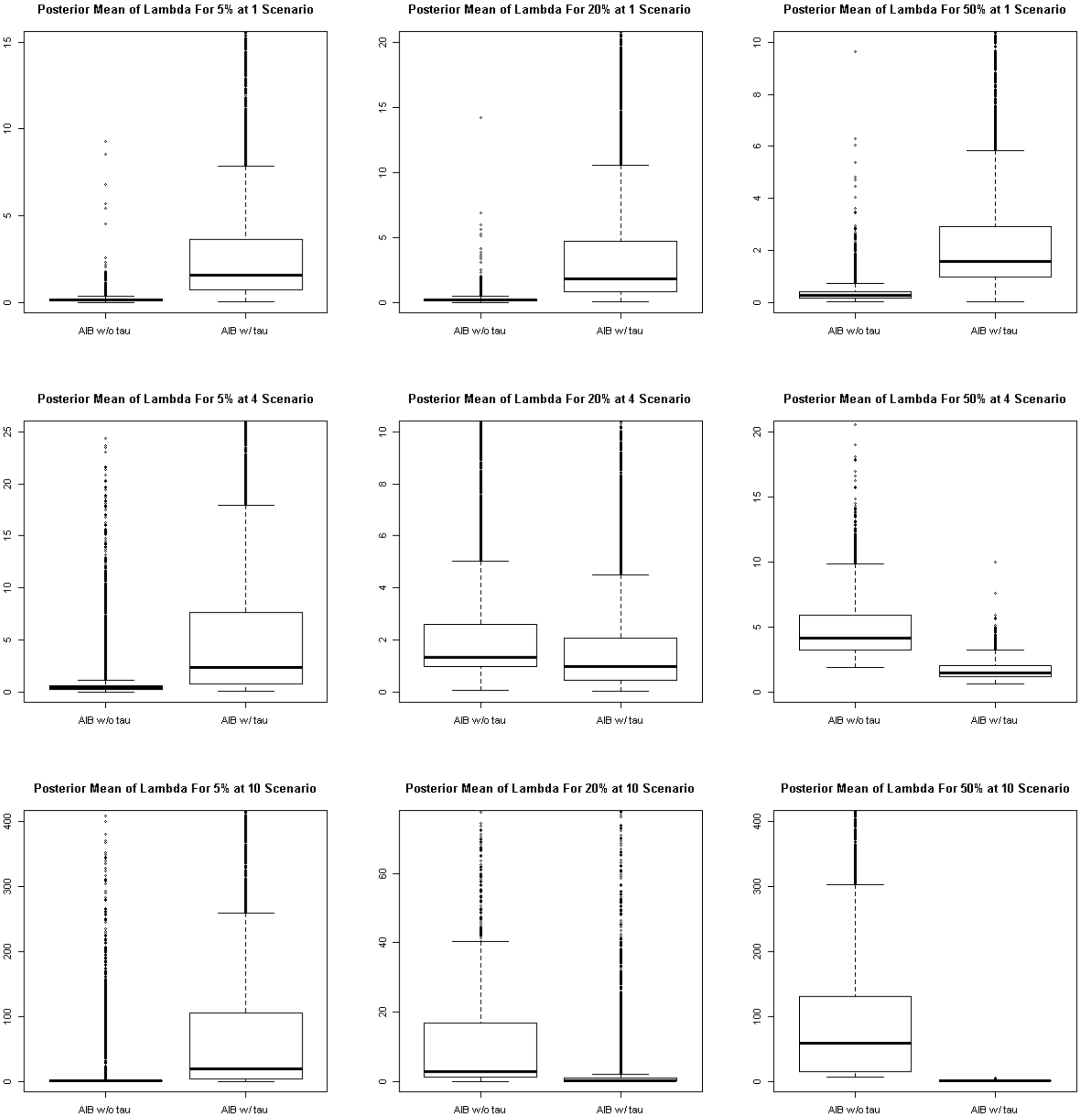

Figure 2.8: Boxplots of posterior means of λ_i's of the AIB without and with τ models under the IID assumptions in different scenarios. All dimensions of $\boldsymbol{\lambda}$ are plotted in one box using 100 datasets for each model. In each dataset, $p = 100$ with 2 observations per dimension. The signal size is fixed at 1 in the upper row, 4 in the middle row and 10 in the lower row. The nonzero percentage is fixed at 5 in the left column, 20 in the middle column and 50 in the right column.

Chapter 3: The Adaptive Inverted-Beta Prior in Normal Linear Regression

3.1 Introduction

In Chapter 2 we introduced the family of adaptive inverted-Beta priors for the normal mean problem with IID observations from

$$\boldsymbol{Y} \mid \boldsymbol{\theta}, \sigma \sim \mathrm{N}(\boldsymbol{\theta}, \sigma^2 I). \tag{3.1}$$

Two natural extensions of this problem are generalizing the mean $\boldsymbol{\theta}$ to a regression form $\boldsymbol{X\beta}$ and generalizing the independent covariance structure to general covariance structures. In this chapter, we focus on the first extension.

The standard multivariate linear regression model can be written as

$$\boldsymbol{Y} = \mu\mathbf{1} + \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \tag{3.2}$$

where $\boldsymbol{Y}$ is the vector of responses, μ is the intercept, $\mathbf{1}$ is the vector of 1's, $\boldsymbol{X}$ is the $n \times p$ design matrix, $\boldsymbol{\beta}$ is the vector of regression coefficients and $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 I)$ is the vector of random errors. An important topic in multivariate regression is the variable selection problem, i.e., choosing relevant predictive variables to the response. Ordinary least squares solutions are standard estimators, but do not exist when $p >$

n. The best subset selection method selects the best model among the 2^p possible models based on a well-defined criterion, for example the model with the largest adjusted R-squared value. However, it suffers from the long computation time and instability as stated in Breiman (1996). Stepwise selection methods are often used as a computational surrogate. We use the forward selection method as an example here. Assume that the response vector $\boldsymbol{Y}$ and the columns of the design matrix $\boldsymbol{X}$ are centralized at their means. For a predetermined subset size k, the forward selection method selects the first variable $\boldsymbol{X}_j$ which minimizes the residual sum of squares

$$S = \sum_{i=1}^{n} (Y_i - b_j X_{i,j})^2,$$

or equivalently maximizes

$$(\sum_{i=1}^{n} X_{i,j} Y_i)^2 / \sum_{i=1}^{n} X_{i,j}^2. \tag{3.3}$$

After the first variable $\boldsymbol{X}_{(1)}$ is selected, it is forced to remain in all further subsets. The next variable is selected among the spaces orthogonal to $\boldsymbol{X}_{(1)}$, so for each variable $\boldsymbol{X}_j$ other than $\boldsymbol{X}_{(1)}$, the candidate can be written as

$$\boldsymbol{X}_{j(1)} = \boldsymbol{X}_j - b_{j(1)} \boldsymbol{X}_{(1)},$$

where $b_{j(1)}$ is the regression coefficient of $\boldsymbol{X}_j$ upon $\boldsymbol{X}_1$. A variable $\boldsymbol{X}_j$ is selected if $\boldsymbol{X}_{j(1)}$ satisfies equation (3.3) with Y_i replaced by $Y_i - b_{(1)} X_{i,(1)}$, and $X_{i,j}$ replaced by $X_{i,j(1)}$. Repeating the above procedures yields a series of selected variables $\boldsymbol{X}_{(1)}, \cdots,$ $\boldsymbol{X}_{(k)}$. The subset size k can be determined by Mallows' C_p criterion. Efroymson's algorithm and the backward elimination method are another two stepwise selection strategies. More details are given in Miller (2002). However, the stepwise selection methods do not explore the whole model space, and thus the final selected model is usually not optimal under any criterion. In contrast, the Lasso is a continuous (in the tuning parameter λ) and stable process for the variable selection problem, but it also has its own disadvantages as stated in Chapter 1.

Bayesian variable selection methods involve placing prior distributions on the unknown parameters. One important method is the "spike-and-slab" prior in Mitchell and Beauchamp (1988) which is studied and developed by George and McCulloch (1993), Chipman (1996) and Clyde, DeSimone and Parmigiani (1996), etc. Another important method is the g-prior in Zellner (1986). Under model (3.2), let $\boldsymbol{\gamma}$ denote a p-dimensional indicator vector where $\gamma_i = 1$ indicates that variable $\boldsymbol{X}_i$ is included and $\gamma_i = 0$ indicates that $\boldsymbol{X}_i$ is excluded. The posterior probability of model $M_{\boldsymbol{\gamma}}$ is defined by

$$p(M_{\boldsymbol{\gamma}} \mid \boldsymbol{Y}) = \frac{p(M_{\boldsymbol{\gamma}})p(\boldsymbol{Y} \mid M_{\boldsymbol{\gamma}})}{\sum_{\boldsymbol{\gamma}'} p(M_{\boldsymbol{\gamma}'})p(\boldsymbol{Y} \mid M_{\boldsymbol{\gamma}'})},$$

where $p(M_{\boldsymbol{\gamma}})$ is the prior probability of model $M_{\boldsymbol{\gamma}}$ and $p(\boldsymbol{Y} \mid M_{\boldsymbol{\gamma}})$ is the marginal likelihood of the data under model $M_{\boldsymbol{\gamma}}$:

$$p(\boldsymbol{Y} \mid M_{\boldsymbol{\gamma}}) = \int_{\boldsymbol{\Theta}_{\boldsymbol{\gamma}}} p(\boldsymbol{Y} \mid \boldsymbol{\theta}_{\boldsymbol{\gamma}}, M_{\boldsymbol{\gamma}})p(\boldsymbol{\theta}_{\boldsymbol{\gamma}} \mid M_{\boldsymbol{\gamma}})\mathrm{d}\boldsymbol{\theta}_{\boldsymbol{\gamma}},$$

where $\boldsymbol{\Theta}_{\boldsymbol{\gamma}} = (\mu, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma)$. Using the Bayes factor for pairs of hypotheses, the posterior probability can also be expressed as

$$p(M_{\boldsymbol{\gamma}} \mid \boldsymbol{Y}) = \frac{p(M_{\boldsymbol{\gamma}})\mathrm{BF}(M_{\boldsymbol{\gamma}} : M_b)}{\sum_{\boldsymbol{\gamma}'} p(M_{\boldsymbol{\gamma}'})\mathrm{BF}(M_{\boldsymbol{\gamma}'} : M_b)}, \tag{3.4}$$

where $\mathrm{BF}(M_{\boldsymbol{\gamma}} : M_b)$ represents the Bayes factor for comparing model $M_{\boldsymbol{\gamma}}$ to a base model M_b. Under model $M_{\boldsymbol{\gamma}}$, Zellner's g-prior can be written as

$$\begin{aligned} p(\mu, \sigma^2) &\propto \frac{1}{\sigma^2}, \\ \boldsymbol{\beta} \mid \sigma &\sim \mathrm{N}(\boldsymbol{\beta}_0, g\sigma(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}), \end{aligned}$$

where $\boldsymbol{\beta}_0$ is the anticipated value of $\boldsymbol{\beta}$, g is the hyperparameter to be determined and $\boldsymbol{X}_\gamma$ is the $n \times p_\gamma$ design matrix. This prior is widely adopted because of its computational efficiency in calculating the marginal likelihood and its simple interpretation as the connection with the design matrix $\boldsymbol{X}_\gamma$. The hyperparameter g can be predetermined by many model selection criteria, such as the Akaike information criterion (AIC), the Bayesian information criterion (BIC) (George and Foster 2000), or using empirical Bayes methods (George and Foster 2000, Clyde and George 2000, Hansen and Yu 2001).

Liang, Paulo, Molina, Clyde and Berger (2008) explored fully Bayes approaches of using g-priors and raised the mixture of g-priors by putting a prior distribution on g. The authors provided two examples. In the first example, they rewrote the Cauchy prior from Zellner and Siow (1980) as:

$$\begin{aligned}\boldsymbol{\beta} \mid \sigma &\sim \mathrm{N}(\boldsymbol{\beta}_0, g\sigma(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1}),\\ \pi(g) &= \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} e^{-n/(2g)}, \end{aligned} \tag{3.5}$$

and developed a new approximation to Bayes factors having simple and tractable expressions to the posterior model probabilities. In the second example, they raised another prior on g, the hyper-g priors:

$$\pi(g) = \frac{a-2}{2}(1+g)^{-a/2}, \quad g > 0, \tag{3.6}$$

which is proper when $a > 2$. This family of priors leads to marginal likelihoods with closed forms in terms of the Gaussian hypergeometric function. Both examples provide computational efficiency and adaptive and nonlinear shrinkage effects.

Assuming that the regression coefficients $\boldsymbol{\beta}$ is sparse, many Bayesian variable selection methods are developed under the prior distributions introduced in Chapter 2, such as the Normal-Jeffreys prior (Figueiredo 2003), the Normal-Exponential-Gamma prior (Griffin and Brown (2005), Griffin and Brown (2007)) and the Normal-Gamma prior (Griffin and Brown (2010)). In this chapter, we implement the AIB prior to normal linear regression models. We show that it can successfully select important predictors and estimate the regression coefficients simultaneously.

The structure of this chapter is as follows: Section 3.2 sets up the regression model under the AIB prior and develops the computational algorithm for the $p > n$ case. Section 3.3 provides simulation studies for the $n > p$ case and the $p > n$ case respectively. Section 3.4 implements our method to the NIR spectroscopy data.

3.2 Model Structure And Computational Details

3.2.1 When $n > p$

For simplicity in this section we assume that the response variable $\boldsymbol{Y}$ and the columns of the design matrix $\boldsymbol{X}$ are centralized at their means. Suppose that the regression response $\boldsymbol{Y}$ satisfies

$$\boldsymbol{Y} \mid \boldsymbol{\beta}, \sigma \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I). \tag{3.7}$$

When $n > p$ it is easy to derive the full conditional posterior distribution of $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} \mid \boldsymbol{Y}, \boldsymbol{X}, \sigma, \tau, \boldsymbol{\lambda} \sim N(\boldsymbol{\alpha}, \boldsymbol{\Sigma}), \tag{3.8}$$

where

$$\boldsymbol{\alpha} = \tau^2 \text{Diag}(\boldsymbol{\lambda})(\tau^2 \text{Diag}(\boldsymbol{\lambda}) + (\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}\hat{\boldsymbol{\beta}}_{OLS}, \tag{3.9}$$

$$\boldsymbol{\Sigma} = \sigma^2\tau^2 \text{Diag}(\boldsymbol{\lambda})(\tau^2 \text{Diag}(\boldsymbol{\lambda}) + (\boldsymbol{X}^T\boldsymbol{X})^{-1})^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}, \tag{3.10}$$

and $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ is the ordinary least squares estimate following a multivariate normal distribution with mean $\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Therefore, using similar steps as in Section 2.3 we can easily update the parameters with the MCMC algorithms.

3.2.2 When $p > n$

When $p > n$, the ordinary least squares estimator does not exist, and we derive the conditional posterior distribution of $\boldsymbol{\beta}$ using the singular value decomposition method (West 2003). Assuming $\boldsymbol{X}_{r\times p}$ is full rank, we can express $\boldsymbol{X}$ as

$$\boldsymbol{X} = U^T D V^T, \tag{3.11}$$

where U is an $n \times n$ matrix such that $U^T U = UU^T = I_{n\times n}$, D is an $n \times n$ diagonal matrix and V is a $p \times n$ matrix such that $V^T V = I_{n\times n}$. Using this decomposition, $\boldsymbol{X\beta}$ can be rewritten as

$$\boldsymbol{X\beta} = U^T D\boldsymbol{\gamma}, \tag{3.12}$$

where $\boldsymbol{\gamma} = V^T\boldsymbol{\beta}$. Now the ordinary least squares estimate of $\boldsymbol{\gamma}$ exists, which is

$$\hat{\boldsymbol{\gamma}} = ((U^T D)^T(U^T D))^{-1}(U^T D)^T\boldsymbol{Y} = D^{-1}U\boldsymbol{Y}. \tag{3.13}$$

Next let $\boldsymbol{\beta} \sim \mathrm{N}(0, \Phi)$, and define $\Phi_0 = V^T\Phi V$ and $\Lambda = \sigma^2 D^{-2}$, then the conditional distributions of $\hat{\boldsymbol{\gamma}}$ given $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}$ given Φ are

$$\hat{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma} \sim N(\boldsymbol{\gamma}, \Lambda), \tag{3.14}$$

$$\boldsymbol{\gamma} \mid \Phi \sim N(\boldsymbol{0}, \Phi_0) \tag{3.15}$$

respectively. Therefore the posterior distribution of $\boldsymbol{\gamma}$ given $\hat{\boldsymbol{\gamma}}$, Φ is

$$\boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi \sim N(\Phi_0(\Phi_0 + \Lambda)^{-1}\hat{\boldsymbol{\gamma}}, (\Lambda^{-1} + \Phi_0^{-1})^{-1}). \tag{3.16}$$

To get the posterior distribution of $\boldsymbol{\beta}$, consider the full singular value decomposition of $\boldsymbol{X}$:

$$\boldsymbol{X} = U^T D^* V^{*T}, \tag{3.17}$$

where D^* is an $n \times p$ matrix with the first r columns same as D and $\mathbf{0}$ for the last $(p-n)$ columns, V^* is a $p \times p$ matrix with the first n columns same as V and the last $(p-n)$ columns such that $V^{*T}V^* = V^*V^{*T} = I_{p\times p}$. For simplicity we write V^* as $V^* = (V \;\; C)$. With the new decomposition, define $\boldsymbol{\gamma}^* = V^{*T}\boldsymbol{\beta}$. Note that the first n elements of $\boldsymbol{\gamma}^*$ are exactly same as $\boldsymbol{\gamma}$. We call the last $(p-n)$ elements of $\boldsymbol{\gamma}^*$, $\boldsymbol{\epsilon}$, which represents those dimensions gaining no information from data. So $\boldsymbol{\gamma}^*$ can be written as

$$\boldsymbol{\gamma}^* = (\boldsymbol{\gamma}^T \quad \boldsymbol{\epsilon}^T)^T. \tag{3.18}$$

Now $\boldsymbol{\beta} = V^*\boldsymbol{\gamma}^*$, and the posterior distribution of $\boldsymbol{\gamma}^*$ given $\hat{\boldsymbol{\gamma}}$ and Φ can be derived as:

$$\begin{aligned} \pi(\boldsymbol{\gamma}^* \mid \hat{\boldsymbol{\gamma}}, \Phi) &= \pi((\boldsymbol{\gamma}, \epsilon) \mid \hat{\boldsymbol{\gamma}}, \Phi) \\ &= \pi((\boldsymbol{\gamma}, \epsilon) \mid \boldsymbol{\gamma}, \Phi)\pi(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi) \\ &= \pi(\epsilon \mid \boldsymbol{\gamma}, \Phi)\pi(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi), \end{aligned} \tag{3.19}$$

where

$$\boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi \sim N(\Phi_0(\Phi_0 + \Lambda)^{-1}\hat{\boldsymbol{\gamma}}, (\Lambda^{-1} + \Phi_0^{-1})^{-1}),$$

$$\boldsymbol{\epsilon} \mid \boldsymbol{\gamma}, \Phi \sim N(C^T\Phi V\Phi_0^{-1}\boldsymbol{\gamma}, C^T\Phi C - C^T\Phi V\Phi_0^{-1}V^T\Phi C).$$

The normality of $\pi(\boldsymbol{\epsilon} \mid \boldsymbol{\gamma}, \Phi)$ and $\pi(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi)$ combined with the fact that the mean of $\pi(\boldsymbol{\epsilon} \mid \boldsymbol{\gamma}, \Phi)$ is a linear function of $\boldsymbol{\gamma}$ imply the normality of $\pi(\boldsymbol{\gamma}^* \mid \hat{\boldsymbol{\gamma}}, \Phi)$ and $\pi(\boldsymbol{\beta} \mid \hat{\boldsymbol{\gamma}}, \Phi)$. To get the conditional posterior mean and covariance of $(\boldsymbol{\beta} \mid \hat{\boldsymbol{\gamma}}, \Phi)$, we get these for $(\boldsymbol{\gamma}^* \mid \hat{\boldsymbol{\gamma}}, \Phi)$ first. For the posterior mean, we have

$$\begin{aligned} E(\boldsymbol{\gamma}^* \mid \hat{\boldsymbol{\gamma}}, \Phi) &= \begin{pmatrix} E(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi) \\ E(\epsilon \mid \hat{\boldsymbol{\gamma}}, \Phi) \end{pmatrix} \\ &= \begin{pmatrix} E(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi) \\ C^T \Phi V \Phi_0^{-1} E(\boldsymbol{\gamma} \mid \hat{\gamma}, \Phi) \end{pmatrix}. \qquad (3.20) \end{aligned}$$

For the posterior covariance, we have

$$\begin{aligned} & Var(\boldsymbol{\gamma}^* \mid \hat{\boldsymbol{\gamma}}, \Phi) \\ = \;& Var(E(\boldsymbol{\gamma}^* \mid \boldsymbol{\gamma}) \mid \hat{\boldsymbol{\gamma}}, \Phi) + E(Var(\boldsymbol{\gamma}^* \mid \boldsymbol{\gamma}) \mid \hat{\boldsymbol{\gamma}}, \Phi) \\ = \;& Var \begin{pmatrix} \boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi \\ C^T \Phi V \Phi_0^{-1} \boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi \end{pmatrix} + E \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & C^T \Phi C - C^T \Phi V \Phi_0^{-1} V^T \Phi C \end{pmatrix}, \qquad (3.21) \end{aligned}$$

where

$$Var \begin{pmatrix} \boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi \\ C^T \Phi V \Phi_0^{-1} \boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi \end{pmatrix} = \begin{pmatrix} V_{\boldsymbol{\gamma}} & V_{\boldsymbol{\gamma}}^T A^T \\ A V_{\boldsymbol{\gamma}} & A V_{\boldsymbol{\gamma}} A^T \end{pmatrix}, \qquad (3.22)$$

and $V_{\boldsymbol{\gamma}} = Var(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi)$, $A = C^T \Phi V \Phi_0^{-1}$.

Let $B = C^T \Phi C - C^T \Phi V \Phi_0^{-1} V^T \Phi C$, after combining the two terms in equation (3.21), we obtain

$$Var(\boldsymbol{\gamma}^* \mid \hat{\boldsymbol{\gamma}}, \Phi) = \begin{pmatrix} V_{\boldsymbol{\gamma}} & V_{\boldsymbol{\gamma}}^T A^T \\ A V_{\boldsymbol{\gamma}} & A V_{\boldsymbol{\gamma}} A^T + B \end{pmatrix}. \qquad (3.23)$$

Applying the equation $\boldsymbol{\beta} = V^* \boldsymbol{\gamma}^*$, we can write the posterior mean and covariance of $\boldsymbol{\beta}$ given $\hat{\boldsymbol{\gamma}}$ and Φ as

$$E(\boldsymbol{\beta} \mid \hat{\boldsymbol{\gamma}}, \Phi) = V E(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi) + C C^T \Phi V \Phi_0^{-1} E(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi), \qquad (3.24)$$

$$Var(\boldsymbol{\beta} \mid \hat{\boldsymbol{\gamma}}, \Phi) = V V_{\boldsymbol{\gamma}} V^T + C A V_{\boldsymbol{\gamma}} V^T + V V_{\boldsymbol{\gamma}}^T A^T C^T + C A V_{\boldsymbol{\gamma}} A^T C^T. \qquad (3.25)$$

Since $V^{*T}V^* = V^*V^{*T} = I$, it is easy to see that $CC^T = I - VV^T$. Plugging it into (3.24) and (3.25), we obtain

$$\begin{aligned} E(\boldsymbol{\beta} \mid \hat{\boldsymbol{\gamma}}, \Phi) &= VE(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi) + \Phi V \Phi_0^{-1} E(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi) - VE(\boldsymbol{\gamma} \mid \hat{\boldsymbol{\gamma}}, \Phi) \\ &= \Phi V (\Phi_0 + \Lambda)^{-1} \hat{\boldsymbol{\gamma}}, \qquad (3.26) \\ Var(\boldsymbol{\beta} \mid \hat{\boldsymbol{\gamma}}, \Phi) &= \Phi - \Phi V \Phi_0^{-1} V^T \Phi + \Phi V \Phi_0^{-1} V_{\boldsymbol{\gamma}} \Phi_0^{-1} V^T \Phi \\ &= \Phi - \Phi V (\Phi_0^{-1} - \Phi_0^{-1} V_{\boldsymbol{\gamma}} \Phi_0^{-1}) V^T \Phi. \qquad (3.27) \end{aligned}$$

Using the Taylor expansion of $V_{\boldsymbol{\gamma}}$:

$$(\Phi_0^{-1} + \Lambda^{-1})^{-1} = \Phi_0 - \Phi_0 \Lambda^{-1} \Phi_0 + \Phi_0 \Lambda^{-1} \Phi_0 \Lambda^{-1} \Phi_0 - \cdots ,$$

we get

$$\begin{aligned} \Phi_0^{-1} - \Phi_0^{-1} V_{\boldsymbol{\gamma}} \Phi_0^{-1} &= \Lambda^{-1} - \Lambda^{-1} \Phi_0 \Lambda^{-1} + \Lambda^{-1} \Phi_0 \Lambda^{-1} \Phi_0 \Lambda^{-1} - \cdots \\ &= (\Phi_0 + \Lambda)^{-1}. \qquad (3.28) \end{aligned}$$

Therefore,

$$Var(\boldsymbol{\beta} \mid \hat{\boldsymbol{\gamma}}, \Phi) = \Phi - \Phi V (\Phi_0 + \Lambda)^{-1} V^T \Phi. \qquad (3.29)$$

3.3 Simulation Studies

3.3.1 Scenario 1 – when $n > p$

In this scenario, we generate $\boldsymbol{\beta}$ from the following fixed nonzero value model:

$$\beta_i \mid w, \delta_0, \delta_1 = \begin{cases} \delta_0, & \text{with probability } 1 - w, \\ \delta_1, & \text{with probability } w, \end{cases} \qquad (3.30)$$

where $\delta_0 = 0$ and δ_1 is the fixed nonzero value. Considering different sparsity levels and signal sizes, we choose the combinations of three nonzero values (δ_1): 1, 4, 10, and three nonzero percentages (w): 5%, 20% and 50%. The rows of the design matrix

$\boldsymbol{X}_{n\times p}$ are identically and independently generated from $\mathrm{N}(\mathbf{0}, 1/(n-p)I)$, then the columns of $\boldsymbol{X}$ are centered. The response vector $\boldsymbol{Y}$ is generated from $\mathrm{N}(\boldsymbol{X\beta}, \sigma^2 I)$ with $\sigma = 1$ and then centered. The scalar $1/(n-p)$ is selected to make the variance of the least squares estimate $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ not too small. For each combination, $p = 50$, $n = 100$ and 100 datasets are generated with 40000 iteration total and a burn-in period of 20000. The estimation performances are evaluated using the sum of squared loss function:

$$\text{Loss} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

We compare the averaged risks of the AIB prior with those of the horseshoe prior (HS), the Strawderman–Berger prior (SB), the NEG(2,1) prior (NEG), and the double-exponential prior (DE). Similar to the first simulation study in Chapter 2, we let all models share one common structure:

$$\begin{aligned} \boldsymbol{\beta} \mid \boldsymbol{\lambda}, \sigma^2, \tau &\sim \mathrm{N}(\mathbf{0}, \sigma^2\tau^2\mathrm{Diag}(\boldsymbol{\lambda})), \\ p(\sigma^2) &\propto \frac{1}{\sigma^2}, \\ \tau &\sim \mathrm{C}^+(0,1), \end{aligned} \tag{3.31}$$

and let the priors on λ_i be $\mathrm{IB}(MN, M(1-N))$, $\mathrm{IB}(1/2,1/2)$, $\mathrm{IB}(1/2,1)$, $\mathrm{IB}(2,1)$ and the standard double-exponential distribution, respectively. The averaged losses across 100 datasets are reported in the upper rows of Table 3.1, and the standard deviations of the averaged losses are reported in the lower rows. The estimates of hyperparameters and σ^2 along with the averaged posterior standard deviations are reported in Table 3.2 and Table 3.3.

Generally speaking, the simulation results are very similar to the those in Chapter 2. All models are consistently better than the MLE. Fixing the sparsity level, the

Scenario		AIB	HS	SB	NEG	DE	MLE
1	5%	2.19	2.32	2.32	2.21	2.22	51.84
		0.071	0.106	0.098	0.071	0.067	1.550
	20%	8.49	8.73	8.76	8.44	8.42	52.71
		0.148	0.183	0.186	0.143	0.138	1.551
	50%	16.19	17.81	17.64	15.90	15.42	52.93
		0.296	0.293	0.298	0.293	0.291	1.229
4	5%	4.57	3.52	3.58	6.22	10.19	54.51
		0.327	0.276	0.279	0.334	0.349	1.566
	20%	17.61	15.35	16.29	21.76	23.95	52.35
		0.781	0.687	0.710	0.721	0.734	1.420
	50%	44.59	44.03	45.44	44.73	44.22	53.11
		1.334	1.663	1.542	1.383	1.320	1.744
10	5%	2.61	2.78	2.84	6.51	14.65	55.85
		0.215	0.204	0.215	0.335	0.533	1.720
	20%	11.68	13.44	13.97	26.04	31.33	52.84
		0.481	0.495	0.511	0.751	0.868	1.441
	50%	33.89	34.17	38.93	43.86	44.99	52.75
		1.155	0.934	1.057	1.187	1.214	1.514

Table 3.1: Comparison of losses among the AIB, the HS, the SB, the NEG and the DE models in normal linear regressions when $n > p$. The losses of the MLE estimate are given as benchmarks. The averaged squared error losses across 100 datasets are reported in the upper rows. The standard deviations of the average losses are reported in the lower rows. For all datasets, $p = 50$ and $n = 100$. The signal sizes are listed in the first column and the nonzero percentages are in the second column.

AIB, the HS and the SB priors have the greatest losses when the signal size is 4, the NEG and the DE priors have the greatest losses when the signal size is 10. Again, this is because that when the signal is 4, the observations of noises and signals are partially mixed. Incorrect discriminations between the noises and signals cause great losses. The NEG prior and the DE prior suffer from the over shrinkage issue, and this issue is more severe when the signal size is large. Fixing the signal size, all methods have the greater losses when the sparsity decreases.

When the signal size is 1, the noises and signals are hardly distinguishable, so incorrect discriminations may not be fatal. All methods are good when the nonzero percentage is 5 or 20. The AIB prior, the NEG prior and the DE prior work better than the HS and the SB priors when the nonzero percentage is 50. As the signal size increases to 4, the AIB prior is worse than the HS and the SB priors, but still better than the NEG and the DE priors, especially when the sparsity is high. When the nonzero percentage is 50, all methods have similar losses. When the signal is 10, the over shrinkage issue of the NEG and the DE priors become more obvious. The AIB prior works consistently best across all sparsity levels. The HS and the SB priors are worse than the AIB prior but better than the NEG and the DE priors.

The estimates of τ in the HS, the SB, the NEG and the DE models show a clear increasing pattern when the signal size or the nonzero percentage increases. In most scenarios, the NEG and the DE priors have similar estimates of τ which are consistently larger than those of the HS and the SB priors. Again, this is because that the local shrinkage parameters λ_i in the NEG and the DE priors offer more shrinkage strength than those in the HS and the SB priors. The estimates of τ from the AIB prior have the similar changing pattern but larger posterior variances in most scenarios. This is because that N is not fixed in the AIB prior which offers more variability to τ. Similarly, the estimates of N have large posterior standard deviations which supports the great variability in the AIB prior. Compared to the simulation study for IID observations, extending the normal mean problem to normal linear regressions involves more variability, and the estimates of hyperparameters have greater posterior variances.

Scenario		M	N	τ(AIB)	τ(HS)	τ(SB)	τ(NEG)	τ(DE)
1	5%	2.887	0.675	0.211	0.012	0.005	0.142	0.121
		1.425	0.139	0.145	0.050	0.019	0.272	0.201
	20%	2.930	0.661	0.459	0.037	0.015	0.340	0.293
		1.521	0.187	1.034	0.083	0.035	0.461	0.323
	50%	3.227	0.630	0.870	0.090	0.038	0.728	0.612
		1.637	0.162	0.878	0.068	0.031	0.379	0.313
4	5%	1.877	0.601	0.629	0.047	0.019	0.661	0.738
		0.953	0.195	0.450	0.057	0.024	0.511	0.512
	20%	2.118	0.639	8.035	0.690	0.264	4.462	3.727
		1.160	0.200	10.027	0.472	0.173	2.269	1.591
	50%	4.741	0.524	11.774	4.965	1.939	16.094	10.868
		1.826	0.133	8.631	2.851	1.066	6.853	4.097
10	5%	1.329	0.491	0.211	0.056	0.022	0.989	2.327
		0.656	0.177	1.066	0.038	0.014	0.457	0.764
	20%	1.033	0.501	12.460	0.998	0.336	14.400	15.440
		0.431	0.163	4.736	0.708	0.232	5.385	4.701
	50%	2.626	0.754	348.044	24.943	8.354	90.676	61.875
		1.739	0.189	213.849	13.314	4.637	36.095	21.505

Table 3.2: Comparison of hyperparameters among the AIB, the HS, the SB, the NEG and the DE models in normal linear regressions when $n > p$. The averaged posterior means across 100 datasets are reported in the upper rows, and the averaged posterior standard deviations are reported in the lower rows. For all datasets, $p = 50$ and $n = 100$. The signal sizes are listed in the first column and the nonzero percentages are in the second column.

The estimates of σ^2 are similarly well for all methods across all combinations. For a fixed signal size, the estimates have a general increasing pattern as the sparsity level decreases. For a fixed sparsity level, a general decreasing pattern can be detected. However, compared to the results in Chapter 2, these patterns are less consistent. Again, this is because that the general design matrix introduces more variability.

Scenario		σ^2(AIB)	σ^2(HS)	σ^2(SB)	σ^2(NEG)	σ^2(DE)
1	5%	0.988	0.989	0.988	0.983	0.982
		0.142	0.143	0.147	0.151	0.145
	20%	1.027	1.040	1.036	1.018	1.010
		0.178	0.175	0.187	0.171	0.165
	50%	1.088	1.144	1.133	1.075	1.052
		0.202	0.198	0.194	0.187	0.191
4	5%	0.944	0.947	0.945	0.938	0.949
		0.128	0.130	0.127	0.133	0.135
	20%	0.928	0.919	0.914	0.944	0.948
		0.234	0.209	0.209	0.239	0.238
	50%	1.033	1.052	1.040	1.036	1.027
		0.204	0.213	0.212	0.203	0.216
10	5%	0.928	0.930	0.928	0.893	0.894
		0.149	0.145	0.149	0.153	0.172
	20%	0.929	0.922	0.914	0.963	0.987
		0.164	0.162	0.164	0.189	0.192
	50%	1.008	1.000	1.017	1.045	1.051
		0.201	0.192	0.197	0.214	0.197

Table 3.3: Comparison of σ^2 among the AIB, the HS, the SB, the NEG and the DE models under regression setups when $n > p$. The averaged posterior means across 100 datasets are reported in the upper rows. The averaged posterior standard deviations are reported in the lower rows. For all datasets, $p = 50$ and $n = 100$. The signal sizes are listed in the first column and the nonzero percentages are in the second column.

3.3.2 Scenario 2 – when $p > n$

For the true regression coefficients, we generate $\boldsymbol{\gamma}$ in (3.12) from the following fixed nonzero value model:

$$\gamma_i \mid w, \delta_0, \delta_1 = \begin{cases} \delta_0, & \text{with probability } 1 - w, \\ \delta_1, & \text{with probability } w, \end{cases} \tag{3.32}$$

and fix $\boldsymbol{\epsilon}$ in (3.18) with zeros to get $\boldsymbol{\gamma}^* = (\boldsymbol{\gamma}^T \quad \mathbf{0}^T)^T$. For the design matrix $\boldsymbol{X}_{n\times p}$, we first generate each row of $\boldsymbol{X}$ from $\mathrm{N}(\mathbf{0}, (1/n)I)$ independently, and then centralize the columns of $\boldsymbol{X}$. The scalar $1/n$ is aimed to make Λ in (3.14) not too small. With the full singular value decomposition, we get V^* in (3.17) and the true values of $\boldsymbol{\beta}$ from $\boldsymbol{\beta} = V^*\boldsymbol{\gamma}^*$. Next $\boldsymbol{Y}$ is generated from $\mathrm{N}(\boldsymbol{X\beta}, \sigma^2 I_{n\times n})$ with $\sigma = 1$.

We consider the same combinations nonzero values and nonzero percentages as in Scenario 1. For each combination, $p = 100$, $n = 50$ and 100 datasets are generated with 40000 iteration total and a burn-in period of 20000. The shrinkage performances are evaluated by the sum of squared error loss function (3.31).

Again we compare the averaged risks among the AIB, the HS, the SB, the NEG and the DE models. For all models, we let

$$\begin{aligned} \boldsymbol{\gamma} \mid \boldsymbol{\lambda}, \sigma^2 &\sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \mathrm{Diag}(\boldsymbol{\lambda})), \\ p(\sigma^2) &\propto \frac{1}{\sigma^2}, \end{aligned} \tag{3.33}$$

and let the priors on λ_i be the same as those in Scenario 1. Notice that in this scenario, we fix $\tau = 1$ in all models. This is because that when $p > n$, the parameters τ^2 and σ^2 are not jointly identifiable. To see this, consider the singular value decomposition (3.11), then the likelihood function (3.7) can be transformed to

$$\boldsymbol{Y}^* \sim \mathrm{N}(\boldsymbol{\theta}, \sigma^2 I), \tag{3.34}$$

where $\boldsymbol{Y}^* = U\boldsymbol{Y}$, $\boldsymbol{\theta} = D\boldsymbol{\gamma}$. Including τ^2 inside the prior of $\boldsymbol{\gamma}$ implies the prior on $\boldsymbol{\theta}$ being $\mathrm{N}(\boldsymbol{0}, \sigma^2\tau^2\mathrm{Diag}(\boldsymbol{\lambda}^*))$, where $\mathrm{Diag}(\boldsymbol{\lambda}^*) = D\mathrm{Diag}(\boldsymbol{\lambda})D$. After integrating $\boldsymbol{\theta}$ out, we get the distribution of $\boldsymbol{Y}^*$ conditional on σ^2, τ^2 and $\boldsymbol{\lambda}^*$:

$$\boldsymbol{Y}^* \mid \sigma^2, \tau^2, \boldsymbol{\lambda}^* \sim \mathrm{N}(\boldsymbol{0}, \sigma^2(\tau^2\mathrm{Diag}(\boldsymbol{\lambda}^*) + I)). \tag{3.35}$$

Consider the following two situations: In the first one, let $\sigma = 1$, $\tau = \tau_0$ and $\lambda_i^* = \omega$, where τ_0 and ω are two positive numbers. This situation corresponds to the one when σ is accurately estimated and

$$\mathrm{Var}(Y_i^* \mid \sigma^2, \tau^2, \lambda_i^*) = \omega + 1. \tag{3.36}$$

In the second situation, let $\sigma = \epsilon$, $\tau = \tau_0/\epsilon$ and $\lambda_i^* = \omega + 1$, where ϵ is a very small constant. This situation corresponds to the one when σ is underestimated, and

$$\mathrm{Var}(Y_i^* \mid \sigma^2, \tau^2, \lambda_i^*) = \omega + (1 + \epsilon). \tag{3.37}$$

As $\epsilon \to 0$, the variance of Y_i^* conditional on σ^2, τ^2 and λ_i^* converges to $(\omega + 1)$, and thus, σ^2 and τ^2 are not identifiable. In fact, in the literature of using sparse priors on the regression coefficients when there are more variables than observations, τ is usually fixed at 1 (Griffin and Brown 2005, Griffin and Brown 2010).

Again the average losses across 100 datasets are reported in Table 3.4 along with the standard deviations of the averaged losses in the lower rows. The estimates of M, N and σ^2 along with the averaged posterior standard deviations are reported in Table 3.5.

Generally speaking, the losses of all methods increase as the sparsity level decreases or the signal size increases. When the signal is 1, the AIB, the NEG and the DE priors work better than the HS and the SB priors. When the signal is 4, the AIB, the NEG and the DE priors are still better than the HS and the SB priors when the nonzero percentage is 5, but when the percentage increases to 20, the AIB and the HS prior work better than the other three methods. The over shrinkage issue of the NEG and the DE priors becomes severe and results in greater losses for these two models. When the sparsity is low, the SB prior works best, the HS prior works better than the AIB prior, and both models work better than the NEG and the DE prior. A similar pattern can be observed when the signal is 10. The AIB prior and the NEG prior work best for the high sparsity scenario, the AIB and the HS prior work best for the moderate high sparsity scenario and the SB prior works best for the low sparsity scenario.

Scenario		AIB	HS	SB	NEG	DE
1	5%	3.28	21.09	30.59	3.61	3.39
		0.362	1.108	1.335	0.142	0.110
	20%	8.58	20.17	29.24	7.95	7.79
		0.237	0.821	1.057	0.161	0.134
	50%	17.08	22.26	30.22	16.06	15.34
		0.277	0.738	0.969	0.182	0.184
4	5%	10.37	21.02	29.70	8.86	11.14
		0.640	0.898	1.068	0.682	0.692
	20%	29.75	26.48	32.82	46.63	49.77
		1.251	1.257	1.359	1.440	1.440
	50%	97.84	71.77	49.09	150.16	129.85
		2.850	1.356	1.276	1.516	1.654
10	5%	11.57	22.88	32.60	8.36	26.48
		1.023	1.016	1.165	1.197	2.552
	20%	26.37	24.74	32.09	133.98	186.65
		1.464	1.197	1.328	6.080	5.985
	50%	315.72	251.56	92.39	799.90	645.10
		19.966	5.987	3.564	7.117	7.503

Table 3.4: Comparison of losses among the AIB, the HS, the SB, the NEG and the DE models in normal linear regressions when $p > n$. The averaged squared error losses across 100 datasets are reported in the upper rows. The standard deviations of the average losses are reported in the lower rows. For all datasets, $p = 100$ and $n = 51$. The signal sizes are listed in the first column and the nonzero percentages are in the second column.

In Table 3.5, the estimates of N of the AIB prior have a clearly decreasing pattern, indicating less shrinkages are offered from the local shrinkage parameters, as the nonzero percentage or the signal size increases. Again, this shows that the AIB prior allows data to determine the shrinkage degrees.

The estimates of σ^2 of all models are not consistently around the true value which is very different from Scenario 1. The estimates from the SB prior are consistently smaller than those from other models, especially when there are 50% nonzero values

Scenario		M	N	σ^2(AIB)	σ^2(HS)	σ^2(SB)	σ^2(NEG)	σ^2(DE)
1	5%	2.714	0.810	0.759	0.315	0.161	0.546	0.485
		1.364	0.177	0.166	0.048	0.027	0.074	0.063
	20%	2.841	0.772	0.869	0.392	0.203	0.666	0.588
		1.668	0.166	0.293	0.113	0.065	0.137	0.115
	50%	3.361	0.714	1.093	0.565	0.294	0.925	0.800
		1.464	0.177	0.363	0.121	0.064	0.179	0.153
4	5%	1.881	0.712	0.686	0.333	0.172	0.637	0.643
		0.992	0.179	0.192	0.062	0.033	0.116	0.107
	20%	2.018	0.550	0.696	0.567	0.302	1.828	1.877
		0.893	0.155	0.362	0.172	0.102	0.526	0.461
	50%	4.106	0.505	4.493	2.858	1.446	6.505	5.459
		1.985	0.189	2.969	1.130	0.591	1.693	1.370
10	5%	1.155	0.696	0.624	0.330	0.170	0.672	1.118
		0.927	0.202	0.361	0.125	0.063	0.212	0.297
	20%	1.381	0.477	0.367	0.551	0.312	5.444	8.108
		0.993	0.178	0.223	0.156	0.088	1.821	1.882
	50%	3.774	0.412	19.417	12.771	5.227	38.450	31.973
		1.956	0.167	12.541	5.505	2.773	9.721	7.570

Table 3.5: Comparison of M, N and σ^2 among the AIB, the HS, the SB, the NEG and the DE models in normal linear regressions when $p > n$. The averaged posterior means across 100 datasets are reported in the upper rows. The averaged posterior standard deviations are reported in the lower rows. For all datasets, $p = 100$ and $n = 51$. The signal sizes are listed in the first column and the nonzero percentages are in the second column.

at 4 or 10. Connecting Table 3.5 with Table 3.4, we can see that the SB model has outstandingly better performances in these two situations. The inaccurate estimates and the advantages of the SB prior are caused by three factors. First of all, σ^2 plays a global shrinkage parameter role in the prior of $\boldsymbol{\gamma}$. Excluding τ from the prior makes the impact from the prior to the estimates of σ^2 stronger. Secondly, when $p > n$, the small number of observations compared to the number of parameters limits the estimation accuracy of σ^2, and σ^2 and λ_i are barely identifiable. To see this, consider the normal mean problem with one observation per dimension:

$$\begin{aligned} y_i &\overset{ind}{\sim} \mathrm{N}(\theta_i, \sigma^2), \\ \theta_i &\overset{ind}{\sim} \mathrm{N}(0, \sigma^2\lambda_i). \end{aligned}$$

The distribution of y_i given σ^2 and λ_i is $\mathrm{N}(0, \sigma^2(\lambda_i + 1))$. When y_i is large and σ^2 is accurately estimated, the total variance of y_i is approximately $(\omega + 1)$, where ω is the estimate of λ_i and is large. In contrast, when the estimate of σ^2 is ω and the estimate of λ_i is $1/\omega$, the total variance of y_i is also $(\omega+1)$, and thus σ^2 and λ_i are not identifiable. In normal linear regressions, when $p > n$ and the columns of $\boldsymbol{X}$ and the response vector $\boldsymbol{Y}$ are centralized, we use n observations to estimate $n-1$ dimensions of $\boldsymbol{\beta}$ and σ^2. As the dimension increases, the help from the extra observation to the estimation of σ^2 becomes smaller and σ^2 and λ_i are less identifiable. This explains why the estimates of σ^2 are not consistently around the true value. At last, the impact from the prior distribution on λ_i cannot be ignored. The SB prior prefers little shrinkage, so the estimates of λ_i are larger than those of other priors as shown in Figure 3.1. This explains why the SB prior always gives the smallest estimates of σ^2 across all scenarios and works better than other methods when the sparsity is low and the signal size is large.

In addition, the estimates of σ^2 in the HS, the SB, the NEG and the DE models have an increasing pattern as the nonzero percentage or the signal size increases. However, this pattern is not clear in the AIB prior. The estimates in the AIB prior have much larger posterior variances than other methods. Again this is caused by the extra variability when N is not fixed in the AIB prior.

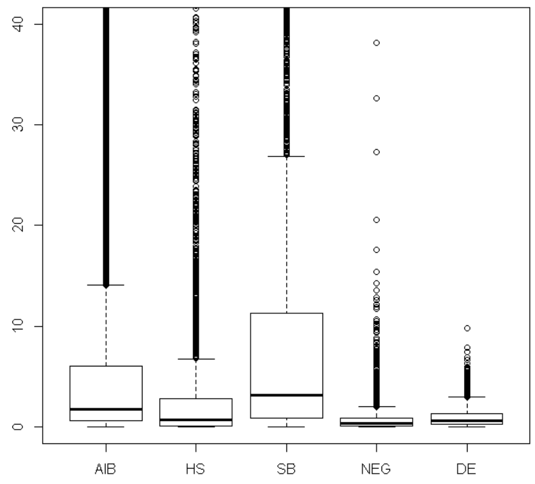

Figure 3.1: Boxplots of posterior draws of one λ_i when the corresponding $\theta_i = 10$. The dataset is selected from the 20% nonzero value at 10 scenario with $p = 100$ and $n = 51$. 40000 iterations are run with a 20000 burn-in period.

3.4 NIR Spectroscopy Data Analysis

In this section, we analyse the NIR spectroscopy data, which is available at http://lib.stat.cmu.edu/datasets/tecator, using the AIB model. The data are records from a Tecator Infratec Food and Feed Analyzer, working in the wavelength range of 850 - 1050 nm, by the Near Infrared Transmission (NIT) principle from the company Tecator. Each observation contains a 100-channel absorbance spectrum along with the contents of moisture, fat and protein of a meat sample. Taking a $-\log 10$ transformation of the original transmittance measured by the spectrometer, gives the absorbance data. The other three components are the percentage of the corresponding content determined by analytic chemistry. The data was originally used by Borggaard and Thodberg (1992), and recently analyzed by Eilers, Li and Marx (2009) and Griffin and Brown (2010). The source data contains 240 observations with the last 25 rows for the purpose of extrapolation test. Following Griffin and Brown (2010), we use the first 215 observations consisting a training dataset with the first 172 observations (all data) and a testing dataset with the last 43 observations, and use the 100 absorbance data to predict the fat content.

In Griffin and Brown (2010), the authors considered both $n > p$ and $p > n$ cases. They used the "all data" as the training dataset for $n > p$ scenario and randomly drew 60 samples from the first 172 observations to construct a $p > n$ dataset. For both scenarios, they drew inferences of the regression coefficients by putting the normal-gamma prior on them and used the posterior means as estimators to predict the fat content using the testing dataset. They compared the roots of mean squared errors (RMSEs) from the normal-gamma prior with those from the Lasso, and found that the normal-gamma prior is much better than the Lasso in both scenarios. When $n > p$, both methods capture the coefficients that are far from 0, but the normal-gamma prior shrinks others strongly to 0 while the Lasso overestimates them. When $p > n$, the normal-gamma prior has similar shrinkage performances, but the Lasso underestimates the large coefficients while overestimating others.

Following Griffin and Brown (2010), we use the "all data" for the $n > p$ scenario, and randomly choose 60 observations from the "all data" to create a $p > n$ dataset. For both scenarios, we use the testing dataset for testing and compare the RMSEs of the AIB prior with those of the HS, the SB, the NEG and the DE models. Again, we fix τ at 1 for all methods when $p > n$. The RMSEs are reported in Table 3.6

along with those in Griffin and Brown (2010) as benchmarks. Since the "small" data is randomly selected, it may be different from that in Griffin and Brown (2010).

	AIB	HS	SB	NEG	DE	NG
All Data	1.97	2.14	2.07	1.92	1.90	1.94
Small	2.54	3.01	2.93	3.25	4.60	2.59

Table 3.6: RMSEs for fat predictions under different models. The "All Data" contains 100 predictors and 172 observations. The "Small" data contains 100 predictors and 60 randomly chosen observations from the training dataset. NG stands for the normal-gamma model in Griffin and Brown (2010) .

Apparently, using "all data" gives much better performances than the "small" data for all methods. When $n > p$, the DE model is the best, the AIB, the NEG and the NG models are a little worse but better than the HS and the SB models. When $p > n$, the AIB and the NG models are better than others. Figure 3.1 plots the posterior means of $\boldsymbol{\beta}$ from the AIB, the NEG and the DE priors. The left graph is for the "all data" and the right one is for the "small" data. From the left one, we can see that all three methods capture the coefficients far from 0. However, the AIB prior shrinks most others to 0 or very close to 0 while the other two methods overestimate them. For the "small" data, the AIB model still captures most large signals while shrinking others close to 0, but the other two methods shrink the large coefficients too much and overestimate others.

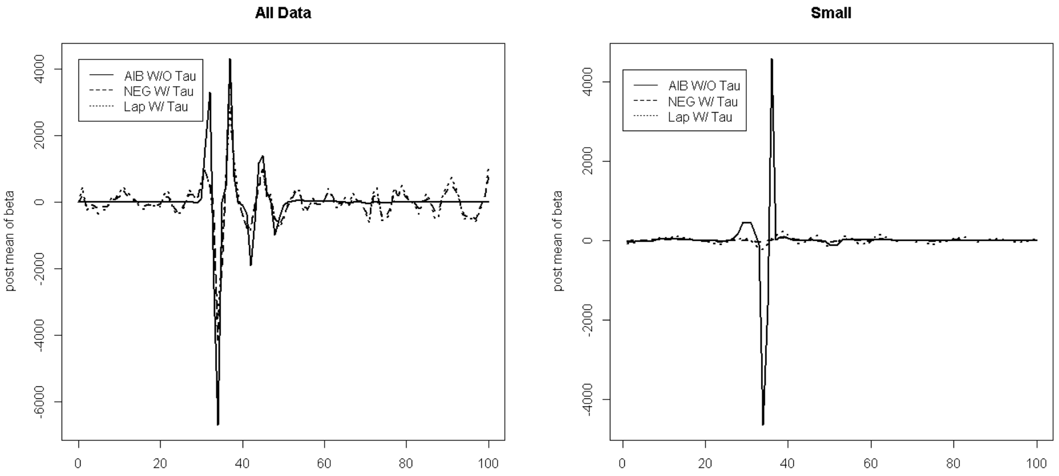

Figure 3.2: Predictor estimation performance comparison among different models. The "All Data" contains 100 predictors and 172 observations, and the "Small" data contains the same 100 predictors but 60 randomly chosen observations from the "All Data".

3.5 Discussion

In this chapter, we implement the AIB prior in normal linear regressions and demonstrate through simulation studies and a data example that the AIB prior is able to accurately estimate the regression coefficients which are far from 0 and shrink others close to 0. When there are more observations than the number of parameters, the AIB prior can yield consistently good performances across different types of data with different sparsity levels and signal sizes. When other priors work well, the AIB prior has similar performances, when they perform normally or poorly, the AIB prior has remarkably improvements. When there are less observations than the number of parameters, we illustrated how to estimate the regression coefficients using the singular value decomposition strategies. In such situation, the AIB prior can still improve the performances of other priors or give similar performances except when the sparsity is low and the signal size is large.

In fact, when $p > n$ and the data appear to be low sparse and have large signals, using empirical Bayes techniques may improve the performances of the AIB prior. For example, assume that the true regression coefficients are sparse. When $\boldsymbol{Y}^* = U\boldsymbol{Y}$ appears to be low sparse and have large signals, instead of putting a Beta(1,1) prior on N, we can modify the prior as Beta(a, b) with a and b such that the distribution puts most densities around $a/(a + b) = n/(\sum Y_i^{*2})$, or simply let $N = n/(\sum Y_i^{*2})$. When the sparsity level is low and the signal size is large, $n/(\sum Y_i^{*2})$ will be small. Similar to the SB prior, using this hyperprior or fixing N at small values forces the prior on λ_i to put more densities on large values which leads to small estimate of σ^2 and consequently may yield better performances.

Moreover, the Bayesian estimator under the AIB priors will not shrink the regression coefficients to exactly 0, but will shrink some to very close to 0. When $n > p$, we can construct a threshoding rule based on the estimates of τ and $\boldsymbol{\lambda}$ for the variable selection purpose in regressions. For example, let

$$\boldsymbol{\kappa} = \hat{\tau}^2 \mathrm{Diag}(\hat{\boldsymbol{\lambda}})(\hat{\tau}^2 \mathrm{Diag}(\hat{\boldsymbol{\lambda}}) + (\boldsymbol{X}^T \boldsymbol{X})^{-1})^{-1}.$$

We can call β_i a signal if $1 - \kappa_i \geq \alpha$ and call β_i a noise otherwise. A natural choice for α would be 0.5 (see, for example, Carvalho et al. (2010)), but more research are needed to find an optimal thresholding rule. For example, α may depend on the sparsity level and the signal size. When the sparsity is high and the signal size is small, α is expected to be greater than 0.5 to express aggressively strong shrinkage. On the other hand, when the sparsity is low and the signal size is large, $\alpha < 0.5$ may be more appropriate to be conservative.

Chapter 4: Conclusion Remarks And Future Work

4.1 Summary

The maximum likelihood estimator (MLE) is a standard estimator for the normal mean estimation problem, but is inadmissible when the parameter's dimension is greater than 2. Shrinking the MLE towards some constant values or subspaces can remarkably improve the performance in high dimensional spaces. In this thesis, we propose a new family of shrinkage priors, the adaptive inverted-Beta priors (AIB), which allows the data to determine the shrinkage degree for each dimension and offers adaptive shrinkage to different dimensions of the parameter. Because the shrinkage power is determined by data, the AIB prior has great flexibility to handle different types of datasets with a large variety of sparsity levels and signal sizes.

The general inverted-Beta prior on λ_i's includes several common shrinkage priors as special cases. These priors are designed for data with certain sparsity properties. We demonstrate through simulation studies and a data example that when these priors work well, the AIB prior can have similar performance, and when these priors work poorly, the AIB prior has the ability to have substantial improvements. Compared to frequetist shrinkage estimators, using the posterior means as estimators automatically guarantees the admissibility under the squared error loss.

The implications of the AIB prior extend beyond the normal mean estimation problem to other settings where shrinkage estimators are valuable. As we showed, the normal linear regression model can be transformed as a normal mean problem by pre-multiplying the response vector by $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ when $n > p$, where $\boldsymbol{X}$ is the design matrix, or using the singular value decomposition strategy when $p > n$.

Moreover, because of the connection between Bayesian approaches and penalized likelihood approaches, the negative logarithm of the AIB prior can be used as a new penalty function.

4.2 Extension of The Adaptive Inverted-Beta Prior for The Normal Mean Problem With General Covariance Structures

4.2.1 Introduction

Another extension of the normal mean estimation problem under IID assumptions is the problem with unknown covariance matrix. Let $\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{\Sigma} \sim \mathrm{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ be a p-dimensional multivariate normal vector with unknown mean $\boldsymbol{\theta}$ and unknown positive definite covariance matrix $\boldsymbol{\Sigma}$. The problem is estimating $\boldsymbol{\theta}$ under loss function (1.1) or other similar invariant loss functions.

As introduced in Chapter 1, Stein (1956) demonstrated that when $\boldsymbol{\Sigma} = I$ and $p \geq 3$, the usual estimator $\boldsymbol{X}$ is still minimax but inadmissible. James and Stein (1961) provided the corresponding modified James–Stein estimator (1.15) and (1.17) when $\boldsymbol{\Sigma}$ is assumed to be $\sigma^2 I$ with unknown σ^2 or totally unknown.

Berger and Bock (1976) considered the above problem using the quadratic loss

$$L(\boldsymbol{\delta}, \boldsymbol{\theta}, \boldsymbol{\Sigma}) = (\boldsymbol{\delta} - \boldsymbol{\theta})^T Q(\boldsymbol{\delta} - \boldsymbol{\theta})/\mathrm{tr}(Q\boldsymbol{\Sigma}),$$

where $Q = \mathrm{diag}(q_1, \cdots, q_p)$ with $q_i > 0$ and tr denotes the trace, and raised a class of minimax estimators. In this paper, they made the independence assumption among the elements of $\boldsymbol{X}$, so $\boldsymbol{\Sigma}$ is assumed to be diagonal with unknown diagonal elements σ_i's. They further assumed that an estimate of $\boldsymbol{\Sigma}$, $\boldsymbol{S} = \mathrm{diag}(s_1, \cdots, s_p)$ is available, with s_i/σ_i following a chi-square distribution with $n_i \geq 3$ degrees of freedom. The s_i's are assumed mutually independent and s_i is assumed independent of X_i. Then estimators of the form

$$\boldsymbol{\delta}(\boldsymbol{X}, W) = (I - r(\boldsymbol{X}, W)\|\boldsymbol{X}\|_W^{-2} Q^{-1} W^{-1})\boldsymbol{X}, \tag{4.1}$$

are minimax and have risks lower than 1 (the risk of $\boldsymbol{X}$) under certain conditions on $r(\boldsymbol{X}, W)$, where $W = \mathrm{diag}(W_1, \cdots, W_p)$ with $W_i = s_i/(n_i - 2)$, and

$$\|\boldsymbol{X}\|_W^2 = \boldsymbol{X}^T W^{-1} Q^{-1} W^{-1} X. \tag{4.2}$$

For practical purpose, the authors recommended using the estimator

$$\delta_i^{c+}(\boldsymbol{X}, W) = (1 - c/(\|\boldsymbol{X}\|_W^2 q_i W_i))^+ X_i, \tag{4.3}$$

under the assumption that $0 \leq c \leq 2(p - 2\tau)$, where $\tau = \mathrm{E}(T^{-1})$, $T = \min(\chi^2_{n_i}/n_i)$ and $\chi^2_{n_i}$, $i = 1, \cdots, p$, denotes independent chi-square random variables with corresponding degrees of freedom n_i. This estimator is recommended for its simplicity and considerable better risk than $\boldsymbol{X}$.

Chetelat and Wells (2012) studied the normal mean estimation with unknown covariance problem under loss function (1.1), but focused on the $p > n$ case. Assume that S is observed along with but independent of $\boldsymbol{X}$, having a Wishart distribution with n degrees of freedom. The authors constructed a class of estimators based on the sufficient statistics $(\boldsymbol{X}, S)$, which generalizes several previous estimators to the

$p \geq n$ setting, including the estimator (4.1) and the James–Stein estimator (1.17). Assuming $\min(p, n) \geq 3$, under certain conditions on a function r, the estimator

$$\boldsymbol{\delta}_r(\boldsymbol{X}, S) = \left(I - \frac{r(\boldsymbol{X}^T S^+ \boldsymbol{X}) S S^+}{\boldsymbol{X}^T S^+ \boldsymbol{X}} \right) \boldsymbol{X} = \boldsymbol{X} + g(\boldsymbol{X}, S) \tag{4.4}$$

is minimax and dominates $\boldsymbol{X}$. Explicitly, when $p > n \geq 3$, the estimator (4.4) dominates $\boldsymbol{X}$ if r is nondecreasing, differentiable and satisfies

$$0 \leq r \leq \frac{2(n-2)}{p-n+3}.$$

More computational details are given in Chetelat and Wells (2012).

When $n > p - 3$, still assuming $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ and observing $S \sim \mathrm{Wishart}(n, \boldsymbol{\Sigma})$, independent of $\boldsymbol{X}$, Lin and Tsai (1973) extended the James-Stein estimator (1.17) by replacing $(p-2)/(n-p+3)$ with a function of $y = \boldsymbol{X}^T S^{-1} \boldsymbol{X}$, $r(y)$. The authors showed that when $p \geq 3$, estimators of the form

$$\boldsymbol{\delta}(\boldsymbol{X}, S) = (1 - r(y)/y) \boldsymbol{X} \tag{4.5}$$

are minimax under loss function (1.1), if $r(y)$ is nonnegative, non-decreasing function and $r(y) \leq 2(p-2)/(n-p+3)$. Furthermore, this paper also showed that under prior

$$\begin{aligned} \boldsymbol{\theta} \mid \lambda, \boldsymbol{\Sigma} &\sim \mathrm{N}\left(\boldsymbol{\theta}, \frac{1-\lambda}{\lambda} \boldsymbol{\Sigma} \right), \\ p(\boldsymbol{\Sigma}^{-1}) &\propto |\boldsymbol{\Sigma}|^{\frac{1}{2}\nu}, \\ p(\lambda) &\propto \lambda^{-a}, \end{aligned} \tag{4.6}$$

where $\nu \leq n$ is an integer and $a < p/2 + 1$, the posterior mean of $\boldsymbol{\theta}$ given $\boldsymbol{X}$ and S has the form (4.6) by letting $r(y) = y\mathrm{E}(\lambda \mid \boldsymbol{X}, S)$. Under certain conditions, this estimator is a generalized Bayes minimax estimator.

Tsukuma (2009) studied the problem of estimating a normal mean matrix with an unknown covariance matrix. The author extended the prior in Lin and Tsai (1973) to include the $p > n$ case and showed that the resulting generalized Bayes estimators are minimax under certain conditions. Let $\boldsymbol{X}$ be an $n \times p$ random matrix with the row vectors mutually independent. Assume that the i-th row vector $\boldsymbol{X}_i$ follows a multivariate normal distribution with mean $\boldsymbol{\theta}_i$ and unknown covariance matrix $\boldsymbol{\Sigma}$, and observe S independent of $\boldsymbol{X}$ following Wishart$(m, \boldsymbol{\Sigma})$, then these models can be written as

$$\begin{aligned} \boldsymbol{X} &\sim \mathrm{N}(\boldsymbol{\Theta}, I \otimes \boldsymbol{\Sigma}), \\ S &\sim \mathrm{Wishart}(m, \boldsymbol{\Sigma}), \end{aligned} \tag{4.7}$$

where $\otimes$ represents the Kronecker product. Let $n \bigwedge p = \min(n,p)$ and $n \bigvee p = \max(n,p)$. Konno (1990), Konno (1991) and Konno (1992) showed that the estimator

$$\boldsymbol{\delta}_K = \begin{cases} (I - RF^{-1}\Phi(F)R^T)\boldsymbol{X} & \text{if } n < p, \\ \boldsymbol{X}(I - QF^{-1}\Phi(F)Q^{-1}) & \text{if } n \geq p, \end{cases} \tag{4.8}$$

is minimax with respect to the loss function

$$L(\boldsymbol{\delta}, \boldsymbol{\Theta}, \boldsymbol{\Sigma}) = \mathrm{tr}(\boldsymbol{\delta} - \boldsymbol{\Theta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\delta} - \boldsymbol{\Theta})^T \tag{4.9}$$

under certain conditions, where R is an $n \times n$ orthogonal matrix such that

$$\boldsymbol{X} S^{-1} \boldsymbol{X}^T = RFR^T, \tag{4.10}$$

$F = \mathrm{diag}(f_1, \cdots, f_{n \bigwedge p})$ is a diagonal matrix based on ordered eigenvalues, $\Phi(F) = \mathrm{diag}(\phi_1(F), \cdots, \phi_{n \bigwedge p}(F))$ with elements $\phi_i(F)$ being functions of F and Q is a $p \times p$ nonsingular matrix such that $Q^T S Q = I$ and $Q^T \boldsymbol{X}^T \boldsymbol{X} Q = F$. Tsukuma (2009) extended the prior (4.6) to

$$\begin{aligned} \boldsymbol{\Theta} &\sim \mathrm{N}(\mathbf{0}, \Omega \otimes \boldsymbol{\Sigma}), \\ p(\Omega) &\propto |I + \Omega|^{-a/2-n}, \\ p(\boldsymbol{\Sigma}^{-1}) &\propto |\boldsymbol{\Sigma}^{-1}|^{(b+p)/2}, \end{aligned} \tag{4.11}$$

if $n < p$, and to

$$\begin{aligned} \boldsymbol{\Theta} &\sim \mathrm{N}(\mathbf{0}, I \otimes \Xi), \\ p(\Xi) &\propto |I + \boldsymbol{\Sigma}^{-1/2} \Xi \boldsymbol{\Sigma}^{-1/2}|^{-a/2-p}, \\ p(\boldsymbol{\Sigma}^{-1}) &\propto |\boldsymbol{\Sigma}^{-1}|^{(b+p)/2}, \end{aligned} \tag{4.12}$$

if $n \geq p$. This paper also showed that the generalized Bayes estimator

$$\boldsymbol{\delta} = \begin{cases} \mathrm{E}_{\boldsymbol{\Theta},\Omega,\boldsymbol{\Sigma}^{-1}|\boldsymbol{X},S}(\boldsymbol{\Theta}\boldsymbol{\Sigma}^{-1})(\mathrm{E}_{\boldsymbol{\Theta},\Omega,\boldsymbol{\Sigma}^{-1}|\boldsymbol{X},S}(\boldsymbol{\Sigma}^{-1}))^{-1}, & \text{if } n < p, \\ \mathrm{E}_{\boldsymbol{\Theta},\Xi,\boldsymbol{\Sigma}^{-1}|\boldsymbol{X},S}(\boldsymbol{\Theta}\boldsymbol{\Sigma}^{-1})(\mathrm{E}_{\boldsymbol{\Theta},\Xi,\boldsymbol{\Sigma}^{-1}|\boldsymbol{X},S}(\boldsymbol{\Sigma}^{-1}))^{-1}, & \text{if } n \geq p, \end{cases} \tag{4.13}$$

belongs to (4.8) and therefore is minimax with respect to loss (4.9) under certain conditions.

4.2.2 Model Structure and Computational Details

Model Structure

Most literature on the normal mean estimation with unknown covariance matrix problem are in the decision theory field and assume observing S independent of $\boldsymbol{X}$. Not many papers discuss how to put prior distributions on $\boldsymbol{\Sigma}$, perhaps because the prior selection of $\boldsymbol{\Sigma}$ is a difficult task due to the large number of parameters in a covariance matrix and the non-negative definite constraint. One common choice is the usual inverse-Wishart prior, because it is a conjugate prior of a multivariate normal distribution. However, under this prior, all standard deviation components share a single degree of freedom. This is a big restriction especially when the dimension is large (Barnard, McCulloch and Meng 2000). Instead of putting a prior distribution on the whole covariance matrix, we consider the standard deviation and the correlation matrix decomposition:

$$\boldsymbol{\Sigma} = SRS, \tag{4.14}$$

where S is a diagonal matrix of standard deviations and R is the correlation matrix. This decomposition strategy offers two benefits: First, it has more flexibility to deal with different standard deviation components. Second, i has the ability to incorporate prior information about the standard deviations. Based on this decomposition, to make the prior distributions diffuse, we put independent Jeffreys priors on the standard deviations and apply the marginally uniform priors from Barnard et al. (2000) on the correlation matrix R:

$$\begin{aligned} \pi(\sigma_i) &\propto \frac{1}{\sigma_i}, \\ \pi(R) &\propto |R|^{-(p+1)} \, (\prod_i r^{ii})^{-\frac{p+1}{2}}, \end{aligned} \tag{4.15}$$

where r^{ii} is the i-th diagonal element of R^{-1}. The prior distribution (4.15) implies a marginal uniform(-1, 1) distribution for each individual correlation element.

Extending the model structure in Chapter 2 to this general covariance matrix case, we modify the prior distribution on $\boldsymbol{\theta}$ as follows:

$$\boldsymbol{\theta} \mid S, R, \boldsymbol{\lambda}, \tau \sim N(\mathbf{0}, \tau^2 S R_H \text{Diag}(\boldsymbol{\lambda}) R_H^T S), \tag{4.16}$$

where R_H is such that $R_H R_H^T = R$. Using this extension, we can see that if $\boldsymbol{Y}$ is pre-multiplied by $R_H^{-1} S^{-1}$, then the transformed parameter $\boldsymbol{\theta}^* = R_H^{-1} S^{-1} \boldsymbol{\theta}$ follows $N(\mathbf{0}, \tau^2 \text{Diag}(\boldsymbol{\lambda}))$ and $\boldsymbol{Y}^* = R_H^{-1} S^{-1} \boldsymbol{Y}$ follows $N(\boldsymbol{\theta}^*, I)$, which is same as model (2.6) with $\sigma = 1$.

Method For Updating The Correlation Matrix

To update the correlation matrix R in the MCMC procedure, we need to pay attention to two restrictions. First, all correlations need to be in the range $[-1, 1]$. Second, the updated correlation matrix R^* need to be positive definite. Let $R_{i,j}(r)$ be the updated correlation matrix by changing the correlation between the i-th component and the j-th component to r. Barnard et al. (2000) showed that

$$f(r) = |R_{i,j}(r)| > 0 \tag{4.17}$$

is a sufficient and necessary condition for $R_{i,j}(r)$ to be positive definite. Without loss of generality, assume $i > j$. To calculate $f(r)$, expanding $R_{i,j}(r)$ along the i-th row yields

$$\begin{aligned} f(r) &= \sum_{k=1}^{p}(-1)^{i+k} r_{ik}|R_{-ik}| \\ &= (-1)^{i+j} r|R_{-ij}| + |R_{-ii}| + \sum_{k \neq i,j}(-1)^{i+k} r_{ik}|R_{-ik}|, \qquad (4.18) \end{aligned}$$

where r_{ik} is the correlation between the i-th component and k-th component and R_{-ik} is the remaining correlation matrix after removing the i-th row and the k-th column. From equation (4.18), it is easy to see that the matrix R_{-ij} contains r. Expanding this matrix by the j-th row shows that $f(r)$ is a quadratic function of r which can be written as

$$f(r) = ar^2 + br + c, \tag{4.19}$$

where

$$\begin{aligned} a &= \frac{1}{2}(f(1) + f(-1) - 2f(0)), \\ b &= \frac{1}{2}(f(1) - f(-1)), \\ c &= f(0). \end{aligned}$$

Therefore, the range of r that makes the updated correlation matrix valid, is determined by the roots of $f(r)$ and ± 1. Furthermore, the only term involving r from expanding R_{-ij} along the j-th row is

$$(-1)^{j+i-1} r |R_{-ij-ji}|, \tag{4.20}$$

where the -1 inside the index comes from the fact that the j-th column has been removed and R_{-ij-ji} stands for the remaining correlation matrix by removing the i-th and j-th rows and columns. Thus, a is always negative and the lower and the upper bound for the range are

$$\begin{aligned} B_L &= max(-1, \frac{-b+\sqrt{b^2-4ac}}{2a}), & (4.21) \\ B_U &= min(1, \frac{-b-\sqrt{b^2-4ac}}{2a}). & (4.22) \end{aligned}$$

After the new value is generated from a uniform distribution Unif(B_L, B_U), we can simply apply the Metropolis-Hasting Algorithm to calculate the acceptance ratio and get the updated R.

4.2.3 Preliminary Simulation Studies

In this section, we conduct simulation studies of the normal mean estimation problem with unknown covariance structures by considering three types of covariance matrices. In the first scenario, $\mathbf{\Sigma}$ is assumed to have an independent structure. In the second scenario, $\mathbf{\Sigma}$ is assumed to have a local-dependent correlation structure. In the third one, $\mathbf{\Sigma}$ is assumed to have an exchangeable correlation structure. Because of the large number of parameters in the correlation matrix R and the time consuming issue in running MCMC procedures, we set $p = 5$ and the studies are preliminary. In the future, with better computation algorithms, the parameter dimension should be increased to yield better simulation studies for high dimensional problems.

Scenario 1 – When $\boldsymbol{\Sigma}$ Has an Independent Structure

In this scenario, we generate $\boldsymbol{\theta}$ using model (3.30) with the combinations of three nonzero values: 1, 4, 10, and three nonzero percentages: 5, 20 and 50. The response vectors $(\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n)$ are independently generated from $\mathrm{N}(\boldsymbol{\theta}, \sigma^2 I)$ with $p = 5$ and $n = 20$. We set $\sigma^2 = 20$ so that $\bar{Y}_i$ has variance 1, where $\bar{\boldsymbol{Y}}$ is the sample average of $(\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n)$. For each combination, 2000 datasets are generated with 40000 iterations total and a burn-in period of 20000. The results are evaluated using the invariant squared error loss function:

$$L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}). \tag{4.23}$$

We compare the estimation performances of the AIB prior with those of the HS, the SB and the NEG priors by letting all priors have the same model structure:

$$\begin{aligned} \boldsymbol{\theta} \mid \boldsymbol{\lambda}, \sigma^2, \tau &\sim \mathrm{N}(\mathbf{0}, \sigma^2 \tau^2 \mathrm{Diag}(\boldsymbol{\lambda})), \\ p(\sigma^2) &\propto \frac{1}{\sigma^2}, \\ \tau &\sim \mathrm{C}^+(0,1), \end{aligned} \tag{4.24}$$

with different priors on λ_i as those in previous chapters. The performances of the MLE estimator are also reported as benchmarks. Similar to the simulation studies in previous chapter, the average losses across 2000 datasets are reported in Table 4.1 along with the standard deviations of the averaged losses in the lower rows. The estimates of hyperparameters are reported in Table 4.2 along with the averaged posterior standard deviations. Figure 4.1 and 4.2 show the boxplots of standard deviations in S and correlations in R. Because the standard deviations of all dimensions and the correlations between any two different dimensions are equal, all dimensions are plotted in one box for each method in both figures.

Scenario		AIB	HS	SB	NEG	MLE
1	5%	4.67	5.33	5.73	4.96	25.24
		0.109	0.122	0.127	0.108	0.331
	20%	6.32	6.90	7.18	6.56	23.78
		0.111	0.122	0.128	0.111	0.310
	50%	11.77	12.23	12.67	11.92	27.21
		0.207	0.221	0.229	0.210	0.416
4	5%	9.69	10.15	10.56	9.84	26.32
		0.278	0.271	0.272	0.268	0.360
	20%	17.73	18.04	17.98	17.98	24.19
		0.312	0.302	0.295	0.300	0.295
	50%	22.61	22.32	22.29	22.48	24.01
		0.294	0.289	0.291	0.295	0.322
10	5%	7.61	8.46	8.49	9.96	23.22
		0.235	0.245	0.235	0.236	0.308
	20%	17.19	17.75	17.60	17.77	23.64
		0.348	0.353	0.340	0.341	0.327
	50%	24.61	24.73	24.59	24.58	25.34
		0.453	0.448	0.451	0.452	0.447

Table 4.1: Comparison of the averaged invariant squared error losses among the AIB, the HS, the SB and the NEG models when $\mathbf{\Sigma} = 20I$. The losses of the MLE estimator are given as benchmarks. The averaged squared error losses across 2000 datasets are reported in the upper rows. The standard deviations of the average losses are reported in the lower rows. For all datasets, $p = 5$ and 20 observations are generated for each dimension. The signal sizes are listed in the first column and the nonzero percentages are in the second column. All values are multiplied by 100.

All methods give better performances than the MLE in all combinations, but the improvements are little when there are 50% nonzero values at 10. Given $p = 5$ it is reasonable to expect the performance differences between these methods and the MLE to small, especially when the sparsity is low and the signal size is large. The improvements may be more obvious when p is large. An increasing pattern can be detected in the losses for all methods as the nonzero percentage increases.

When the signal is 1, the AIB prior and the NEG prior give better performances than the HS and the SB priors. But the differences decrease as the nonzero percentage increases. When the signal is 4 or 10, the AIB prior and the NEG prior give better performance when the data is highly sparse. When the nonzero percentage increases to 20 or 50, these methods have similar performances.

Compared to the first simulation study in Chapter 2, all models are less better than the MLE, especially when the sparsity is low and the signal size is large. The low dimension is one factor to this change. Another factor is that without the independence assumption, we have more parameters need to be estimated. This automatically increases the estimation difficulty which leads to less improvements compared to the MLE.

The patterns in the estimates of the hyperparameters are very similar to those in Chapter 2. When the nonzero percentage or the signal size increases, the estimates of τ of all methods increase and the estimates of N decrease. Both patterns indicate that less shrinkage is given as the sparsity decreases. Among the HS, the SB and the NEG priors, the SB prior gives smallest estimates of τ and the NEG prior gives largest estimates of τ across all combinations, which is consistent with the simulations in Chapter 2 and Chapter 3. The estimates of τ in the AIB prior have larger posterior variances than those of other methods. Again, this is because that unfixing

Scenario		M	N	τ(AIB)	τ(HS)	τ(SB)	τ(NEG)
1	5%	2.47	0.66	0.37	0.22	0.13	0.29
		1.419	0.204	0.508	0.165	0.103	0.218
	20%	2.46	0.66	0.40	0.24	0.14	0.32
		1.444	0.198	0.625	0.226	0.137	0.311
	50%	2.46	0.65	0.46	0.28	0.17	0.39
		1.473	0.207	0.575	0.341	0.210	0.360
4	5%	2.46	0.65	0.45	0.28	0.17	0.39
		1.410	0.189	0.550	0.371	0.210	0.383
	20%	2.46	0.62	0.73	0.47	0.30	0.69
		1.421	0.196	0.975	0.521	0.251	0.530
	50%	2.44	0.57	0.96	0.68	0.45	1.04
		1.393	0.210	0.614	0.325	0.187	0.333
10	5%	2.48	0.64	0.57	0.39	0.25	0.56
		1.408	0.197	0.225	0.177	0.084	0.191
	20%	2.46	0.57	1.07	0.78	0.53	1.26
		1.487	0.189	0.307	0.174	0.104	0.209
	50%	2.39	0.51	1.58	1.19	0.83	2.12
		1.337	0.217	3.216	0.908	0.582	1.173

Table 4.2: Comparison of hyperparameters among the AIB, the HS, the SB and the NEG models when $\mathbf{\Sigma} = 20I$. The averaged posterior means across 2000 datasets are reported in the upper rows. The averaged posterior standard deviations are reported in the lower rows. For all datasets, $p = 5$ and 20 observations are generated for each dimension. The signal sizes are listed in the first column and the nonzero percentages are in the second column.

N introduces more variability. The estimates of τ in the AIB prior consistently increase when those of N decrease. This shows that the global shrinkage parameter τ corporates with the local shrinkage parameters λ_i to balance the shrinkage powers.

As shown in Figures 4.1 and 4.2, all methods similar estimates of the standard deviations and the correlations. The standard deviations are slightly overestimated, while most estimates of correlations are around the true value.

Scenario 2 – When $\boldsymbol{\Sigma}$ Has a Local-Dependent Correlation Structure

In this scenario, we use the same model as in Scenario 1 to generate $\boldsymbol{\theta}$ with the same combinations of nonzero values and sparsity levels. Again 20 response vectors $(\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n)$ are independently generated from $\mathrm{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ with $p = 5$. The standard deviations of all dimensions are still $\sqrt{20}$, but the correlations along the off-diagonal lines of R are set to be 0.5. For each combination, 2000 datasets are generated with 40000 iterations total and a burn-in period of 20000. The invariant squared error loss function (4.23) is used to compare the estimation performances among the AIB, the HS, the SB and the NEG models. The model structure and priors on λ_i are the same with those in Scenario 1. The average losses are reported in Table 4.3 and the estimates of hyperparameters are reported in Table 4.4. Figure 4.3 and Figure 4.4 show the boxplots of standard deviation and correlation estimates. For the same reason, one box for each method is reported in Figure 4.3. Two boxes for each method are reported in Figure 4.4 for the two correlation values: 0, 0.5.

All methods give better performance than the MLE in most combinations but the improvements are very little when there are 50% nonzero values at 4 or 10. An increasing pattern can be detected in the losses for all methods as the nonzero percentage increases. When the data are highly sparse, all models have the largest losses when the signal is 4. Again, this is because that the noises and signals are partially mixed. Incorrect discriminations will cause great losses. When the nonzero percentage is 20 or 50, this pattern is not obvious. When the signal is 1, the AIB prior gives better performance than other three models. But the differences are smaller as the nonzero percentage increases. When the signal is 4 or 10, all models give similar

Scenario		AIB	HS	SB	NEG	MLE
1	5%	5.24	5.99	6.43	5.73	25.56
		0.141	0.159	0.168	0.151	0.354
	20%	7.62	8.03	8.23	7.82	24.44
		0.145	0.156	0.157	0.146	0.328
	50%	11.81	12.21	12.28	11.97	25.83
		0.193	0.205	0.208	0.197	0.378
4	5%	10.28	10.67	10.63	10.40	24.17
		0.354	0.343	0.326	0.343	0.315
	20%	19.86	19.67	19.23	19.51	24.41
		0.439	0.415	0.405	0.428	0.335
	50%	24.31	24.04	23.64	24.09	24.71
		0.355	0.348	0.339	0.345	0.362
10	5%	8.49	9.19	9.27	8.76	24.76
		0.268	0.280	0.272	0.267	0.359
	20%	19.40	19.83	19.62	19.39	23.48
		0.417	0.426	0.412	0.413	0.347
	50%	24.56	24.79	24.51	24.60	25.52
		0.346	0.349	0.343	0.348	0.342

Table 4.3: Comparison of invariant squared error losses among the AIB, the HS, the SB and the NEG models when the off-diagonal elements of the correlation matrix R are 0.5. The losses of the MLE estimate are given as benchmarks. The averaged losses across 2000 datasets are reported in the upper rows. The standard deviations of the average losses are reported in the lower rows. For all datasets, $p = 5$ and $n = 20$. The signal sizes are listed in the first column and the nonzero percentages are in the second column. All values are multiplied by 100.

performances. Compared to Scenario 1, we see that the losses are larger and the differences among different models are less obvious. This shows that the correlated observations increase the estimation difficulty and the advantage of the AIB prior still exists but becomes less. It is reasonable to expect that the improvements by using the AIB prior becomes more remarkable when p is large.

Scenario		M	N	τ(AIB)	τ(HS)	τ(SB)	τ(NEG)
1	5%	2.47	0.66	0.37	0.22	0.13	0.30
		1.385	0.188	0.405	0.219	0.108	0.221
	20%	2.46	0.66	0.40	0.25	0.14	0.33
		1.420	0.180	0.542	0.216	0.116	0.237
	50%	2.46	0.66	0.46	0.28	0.17	0.38
		1.480	0.212	0.531	0.362	0.203	0.421
4	5%	2.48	0.64	0.50	0.32	0.20	0.46
		1.467	0.211	0.451	0.292	0.175	0.389
	20%	2.46	0.61	0.75	0.50	0.32	0.72
		1.433	0.196	0.634	0.526	0.314	0.515
	50%	2.51	0.57	1.00	0.74	0.49	1.08
		1.427	0.195	0.800	0.648	0.332	0.602
10	5%	2.49	0.63	0.63	0.43	0.28	0.65
		1.499	0.200	0.425	0.259	0.140	0.249
	20%	2.47	0.55	1.23	0.93	0.64	1.53
		1.390	0.210	2.081	1.345	0.679	1.655
	50%	2.47	0.50	1.73	1.35	0.93	2.28
		1.418	0.205	1.732	1.083	0.559	1.220

Table 4.4: Comparison of hyperparameters among the AIB, the HS, the SB and the NEG models when the correlations matrix has 0.5 along the off-diagonal lines. The averaged posterior means across 2000 datasets are reported in the upper rows. The averaged posterior standard deviations are reported in the lower rows. For all datasets, $p = 5$ and 20 observations are generated for each dimension. The signal sizes are listed in the first column and the nonzero percentages are in the second column.

Similar to scenario 1, as the nonzero percentage or the signal size increases, the estimates of τ of all methods increase and the estimates of N decrease. Excluding the AIB prior, the SB prior still gives the smallest estimates of τ and the NEG prior gives the largest estimates across all combinations. The estimates of τ in the AIB prior have larger posterior variance than those of other methods. Compared to Scenario 1, the non-equal correlations increase the posterior standard deviations of τ in the HS, the SB and the NEG models. But those of N and τ in the AIB prior decrease. This shows that unfixing N and τ is more stable to correlation changes, which is another unique advantage of the AIB prior.

Figure 4.3 and Figure 4.4 show that the estimates of σ_i and correlations are similar for all methods. The standard deviations are still slightly overestimated, the estimates of 0 correlations are good, but the 0.5 correlations are underestimated. Compared to Scenario 1, involving nonzero correlations improves the estimates of σ_i.

Scenario 3 – When $\mathbf{\Sigma}$ Has an Exchangeable Correlation Structure

In this scenario, we use the same simulation setup as in previous scenarios, but let the covariance matrix $\mathbf{\Sigma}$ have the $aI + bJ$ form where J represents the square matrix of 1. We set $a = 10$ and $b = 10$ so that $\bar{Y}_i$ has variance 1. Now the correlation between any pair of different dimensions is 0.5. Again, we use the invariant squared error loss function (4.23) to compare the estimation performances among the AIB, the HS, the SB and the NEG models. The averaged losses are reported in Table 4.5 and the estimates of hyperparameters are reported in Table 4.6. Figure 4.5 and 4.6 show the boxplots of standard deviations and correlations. As in Scenario 1, one box for each method is reported in both figures.

When all dimensions are correlated, all models give better performance than the MLE in most scenarios. But when there are 50% nonzero values at 4 or 10, they are not distinguishable from the MLE. For fixed signal sizes, all models have increased losses as the sparsity decreases. When the signal is 1, the AIB prior gives consistently better performance than other three models. But the differences decrease as the nonzero percentage increases. When the signal is 4 or 10, all models give similar performances. Compared to the previous scenarios, all models have larger losses

Scenario		AIB	HS	SB	NEG	MLE
1	5%	4.83	5.43	5.69	5.09	23.42
		0.133	0.145	0.151	0.134	0.311
	20%	7.02	7.54	7.76	7.16	23.41
		0.166	0.181	0.184	0.168	0.355
	50%	8.62	8.92	9.04	8.68	24.35
		0.115	0.123	0.128	0.117	0.293
4	5%	9.25	9.81	10.08	9.34	25.37
		0.308	0.301	0.296	0.290	0.347
	20%	23.64	23.55	23.18	23.31	27.67
		0.515	0.494	0.472	0.497	0.354
	50%	25.42	25.62	24.72	24.72	24.75
		0.342	0.341	0.332	0.331	0.316
10	5%	11.09	11.92	11.63	11.32	23.91
		0.356	0.364	0.347	0.351	0.366
	20%	17.77	18.26	18.15	17.67	24.03
		0.350	0.353	0.346	0.343	0.290
	50%	24.38	24.68	24.53	24.46	24.15
		0.347	0.352	0.347	0.347	0.334

Table 4.5: Comparison of invariant squared error losses among the AIB, the HS, the SB and the NEG models when $\mathbf{\Sigma} = 10I + 10J$. The losses of the MLE estimate are given as benchmarks. The averaged losses across 2000 datasets are reported in the upper rows. The standard deviations of the averaged losses are reported in the lower rows. For all datasets, $p = 5$ and $n = 20$. The signal sizes are listed in the first column and the nonzero percentages are in the second column. All values are multiplied by 100.

when there are 20% or 50% nonzero values at 4, especially obvious for the 20% case. This is because that the nonzero correlations increase the difficulty of estimation, which makes the incorrect discriminations more harmful.

Similar to the previous 2 scenarios, as the nonzero percentage or the signal size increases, the estimates of τ of all methods have an increasing trend and the estimates of N have a decreasing trend. Excluding the AIB prior, the SB prior gives the smallest

Scenario		M	N	τ(AIB)	τ(HS)	τ(SB)	τ(NEG)
1	5%	2.47	0.66	0.38	0.23	0.13	0.30
		1.407	0.201	0.469	0.270	0.171	0.273
	20%	2.47	0.66	0.42	0.26	0.15	0.34
		1.379	0.194	0.530	0.187	0.118	0.247
	50%	2.48	0.66	0.44	0.27	0.16	0.36
		1.459	0.204	0.900	0.504	0.303	0.540
4	5%	2.44	0.65	0.48	0.30	0.19	0.43
		1.429	0.196	0.538	0.279	0.156	0.310
	20%	2.48	0.61	0.72	0.51	0.33	0.74
		1.385	0.203	0.528	0.328	0.190	0.357
	50%	2.52	0.58	0.97	0.70	0.47	1.04
		1.573	0.213	1.263	0.491	0.376	0.583
10	5%	2.47	0.62	0.66	0.45	0.30	0.69
		1.442	0.217	1.512	0.993	0.516	1.116
	20%	2.48	0.57	1.11	0.85	0.57	1.35
		1.388	0.192	0.356	0.189	0.119	0.228
	50%	2.59	0.49	1.79	1.45	1.01	2.41
		1.370	0.207	1.340	0.823	0.445	0.841

Table 4.6: Comparison of hyperparameters among the AIB, the HS, the SB and the NEG models when $\mathbf{\Sigma} = 10I + 10J$. The averaged posterior means across 2000 datasets are reported in the upper rows. The averaged posterior standard deviations are reported in the lower rows. For all datasets, $p = 5$ and $n = 20$. The signal sizes are listed in the first column and the nonzero percentages are in the second column.

estimates of τ while the NEG prior gives the largest across all combinations. The estimates of τ in the AIB prior have larger posterior variances than those of other methods.

All methods yield similar estimates of the standard deviations and the correlations. Similar to the previous scenarios, the standard deviations are overestimated a little and the nonzero correlations are underestimated.

4.2.4 Real Data Analysis

Introduction of Portfolio Choice Problems

Portfolio choice problems are popular in financial research. The mean-variance paradigm of Markowitz (1952) is one of the most common formulations of portfolio choice problems. Consider N risky assets with random return vector $\boldsymbol{R}_t$ and a riskfree asset with known return R_t^f at time t. Define the excess return vector

$$\boldsymbol{r}_t = \boldsymbol{R}_t - R_t^f \mathbf{1}_N, \tag{4.25}$$

where $\mathbf{1}_N$ is a vector of 1 with length N. Suppose the excess returns are independently and identically distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and denote the weight vector on the N risky assets at time t by $\boldsymbol{w}$ and the weight on the riskfree asset by $1 - \mathbf{1}_N^T \boldsymbol{w}$. Then the mean-variance paradigm is to choose a weight vector $\boldsymbol{w}$ to minimize the variance of the portfolio return

$$R_{p,t+1} = \boldsymbol{w}^T \boldsymbol{R}_{t+1} + (1 - \mathbf{1}_N^T \boldsymbol{w}) R_{t+1}^f = \boldsymbol{w}^T \boldsymbol{r}_{t+1} + R_{t+1}^f \tag{4.26}$$

at time $t+1$ for a pre-determined target expected portfolio return $R_{t+1}^f + \bar{\mu}$. That is to minimize

$$\mathrm{Var}(R_{p,t+1}) = \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w}, \tag{4.27}$$

subject to

$$\mathrm{E}(R_{p,t+1}) = \boldsymbol{w}^T \boldsymbol{\mu} + R_{t+1}^f = R_{t+1}^f + \bar{\mu}. \tag{4.28}$$

The solution to the above problem when both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are known is

$$\boldsymbol{w}_{Opt} = \frac{\bar{\mu}}{\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}. \tag{4.29}$$

The mean-variance problem can be rewritten as the following maximization problem:

$$\max_{\boldsymbol{w}} \mathrm{E}(r_{p,t+1}) - \frac{\gamma}{2} \mathrm{Var}(r_{p,t+1}), \tag{4.30}$$

where $r_{p,t+1} = \boldsymbol{w}^T\boldsymbol{r}_{t+1}$ is the portfolio excess return and γ is the relative risk aversion coefficient. The two problems are equivalent when

$$\frac{\bar{\mu}}{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}} = \frac{1}{\gamma}. \tag{4.31}$$

In practice, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are often unknown. One way to estimate $\boldsymbol{w}_{Opt}$ in this case is estimating $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ first with the observed data and then plugging the estimates into (4.29). However this method does not consider the uncertainty of the parameters, so the estimated weight may be substantially different from the best weight when the estimated parameters are substantially different from the truth. As a nature solution to this issue, Bayesian approaches are applied in this field. Following Zellner and Chetty (1965), a Bayesian estimate of the weight is defined by

$$\begin{aligned}\hat{\boldsymbol{w}}_B &= \arg\max_{\boldsymbol{w}} \int_{\boldsymbol{R_{t+1}}} U(\boldsymbol{w})p(\boldsymbol{R_{t+1}} \mid \boldsymbol{X})\mathrm{d}\boldsymbol{R_{t+1}} \\ &= \arg\max_{\boldsymbol{w}} \int_{\boldsymbol{R_{t+1}}} \int_{\boldsymbol{\mu}} \int_{\boldsymbol{\Sigma}} U(\boldsymbol{w})p(\boldsymbol{R_{t+1}}, \boldsymbol{\mu}, \boldsymbol{\Sigma} \mid \boldsymbol{X})\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\Sigma}\mathrm{d}\boldsymbol{R_{t+1}}, \end{aligned} \tag{4.32}$$

where $U(\boldsymbol{w})$ is the utility of holding the portfolio with weight $\boldsymbol{w}$ at time $t+1$ and $\boldsymbol{X}$ is the data available. Various prior densities of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given in this literature. For examples, the standard diffuse prior on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ has the form

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \mid \boldsymbol{\Sigma} \mid^{-\frac{N+1}{2}}, \tag{4.33}$$

the conjugate prior has the form

$$\begin{aligned}\boldsymbol{\mu} \mid \boldsymbol{\Sigma} &\sim \mathrm{N}\left(\boldsymbol{\mu}_0, \frac{1}{\tau}\boldsymbol{\Sigma}\right), \\ \boldsymbol{\Sigma} &\sim \mathrm{IW}(\boldsymbol{\Sigma}_0, \nu_0), \end{aligned} \tag{4.34}$$

where IW stands for an inverted Wishart distribution and the hyperparameters $\boldsymbol{\mu}_0$, τ, $\boldsymbol{\Sigma}_0$, ν_0 are assumed known or estimated using empirical Bayes methods. More details are given in Jorion (1986) and Avramov and Zhou (2010).

Considering the relationship between $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and the best weight in (4.29), an alternative way to implement Bayesian approaches is to put priors on $\boldsymbol{\Sigma}$ and the weight $\boldsymbol{w}$ directly. Since the weight $\boldsymbol{w}$ is of our interest, putting priors in this way allows investors to apply their prior knowledge of wealth allocation. Tu and Zhou (2010) gives an example of putting priors in this way. The prior on $\boldsymbol{w}$ is

$$\boldsymbol{w} \sim \mathrm{N}\left(\boldsymbol{w}_0, \frac{\sigma_\rho^2}{\gamma s^2}\boldsymbol{\Sigma}^{-1}\right), \tag{4.35}$$

or equivalently, using (4.29), the prior on $\boldsymbol{\mu}$ is

$$\boldsymbol{\mu} \sim \mathrm{N}\left(\gamma\boldsymbol{\Sigma}\boldsymbol{w}_0, \frac{\sigma_\rho^2}{s^2}\boldsymbol{\Sigma}\right), \tag{4.36}$$

where σ_ρ is a known parameter reflecting the degree of uncertainty about $\boldsymbol{w}$ or $\boldsymbol{\mu}$, s^2 is the average of the diagonal elements of $\boldsymbol{\Sigma}$ and a standard Wishart distribution on $\boldsymbol{\Sigma}$. When there is no data or prior information about the risky assets available, an equal weight is a reasonable choice for $\boldsymbol{w}_0$. That is

$$\boldsymbol{w}_0 \propto \frac{1}{N}\mathbf{1}_N. \tag{4.37}$$

Other choices of $\boldsymbol{w}_0$ include the value-weighted market portfolio weight $\boldsymbol{w}_m$ and the data determined weight

$$\boldsymbol{w}_0 = \frac{1}{\gamma}\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}, \tag{4.38}$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the sample mean and covariance matrix of the available excess returns.

Similar to Tu and Zhou (2010), we put priors on $\boldsymbol{w}$ and $\boldsymbol{\Sigma}$, but instead of using a standard Wishart distribution on $\boldsymbol{\Sigma}$, we implement the standard deviation and correlation decomposition strategy described in section 4.2.2. Without loss of generality, suppose $\gamma = 1$ and the observed excess returns are independently and identically distributed as

$$\boldsymbol{r}_t \mid \boldsymbol{w}, \boldsymbol{\Sigma} \sim \mathrm{N}(\boldsymbol{\Sigma}\boldsymbol{w}, \boldsymbol{\Sigma}), \tag{4.39}$$

where $\boldsymbol{\Sigma} = SR_H R_H^T S$. Assuming that the best weight is sparse, we put the AIB prior on $\boldsymbol{w}$

$$\begin{aligned} \boldsymbol{w} \mid \tau, \boldsymbol{\lambda}, S &\sim \mathrm{N}(\mathbf{0}, \tau^2 S^{-1} \mathrm{diag}(\boldsymbol{\lambda}) S^{-1}), \\ \lambda_i \mid M, N, &\sim \mathrm{IB}(MN, M(1-N)), \\ \tau &\sim \mathrm{C}^+(0,1), \\ p(\sigma_i) &\propto \frac{1}{\sigma_i}, \end{aligned} \tag{4.40}$$

and the prior (4.15) on R. In model (4.40), we assume that the best weight is sparse in priori and the elements are mutually independent. In fact, these are two common assumptions in practice, especially when N is large.

Analysis of Fama-French 25 Portfolios Formed on Size and Book-to-market

We consider the monthly returns of the Fama-French 25 portfolios formed on size and book-to-market, covering from August 1963 to July 2007. The portfolios are the intersections of 5 portfolios formed on size (market equity) and 5 portfolios formed on prior returns. More information is available from Ken French's website: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html. An old version of this data is studied by DeMiguel, Garlappi and Uppal (2006). In this paper, the authors claimed that "there is no single model that consistently delivers a Sharpe ratio or a certainty equivalent return that is higher than that of the 1/N portfolio, which also has a very low turnover". To point against this argument, Chevrier and

McCulloch (2008) applied Bayesian modelling techniques on this dataset but updated to include the period up to July 2007. Using economically motivated priors (EMP) on the parameters, they found that their approach outperforms the equal weight portfolio by about 30%. In fact, both paper studied other datasets from the above website, and the approach in Chevrier and McCulloch (2008) is consistently better than the $1/N$ portfolio.

Following Chevrier and McCulloch (2008), we subtract the riskfree rate from the monthly returns and augment the data with the Fama French 3 factors which makes the total number of risky assets $N = 25 + 3 = 28$. We use the data from August of year $t - 10$ to July of year t to fit the model, and use the data from August of year t to July of year $t + 1$ to get the out-of-sample measures, the Sharpe Ratios:

$$\mathrm{SR} = \frac{\hat{\mu}}{\hat{\sigma}}, \tag{4.41}$$

where $\hat{\mu}$ is the mean of the monthly portfolio excess returns and $\hat{\sigma}$ is the standard deviation of the monthly portfolio excess returns. We compare the out-of-sample Sharpe ratios of the AIB prior to those of the HS, the SB and the NEG priors.

Every August, we use the past 10 years data to estimate the asset allocation strategy and hold the determined portfolio for a year. The portfolios are rebalanced every year. The results are reported in Table 4.7. The column named “EMP” stands for the model in Chevrier and McCulloch (2008) and $1/N$ stands for equal weight portfolio. We get both results from Chevrier and McCulloch (2008). Notice that the EMP and the equal weight portfolios assume that all weights are greater than 0, while model (4.40) does not have this constraint. In other words, the weights in our method can be negative meaning assets can be held short. More explicitly, for example, when the weight on a stock is negative, it means that we borrow certain amount of this stock from others, sell it at the beginning of August year t, and purchase back the same amount at the end of July year $t+1$ and return it to the lenders. Thus, using model (4.40) is expected to improve the performances of the EMP and the equal weight portfolios remarkably.

AIB	HS	SB	NEG	EMP	1/N
93.6	93.0	92.5	92.3	71.3	57.1

Table 4.7: Out-of-sample Sharpe ratios comparison for the 1963 - 2007 period. All numbers are annualized and in percent. EMP stands for the model in Chevrier and McCulloch (2008) and $1/N$ stands for equal weight portfolio.

Table 4.7 shows that, all four models outperforms the EMP model about 30%, and outperforms the equal weight portfolio almost 65%. Among these four models, the AIB prior works a little better.

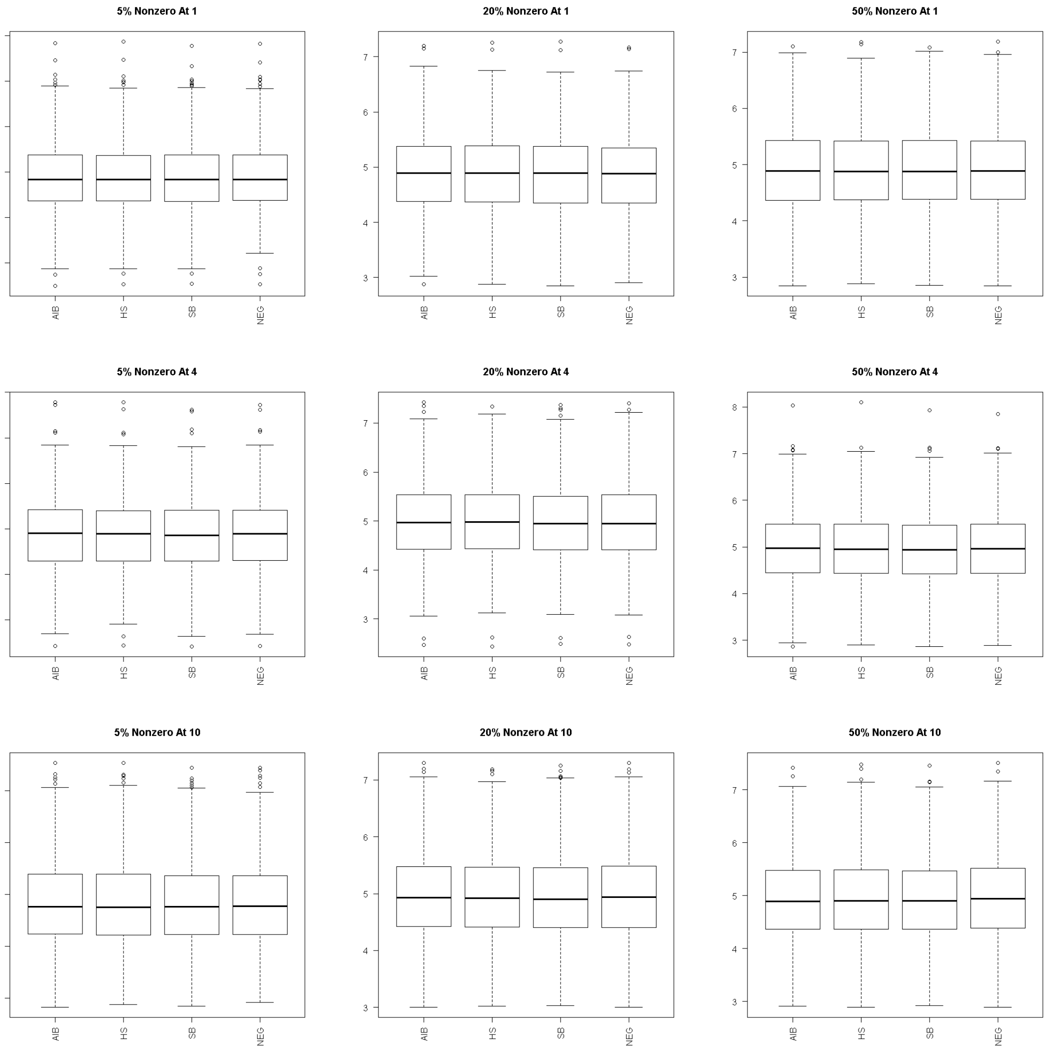

Figure 4.1: Boxplots of posterior means of σ_i's in the AIB, the HS, the SB and the NEG models when $\mathbf{\Sigma} = 20I$. All dimensions are plotted in one box using 2000 datasets for each model. In all datasets, $p = 5$ and 20 observations are generated for each dimension. The signal size is fixed at 1 in the upper row, 4 in the middle row and 10 in the lower row. The nonzero percentage is fixed at 5 in the left column, 20 in the middle column and 50 in the right column.

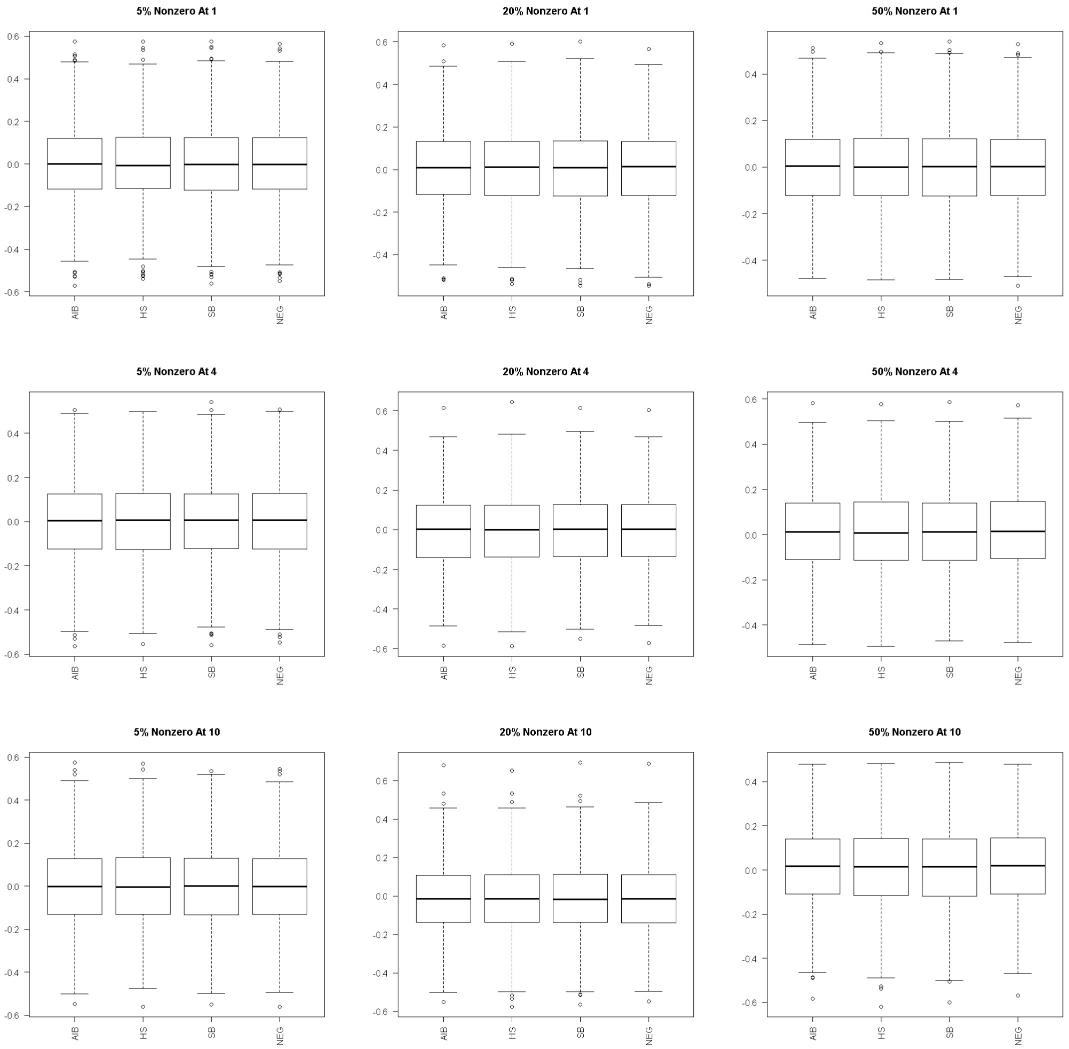

Figure 4.2: Boxplots of posterior means of correlations between different dimensions in the AIB, the HS, the SB and the NEG models when $\mathbf{\Sigma} = 20I$. All dimensions are plotted in one box using 2000 datasets for each model. In all datasets, $p = 5$ and 20 observations are generated for each dimension. The signal size is fixed at 1 in the upper row, 4 in the middle row and 10 in the lower row. The nonzero percentage is fixed at 5 in the left column, 20 in the middle column and 50 in the right column.

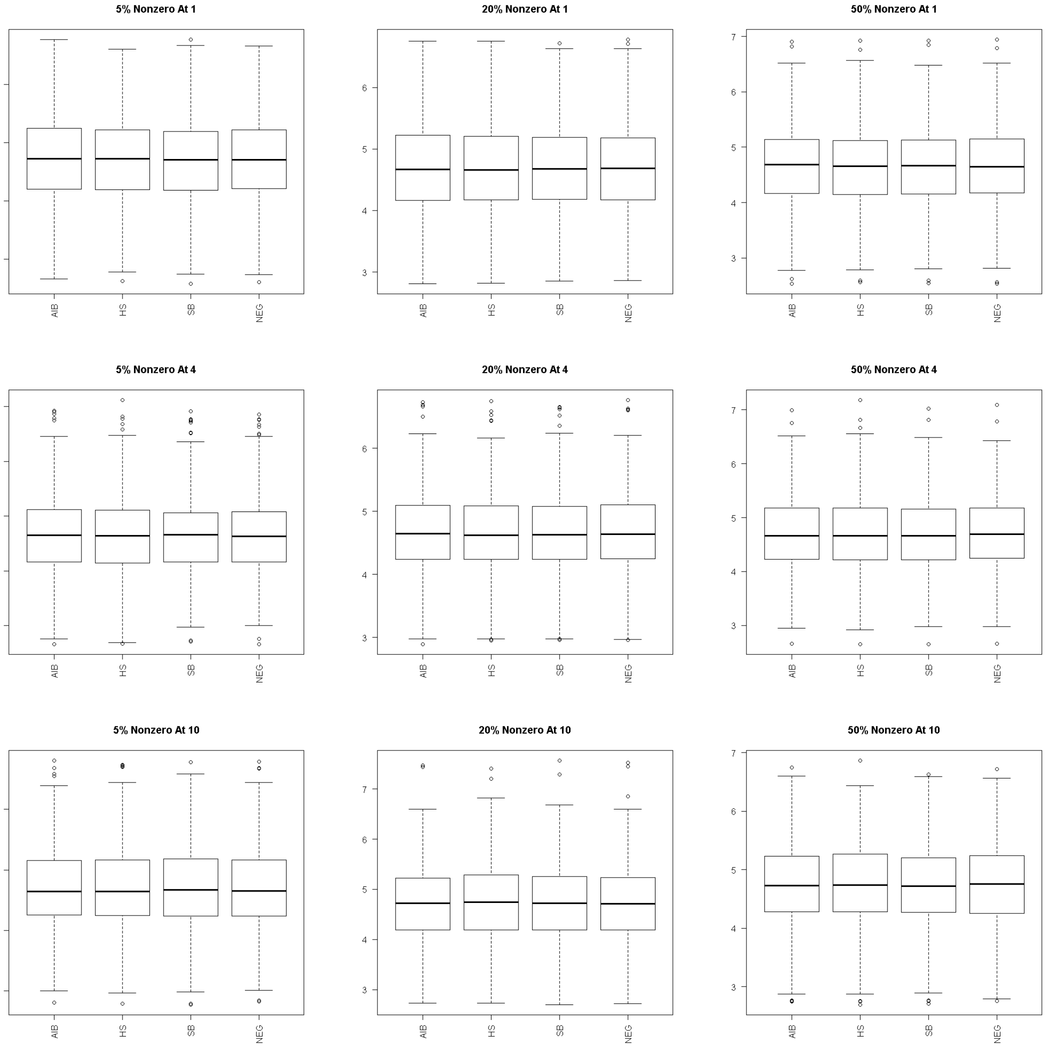

Figure 4.3: Boxplots of posterior means of σ_i in the AIB, the HS, the SB and the NEG models when the correlation matrix has 0.5 along the off-diagonal lines. The true σ_i's are $\sqrt{20}$. All dimensions are plotted in one box using 2000 datasets for each model with $p = 5$ and $n = 20$. The signal size is fixed at 1 in the upper row, 4 in the middle row and 10 in the lower row. The nonzero percentage is fixed at 5 in the left column, 20 in the middle column and 50 in the right column.

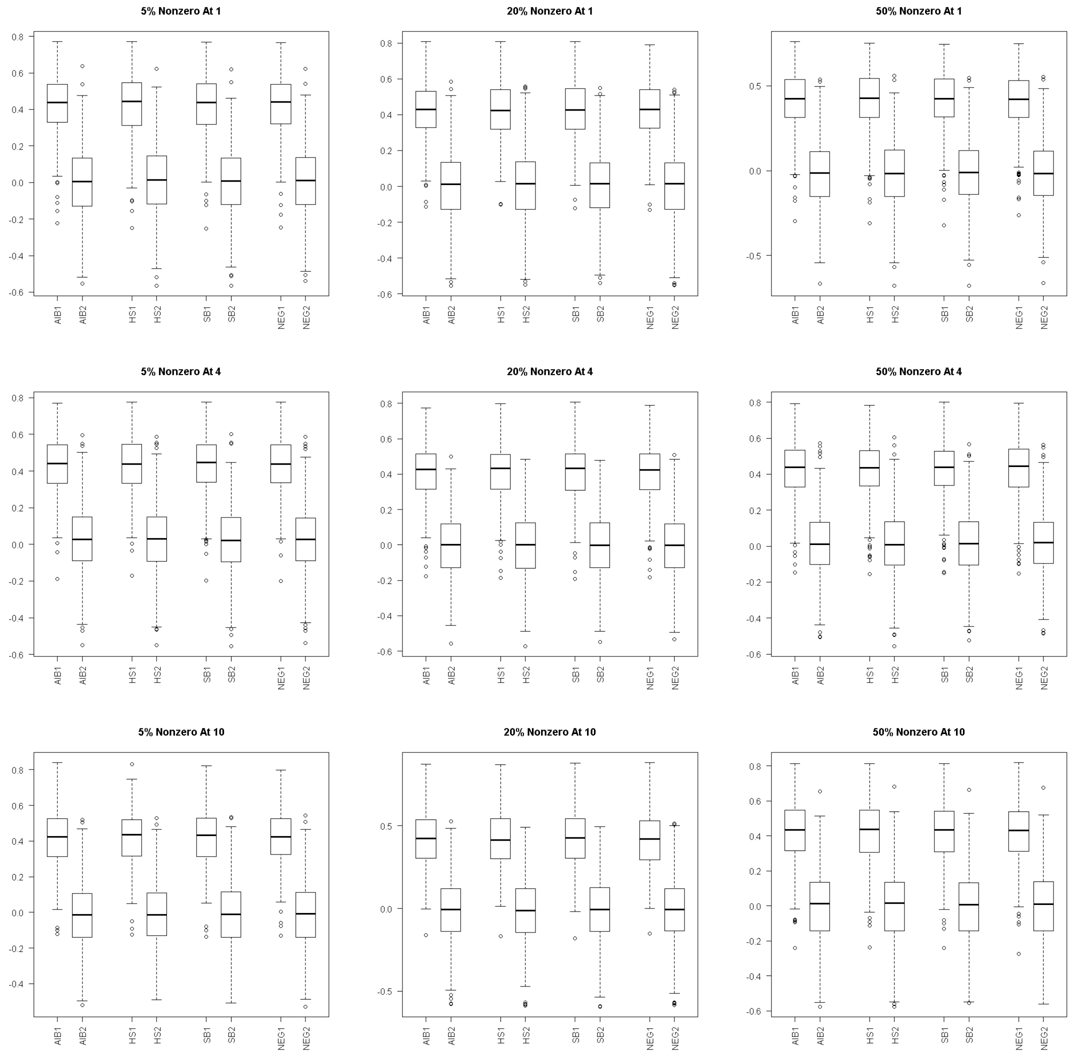

Figure 4.4: Boxplots of posterior means of correlations between different dimensions in the AIB, the HS, the SB and the NEG models. The true correlation matrix has 0.5 along the off-diagonal lines and 0 for other non-diagonal correlations. Two boxes are plotted using 2000 datasets for each model with $p = 5$ and $n = 20$. The first one plots the posterior means of correlations with true values 0.5 and the second one plots those ones with true values 0. The signal size is fixed at 1 in the upper row, 4 in the middle row and 10 in the lower row. The nonzero percentage is fixed at 5 in the left column, 20 in the middle column and 50 in the right column.

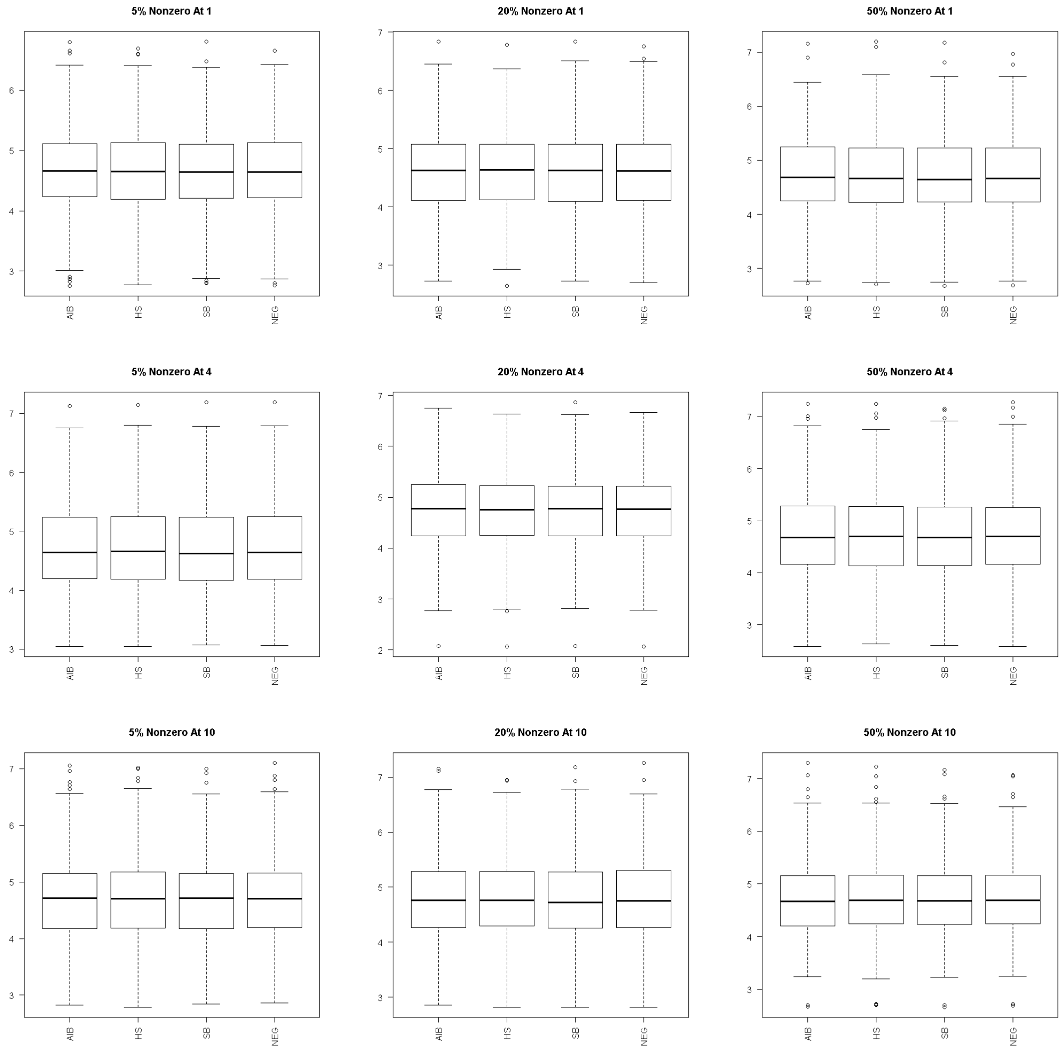

Figure 4.5: Boxplots of posterior means of σ_i's in the AIB, the HS, the SB and the NEG models when $\mathbf{\Sigma} = 10I + 10J$. The true σ_i's are $\sqrt{20}$. All dimensions are plotted in one box using 2000 datasets for each model. In all datasets, $p = 5$ and 20 observations are generated for each dimension. The signal size is fixed at 1 in the upper row, 4 in the middle row and 10 in the lower row. The nonzero percentage is fixed at 5 in the left column, 20 in the middle column and 50 in the right column.

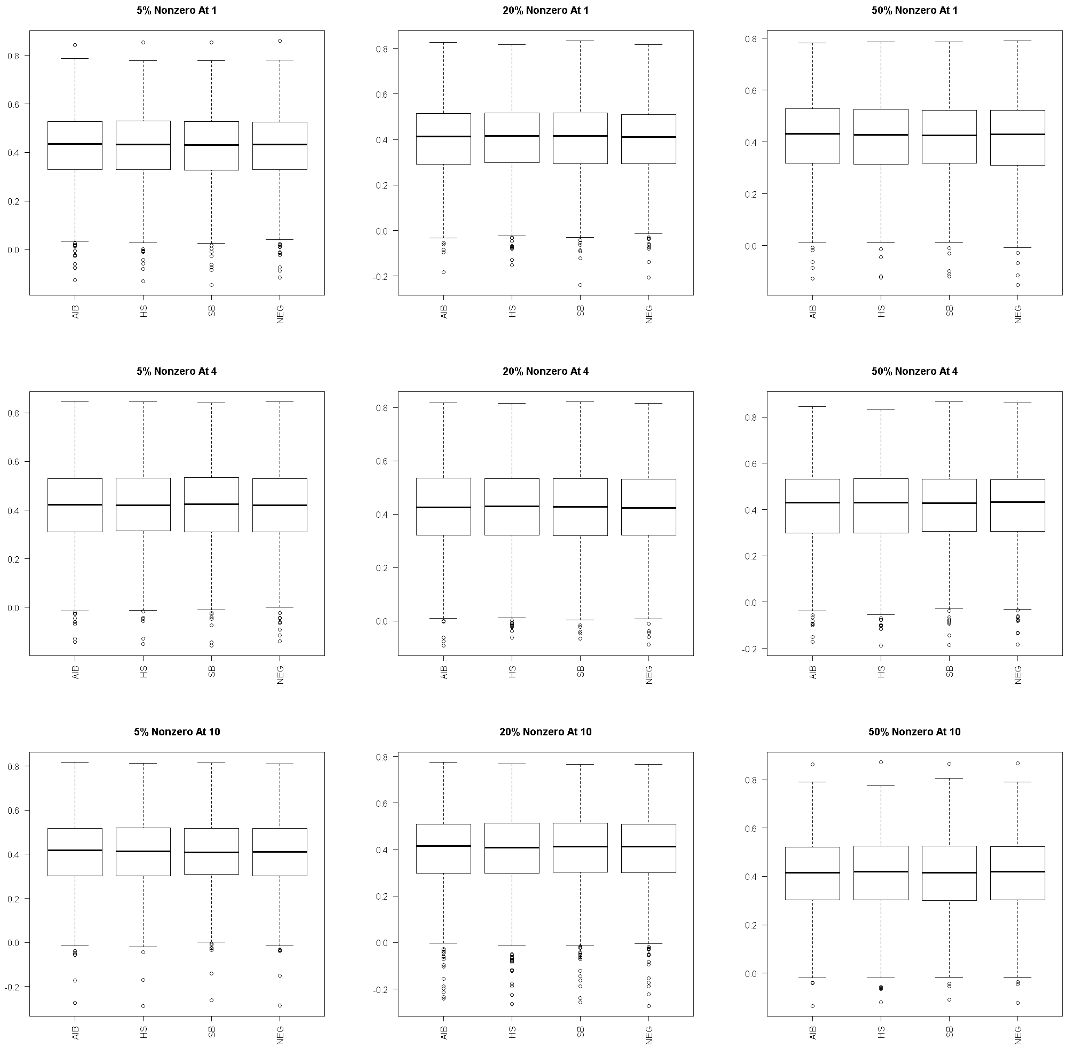

Figure 4.6: Boxplots of posterior means of correlations between different dimensions in the AIB, the HS, the SB and the NEG models. The true correlations between different dimensions are all 0.5. All dimensions are plotted in one box using 2000 datasets for each model. In all datasets, $p = 5$ and 20 observations are generated for each dimension. The signal size is fixed at 1 in the upper row, 4 in the middle row and 10 in the lower row. The nonzero percentage is fixed at 5 in the left column, 20 in the middle column and 50 in the right column.

Bibliography

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, ninth dover printing, tenth gpo printing edn. Dover, New York.

Armagan, A., Dunson, D. and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica* **23**, 119–143.

Avramov, D. and Zhou, G. (2010). Bayesian portfolio analysis. *Annual Review of Financial Economics* **2**, 25–47.

Baranchik, A. (1964). Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution. *Technical report.* Department of Statistics, Stanford University.

Barnard, J., McCulloch, R. and Meng, X. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica* **10**, 1281–1131.

Berger, J. (1980). A robust generalized Bayes estimator and confidence region for a multivariate normal mean. *The Annals of Statistics* **8**, 716–761.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis.* Springer.

Berger, J. O. and Bock, M. E. (1976). Combining independent normal mean estimation problems with unknown variances. *The Annals of Statistics* **4**, 642–648.

Borggaard, C. and Thodberg, H. H. (1992). Optimal minimum neural interpretation of spectra. *Analytical Chemistry* **64**, 545–551.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics* **24**, 2350–2383.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *The Annals of Mathematical Statistics* **42**, 855–903.

Brown, P. J. and Vannucci, M. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society* **60**, 627–641.

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

Chetelat, D. and Wells, M. T. (2012). Improved multivariate normal mean estimation with unknown covariance when p is greater than n. *The Annals of Statistics* **40**, 3137–3160.

Chevrier, T. and McCulloch, R. E. (2008). Using economic theory to build optimal portfolios. Working paper.

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics* **24**, 17–36.

Clyde, M. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistics Society* **62**, 681–698.

Clyde, M., DeSimone, H. and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association* **91**, 1197–1208.

DeMiguel, V., Garlappi, L. and Uppal, R. (2006). 1/N. In *EFA 2006 Zurich Meetings* (M. Spiegel (ed.)). Available at SSRN: http://ssrn.com/abstract=911512 or http://dx.doi.org/10.2139/ssrn.911512.

Eilers, P. H. C., Li, B. and Marx, B. D. (2009). Multivariate calibration with single index regression. *Chememetrics and Intelligent Laboratory Systems* **96**, 196–202.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Figueiredo, M. A. T. (2003). Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**, 1150–1159.

Fourdrinier, D., Strawderman, W. E. and Wells, M. T. (1998). On the construction of Bayes minimax estimators. *The Annals of Statistics* **26**, 660–671.

George, E. (1986). Minimax Multiple Shrinkage Estimation. *The Annals of Statistics* **14**, 188–205.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–747.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica* **7**, 339–373.

Gordy, M. B. (1998). A generalization of generalized beta distributions. Finance and Economics Discussion Series 1998-18, Board of Governors of the Federal Reserve System (U.S.).

Griffin, J. and Brown, P. (2005). Alternative prior distributions for variable selection with very many more variables than observations. *Technical report.* University of Warwick.

Griffin, J. E. and Brown, P. J. (2007). Bayesian adaptive lassos with non-convex penalization. Working paper.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* **5**, 171–188.

Hansen, M. H. and Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association* **96**, 746–774.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 361–379.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics* **32**, 1594–1649.

Jorion, P. (1986). Bayes-Stein estimation for portfolio analysis. *Journal of Financial and Quantitative Analysis* **21**, 279–292.

Judge, G. G. and Bock, M. E. (1978). *The Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics.* North-Holland Publishing Co., Amsterdam.

Konno, Y. (1990). Families of minimax estimators of matrix of normal means with unknown covariance matrix. *Japan Statistics Society* **20**, 191–201.

Konno, Y. (1991). On estimation of a matrix of normal means with unknown covariance matrix. *Journal of Multivariate Analysis* **36**, 44–55.

Konno, Y. (1992). *Improved estimation of matrix of normal mean and eigenvalues in the multivariate F-distribution.* PhD thesis. University of Tsukuba.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*, Springer Texts in Statistics. Springer.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g priors for Baysian variable selection. *Journal of the American Statistical Association* **103**, 410–423.

Lin, P. and Tsai, H. (1973). Generalized Bayes minimax estimators of the multivariate normal mean with unknown covariance matrix. *The Annals of Statistics* **1**, 142–145.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance* **7**, 77–91.

Miller, A. (2002). *Subset selection in regression.* CRC Press.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association* **83**, 1023–1032.

Park, T. and Dyk, D. A. (2009). Partially collapsed Gibbs Sampler: illustrations and applications. *Journal of Computational and Graphical Statistics* **19**, 283–305.

Polson, N. G. and Scott, J. G. (2009). Alternative global-local shrinkage priors using hypergeometric-beta mixtures. Working paper.

Polson, N. G. and Scott, J. G. (2010). On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis* **7**, 887–902.

Scheipl, F. and Kneib, T. (2008). Locally adaptive Bayesian P-splines with a normal-exponential-gamma prior. *Technical report.* Department of Statistics, University of Munich.

Shao, P. and Strawderman, W. (1994). Improving on the James-Stein Positive-Part Estimator. *The Annals of Statistics* **22**, 1517–1538.

Slater, L. J. (1960). *Confluent Hypergeometric Functions.* Cambridge University Press, New York.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1**, 197–206.

Stein, C. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *The Annals of Statistics* **9**, 1135–1151.

Strawderman, W. (1971). Proper Bayes Minimax Estimators of the Multivariate Normal Mean. *The Annals of Mathematical Statistics* **42**, 385–388.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society* **58**, 267–288.

Tsukuma, H. (2009). Generalized Bayes minimax estimation of the normal mean matrix with unknown covariance matrix. *Journal of Multivariate Analysis* **100**, 2296–2304.

Tu, J. and Zhou, G. (2010). Incorporating economic objectives into Bayesian priors: Portfolio choice under parameter uncertainty. *Journal of Financial and Quantitative Analysis* **45**, 959–986.

West, M. (2003). Bayesian factor regression models in the "large p, small n" paradigm. In *Bayesian Statistics* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (eds)), pp. 723–732, Oxford University Press.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayeisan Inference and Decision Techniques* pp. 233–243.

Zellner, A. and Chetty, V. K. (1965). Prediction and decision problems in regression models from the Bayesian point of view. *Journal of the American Statistical Association* **60**, 608–616.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadistica Y de Investigacion Operativa* **31**, 585–603.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society* **67**, 301–320.