A STUDY OF THE PREDICTION PERFORMANCE AND MULTIVARIATE EXTENSIONS OF THE HORSESHOE ESTIMATOR

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Yunfan Li

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Anindya Bhadra, Chair
Department of Statistics
Dr. Bruce A. Craig, Co-Chair
Department of Statistics
Dr. Jun Xie
Department of Statistics
Dr. Michael Zhu
Department of Statistics

Approved by:
Dr. Jun Xie
Graduate Chair

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

ABSTRACT

Li, Yunfan PhD, Purdue University, May 2019. A Study of the Prediction Performance and Multivariate Extensions of the Horseshoe Estimator. Major Professor: Anindya Bhadra and Bruce A. Craig.

The horseshoe prior has been shown to successfully handle high-dimensional sparse estimation problems. It both adapts to sparsity efficiently and provides nearly unbiased estimates for large signals. In addition, efficient sampling algorithms have been developed and successively applied to a vast array of high-dimensional sparse estimation problems. In this dissertation, we investigate the prediction performance of the horseshoe prior in sparse regression, and extend the horseshoe prior to two multivariate settings.

We begin with a study of the finite sample prediction performance of shrinkage regression methods, where the risk can be unbiasedly estimated using Stein's approach. We show that the horseshoe prior achieves an improved prediction risk over global shrinkage rules, by using a component-specific local shrinkage term that is learned from the data under a heavy-tailed prior, in combination with a global term providing shrinkage towards zero. We demonstrate improved prediction performance in a simulation study and in a pharmacogenomics data set, confirming our theoretical findings.

We then shift to extending the horseshoe prior to handle two high-dimensional multivariate problems. First, we develop a new estimator of the inverse covariance matrix for high-dimensional multivariate normal data. The proposed graphical horseshoe estimator has attractive properties compared to other popular estimators. The most prominent benefit is that when the true inverse covariance matrix is sparse, the graphical horseshoe estimator provides estimates with small information divergence from the sampling model. The posterior mean under the graphical horseshoe prior can also be almost unbiased under certain conditions. In addition to these theoretical results, we provide a full Gibbs sampler for implementation. The graphical horseshoe estimator compares favorably to existing techniques in simulations and in a human gene network data analysis.

In our second setting, we apply the horseshoe prior to the joint estimation of regression coefficients and the inverse covariance matrix in normal models. The computational challenge in this problem is due to the dimensionality of the parameter space that routinely exceeds the sample size. We show that the advantages of the horseshoe prior in estimating a mean vector, or an inverse covariance matrix, separately are also present when addressing both simultaneously. We propose a full Bayesian treatment, with a sampling algorithm that is linear in the number of predictors. Extensive performance comparisons are provided with both frequentist and Bayesian alternatives, and both estimation and prediction performances are verified on a genomic data set.

1. INTRODUCTION

In considering the normal means model, James and Stein (1961) proposed an admissible alternative to the classical least squares estimator. This method shrinks the least square estimates toward each other (i.e., the overall mean), consequently reducing overall risk under quadratic loss. As a result of this discovery, other shrinkage rules were proposed and gradually gained popularity because they were shown to outperform unbiased estimators with respect to many other loss functions that sum errors over all coordinates (e.g. sum of absolute errors) (Efron, 1975).

More recently, shrinkage rules have been widely applied to high-dimensional regression problems, as they reduce model complexity, as well as improve statistical risk properties. According to Fan and Li (2001), a good estimator is expected to shrink small estimated parameters (ideally to zero) in order to reduce model complexity, as well as provide a nearly unbiased estimate when the unknown parameter is large. Many shrinkage methods have gained popularity in high-dimensional data analysis, each with its own merits. For instance, ridge regression (Hoerl and Kennard, 1970) is one of the most prevailing shrinkage methods. Some other methods, including lasso (Tibshirani, 1996), adaptive lasso (Zou, 2006), and smoothly clipped absolute deviation (SCAD) regression (Fan and Li, 2001), simultaneously perform shrinkage and variable selection.

Bayesian methods naturally incorporate shrinkage through prior distributions, and a vast number of Bayesian models for high-dimensional estimation have been proposed. For example, the discrete spike-and-slab prior, which is a mixture of a point mass at zero and a nonnull density, can be represented as a shrinkage rule (Scott and Berger, 2006). However, exploring the posterior under discrete mixture priors is burdensome in high dimensions since block updates are generally not feasible. To avoid these posterior sampling issues, many continuous priors have been proposed.

These continuous priors are usually scale mixtures of normal distribution, and allow easy posterior sampling. For instance, the double-exponential prior was proposed in the 1990s (Carlin and Polson, 1991; Pericchi and Smith, 1992), and regained interest as a Bayesian alternative to lasso (Tibshirani, 1996). Another set of continuous priors combine a global shrinkage parameter with local shrinkage parameters, and these global-local shrinkage rules include the horseshoe prior described below.

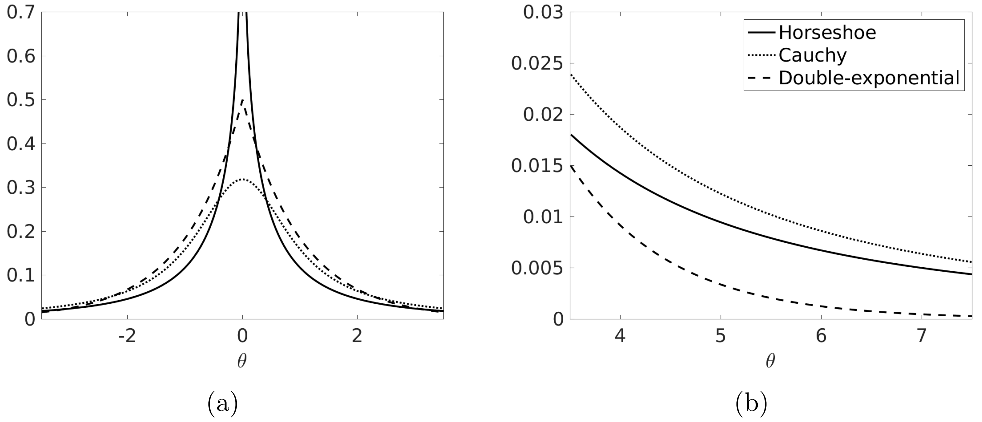
1.1 The horseshoe prior and its theoretical properties

This thesis focuses on the horseshoe prior for high-dimensional sparse problems. The horseshoe prior was introduced to provide a new approach to sparsity (Carvalho et al., 2010). The prior was initially proposed for use in the sparse normal means problem, $y_i|\theta_i \sim \text{Normal}(\theta_i, \sigma^2)$ and the n-dimensional mean vector θ is expected to be sparse. The horseshoe prior is in the family of multivariate scale mixtures of normals. Specifically, each θ_i is mixed over its own local scale parameter, and a scale parameter shared by all parameters, which is often referred to as the global shrinkage parameter. The horseshoe prior for the normal means problem can be written as:

$$\theta_i|\lambda_i \sim \text{Normal}(0, \lambda_i^2\tau^2), \quad \lambda_i \sim C^+(0,1), \quad \tau|\sigma \sim C^+(0,\sigma),$$

where $C^+(0,1)$ is a standard half-Cauchy distribution with probability density proportional to $1/\pi(1+x^2)$. Here, the λ_i's are the local scale parameters, and τ is the global scale parameter shared by all the θ_i's. A smaller τ results in a higher prior probability of shrinking all the θ_i's towards zero, while large λ_i for some θ_i's encourage less shrinking of the corresponding parameters. The induced prior on shrinkage is between zero and one, and has a mode near zero and one, hence the horseshoe name.

The Cauchy distribution has both a mode at zero and slow decaying tails. Consequently, the induced marginal prior on θ_i is both infinite near zero and has heavy tails decaying polynomially (Carvalho et al., 2010). By putting infinite density near zero, the horseshoe prior poses large prior mass near the true parameter when $\theta_i = 0$, and achieves fast convergence to the correct estimate of the sampling density in terms of Kullback-Leibler divergence in sparse problems. The horseshoe convergence is faster than priors with bounded density near zero, for instance, the double exponential prior. The heavy tails of the marginal prior induce robustness for large signals. More specifically, the horseshoe estimator is asymptotically unbiased when the signal is large. This unbiasedness can not be achieved by any prior with exponentially decaying tails (Carvalho et al., 2010). The fast convergence combined with the asymptotic unbiasedness make the horseshoe prior robust and highly adaptive in sparse problems.



(a)

(b)

Figure 1.1.: Comparison of the horseshoe, Cauchy, and double-exponential priors on θ, (a) near the origin and (b) in the tails.

Figure 1.1 plots the density of the horseshoe prior with the standard double-exponential and standard Cauchy densities. The horseshoe prior is spiked at zero, and has a polynomially decaying tail like the Cauchy density. The double-exponential tail decays much faster. Figure 1.2 shows the horseshoe induced prior and posterior density of the shrinkage factor $\kappa_i = 1/(1 + \lambda_i^2 \tau^2)$. Parts (a) and (b) show the prior density when $\tau^2 = 0.01$ and $\tau^2 = 0.1$, respectively. The spike near 1 encourages efficient shrinkage of all parameters, and the spike near 0 allows shrinkage of some parameters to be small. The shape of this prior is governed by the global shrinkage parameter τ^2. A small τ^2 shifts the density toward 1 and encourages shrinkage, and a large τ^2 puts higher density near 0. When $\tau^2 = 1$ or $n = 1$, the prior on κ

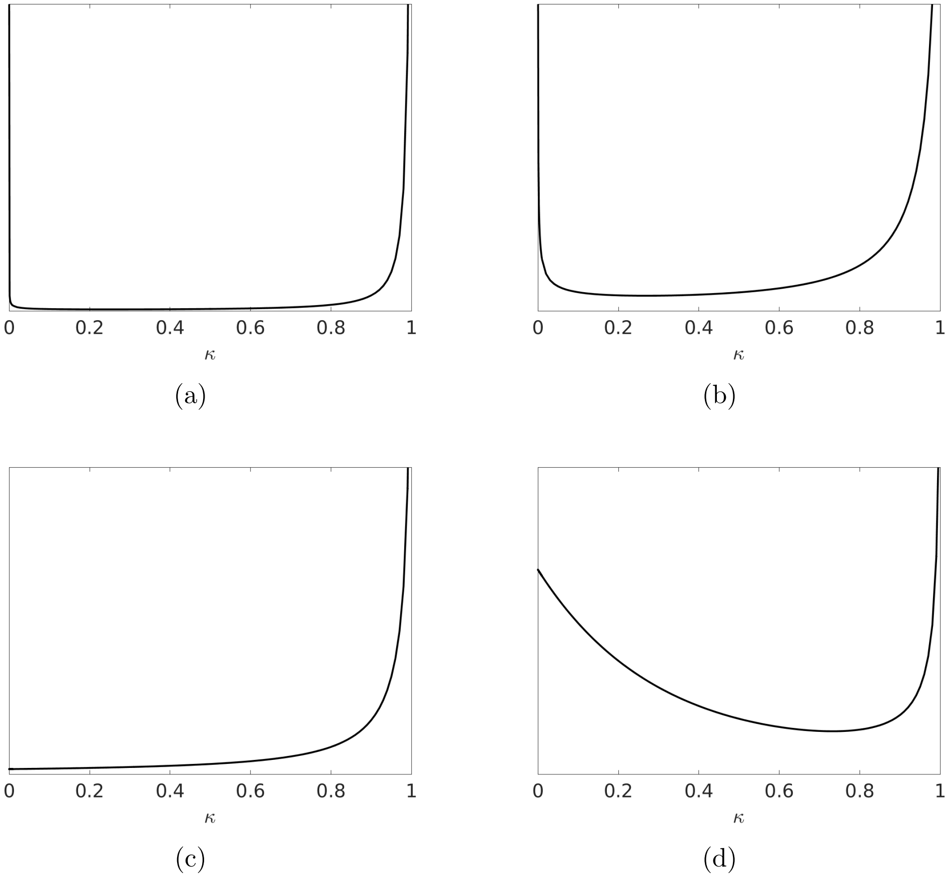


Figure 1.2.: The horseshoe implied density of the shrinkage factor $\kappa_i = 1/(1+\lambda_i^2\tau^2)$, *a priori* (a) when $\tau^2 = 0.01$, (b) when $\tau^2 = 0.1$, and *a posteriori* when (c) $y_i^2/\sigma^2 = 1$ and $\tau^2 = 0.1$, and (d) $y_i^2/\sigma^2 = 9$ and $\tau^2 = 0.1$.

becomes a Beta$(1/2, 1/2)$ distribution. Parts (c) and (d) show the posterior of κ when the observed signal strengths y_i^2/σ^2 are small and large, respectively. The posterior emphasizes heavy shrinkage when the observed value is relatively small, and suggests far less shrinkage when the observed value is large. More discussion on the shrinkage factor κ will be made in Chapter 2.

Many theoretical studies of the horseshoe prior concentrate on the normal means problem. For instance, Datta and Ghosh (2013) studied Bayes risk for the two-group normal means model under additive $0-1$ loss. They derived asymptotic Type I and Type II error rates, and proved that the Bayes risk induced by the horseshoe prior attains the risk of the Bayes oracle up to a constant provided the global shrinkage parameter is chosen to suit the sparsity of the data. Numerical examples using the full Bayes estimate confirmed their theoretical results. Ghosh et al. (2016) extended the results on Bayes risk to the general class of one-group priors, which includes the horseshoe, three parameter beta (Armagan et al., 2011), and normal-exponential-gamma (Griffin and Brown, 2010) priors. They further showed that under very mild assumption on underlying sparsity, the induced decisions using an empirical Bayes estimate of the global shrinkage parameter asymptotically attain the optimal Bayes risk.

van der Pas et al. (2014) showed that the horseshoe estimator achieves minimax quadratic risk when the true sparsity is known, and that the variance of the posterior distribution corresponding to the horseshoe prior has an upper bound of the order of the minimax risk. They also gave conditions under which the horseshoe estimator, combined with an empirical Bayes estimator of τ, still attains the minimax rate. van der Pas et al. (2017b) dropped the assumption of known sparsity level, and proved that the maximum marginal likelihood estimator (MMLE) is an effective estimator of the sparsity level. They also considered a hierarchical Bayes method, and showed that both MMLE and the hierarchical Bayes procedure adapt to the number p_n of nonzero means, and lead to minimax optimal estimation of the normal means.

In addition, van der Pas et al. (2017a) studied uncertainty quantification in the normal means model under the horseshoe prior. They showed that credible balls and marginal credible intervals, resulting from the horseshoe prior, have good frequentist coverage properties as well as optimal size, when the global sparsity level of the prior is set properly. They also showed that few zero parameters are falsely selected and most large signals are correctly selected under the horseshoe prior. However, most of the remaining parameters with moderate signals are not selected, constituting false negatives, so that model selection under the horseshoe prior is conservative.

1.2 Computation and other developments

In terms of computation, Makalic and Schmidt (2016) proposed a simple sampler for the horseshoe estimator. They used the scale mixture representation of the half-Cauchy distribution. That is, if $x^2|a \sim \text{InvGamma}(1/2, 1/a)$ and $a \sim \text{InvGamma}(1/2, 1)$, then $x \sim C^+(0, 1)$. By adding inverse gamma distributed auxiliary variables, a full Gibbs sampler is obtained for the horseshoe estimator. They also extended the sampler to other models and other priors, including horseshoe logistic regression, horseshoe negative binomial regression, and the horseshoe+ prior where there is one more half-Cauchy distributed hyperprior (Bhadra et al., 2017).

Bhattacharya et al. (2016) proposed a fast sampling method for normal scale mixture priors, including the horseshoe prior, in high-dimensional regression. Normal scale mixture priors are popular in sparse regression problems because of their computational efficiency and simplicity, but computation can still be intense in high dimensions. For earlier algorithms that relied on Cholesky factorization, computational complexity is $O(p^3)$ for dimension p. The authors proposed an algorithm that utilizes matrix multiplication and linear system solutions, and achieves $O(n^2p)$ complexity, where n is the sample size, an improvement from $O(p^3)$ when $p > n$. This fast sampling method enables computation of normal scale mixture priors in higher dimensions.

The horseshoe prior has been used to solve many problems beyond normal means and linear models. Piironen and Vehtari (2017) and Wei (2017) both applied the horseshoe prior in logistic regression. Piironen and Vehtari (2017) introduced a small regularization to the largest nonzero coefficients to deal with data separation in logistic regressions. Wei (2017) showed that under the horseshoe prior, the posterior for the coefficients concentrates around the truth with respect to the Euclidean norm, and the procedure selects the correct predictors like a point-mass prior, under some conditions. Magnusson et al. (2016) considered supervised classification with a large number of classes, with the application of classifying documents. They proposed a full Bayesian method where all classes are independently modeled using binary probit models, combined with the horseshoe prior on coefficients of classes on semantic meanings of topics. The horseshoe prior enables higher accuracy in prediction, easier interpretation of coefficients, and the ability to easily handle many additional covariates in the model. Peltola et al. (2014) compared the performance of the normal, Laplace, and horseshoe priors in survival analysis, and found that the horseshoe prior has the best predictive performance. The horseshoe prior also shrinks strong predictors less than the other priors.

Bhadra et al. (2016) considered estimating non-linear low-dimensional functions of normal means in the sparse high-dimensional normal means model. The difficulty in this problem is that non-informative priors could become highly informative in the non-linear transformation from high-dimensional space to a low-dimensional space. The authors showed that the horseshoe prior is a good candidate to be the default prior distribution in this problem because of its slow decaying tails, satisfying the regularly varying condition. The slow decaying tails are preserved in the many-to-one functions in the transformation, making the prior on low-dimensional space non-informative as well. In addition, the horseshoe prior is also more appropriate than some other heavy-tailed priors (e.g. a multivariate-t prior) in this problem because the global-local shrinkage allows heavier shrinkage towards zero. In addition, readers looking for a more detailed survey could go to Bhadra et al. (2019), where the authors discussed theoretical optimality, computation, and methodological developments of the horseshoe prior.

1.3 Outline of Thesis

Bayes risk under additive $0 - 1$ loss, quadratic risk, and uncertainty quantification have been studied for the horseshoe prior in the normal means model. Beyond this simple model, the properties of this global-local priors remain largely unexplored. Chapter 2 considers prediction risk of horseshoe regression and compares it to some global shrinkage rules. Polson and Scott (2012) proposed a unifying framework of studying regularized regressions, and showed connections among some disparate methods, including ridge regression, principal component regression, partial least squares, and the g-prior. They also gave intuition as well as presented examples of when full Bayesian methods should work better. In Chapter 2, we consider the framework proposed by Polson and Scott (2012), and provide expressions of Stein's Unbiased Risk Estimator (SURE) of prediction risk for a few regularized regressions, including horseshoe regression. We show certain cases where horseshoe regression is expected to perform better, in terms of prediction risk, than regularized regressions with global shrinkage only, and demonstrate these cases with numerical examples.

The horseshoe prior has been applied to a wide variety of problems including generalized regressions, classification and survival analysis, where the horseshoe prior density yields good results. Chapter 3 extends the horseshoe prior to estimate the inverse covariance matrix in zero-mean Gaussian graphical models. This problem can be expressed as a set of partial regression problems (Pourahmadi, 2011). Therefore the horseshoe prior, being efficient in solving sparse regressions, is expected to work well in sparse inverse covariance estimation. The primary challenge of applying the horseshoe prior in this problem is maintaining symmetric and positive definite samples of the inverse covariance matrix. Chapter 3 proposes a full-Bayesian method that applies the horseshoe prior on the elements of the inverse covariance matrix. Because the horseshoe prior has heavy tails and is peaked at zero, this novel method provides posterior estimates with small information divergence from the distribution with the true parameters, and is tail robust.

Chapter 4 then considers multivariate regressions where both the coefficients and the inverse covariance matrix need to be estimated. This framework is known as the seemingly unrelated regression (Zellner, 1986). Some applications, such as eQTL analysis, have high dimensional data and expect both the mean/regression coefficients and the inverse covariance to be sparse. We apply the horseshoe prior to both the mean/regression coefficients and elements in the inverse covariance matrix, to induce efficient shrinkage towards sparsity. Chapter 4 gives an MCMC sampling scheme to this model, and demonstrates that application of the horseshoe prior results in better estimation, prediction, and variable selection than other shrinkage rules for this problem.

2. PREDICTION RISK FOR THE HORSESHOE REGRESSION

Prediction using shrinkage regression techniques such as ridge regression (Hoerl and Kennard, 1970) and principal components regression, or PCR (Jolliffe, 1982), remain popular in high-dimensional problems. They enjoy a number of advantages over selection-based methods, such as the lasso (Tibshirani, 1996), and comfortably outperform them in predictive performance in certain situations. Prominent among these is when the predictors are correlated and the resulting lasso estimate is unstable, but ridge or PCR estimates are not (Hastie et al., 2009). Polson and Scott (2012) showed, following a representation originally devised by Frank and Friedman (1993), that many commonly used high-dimensional shrinkage regression estimates, such as the estimates of ridge regression, regression with the g-prior (Zellner, 1986) and PCR, can be viewed as posterior means under a unified framework of a "global" shrinkage prior on the regression coefficients that are suitably orthogonalized. They went on to demonstrate these global shrinkage regression models suffer from two major difficulties: (i) the amount of relative shrinkage is monotone in the singular values of the design matrix and (ii) the amount of shrinkage does not depend on values of the response variables. Both of these factors can contribute to poor out of sample prediction performance, which they demonstrated numerically.

Polson and Scott (2012) further provided numerical evidence that both of these difficulties mentioned above can be resolved by allowing a "local," component-specific shrinkage term that can be learned from the data, in conjunction with a global shrinkage parameter as used in ridge or PCR, giving rise to the so-called "global-local" shrinkage regression models. Specifically, Polson and Scott (2012) demonstrated by simulations that using the horseshoe prior of Carvalho et al. (2010) on the regression coefficients performed well over a variety of competitors in terms of predictive performance, including the lasso, ridge, PCR and sparse partial least squares (Chun and Keles, 2010). However, a theoretical investigation of the conditions required for a global-local shrinkage regression model to outperform a global shrinkage regression model such as ridge or PCR in terms of predictive performance has been lacking. The goal of this work is to bridge this methodological and theoretical gap by developing formal tools for comparing the predictive performances of shrinkage regression methods.

Developing a formal measure to compare predictive performance of competing regression methods is important in both frequentist and Bayesian settings. This is because the frequentist tuning parameter, or the Bayesian hyper-parameter, can then be chosen to minimize the prediction risk, if prediction of future observations is the main modeling goal. A measure of quadratic risk for prediction in regression models can be obtained either through model-based covariance penalties or through nonparametric approaches. Examples of covariance penalties include Mallow's C_p (Mallows, 1973), Akaike's information criterion (Akaike, 1974), risk inflation criterion (Foster and George, 1994) and Stein's unbiased risk estimate or SURE (Stein, 1981). Nonparametric penalties include the generalized cross validation of Craven and Wahba (1978), which has the advantage of being model free but usually produces a prediction error estimate with high variance (Efron, 1983). The relationship between the covariance penalties and nonparametric approaches were further explored by Efron (2004), who showed the covariance penalties to be a Rao-Blackwellized version of the nonparametric penalties. Thus, Efron (2004) concluded that model-based penalties such as SURE or Mallow's C_p (the two coincide for models where the fit is linear in the response variable) offer substantially lower variance in estimating the prediction error, assuming, of course, the model is true. From a computational perspective, calculating SURE, when it is explicitly available, is substantially less burdensome than performing cross validation, which usually requires several Monte Carlo replications. Furthermore, SURE, which is a measure of quadratic risk in prediction, also has connections with the Kullback-Leiber risk for the predictive density (George et al., 2006).

Given these advantages enjoyed by SURE, we devise a general, explicit and numerically stable technique for computing SURE for regression models that can be employed to compare the performances of global as well as global-local shrinkage regressions. The key technique to our innovation is an orthogonalized representation first employed by Frank and Friedman (1993), which results in particularly simple and numerically stable formulas for SURE. Using the developed tools for SURE, we demonstrate that the suitable conditions for success of a global-local regression model over global regression models in prediction arise when a certain eigen-sparse structure is present in the design. Specifically, our major finding is that when a certain principal component corresponding to a low singular value of the design matrix is a strong predictor of the outcomes, global shrinkage methods necessarily shrink these components too much, whereas global-local models do not. This results in a substantially increased SURE for global regression over global-local regression, explaining why global-local shrinkage can overcome the two major difficulties encountered by global shrinkage regression methods.

The rest of the article is organized as follows. In Section 2.1, we demonstrate how several standard shrinkage regression estimates can be reinterpreted as posterior means in an orthogonalized representation of the design matrix. Using this representation, we derive explicit expressions for SURE for global and global-local shrinkage regressions in Sections 2.2 and 2.3 respectively. The main theoretical findings are presented in the form of three theorems in Section 2.3 which formally demonstrate global-local shrinkage regressions to have lower prediction risk over global shrinkage methods in a sparse setting. A simulation study is presented in 2.4 and prediction performance of several competing approaches are assessed in a pharmacogenomics data set in Section 2.5. We conclude by pointing out some possible extensions of the current work in Section 2.6.

2.1 Shrinkage regression estimates as posterior mean

Consider the high-dimensional regression model

$$y = X\beta + \epsilon, \tag{2.1}$$

where $y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times p}, \beta \in \mathbb{R}^p$ and $\epsilon \sim \text{Normal}(0, \sigma^2\mathbf{I}_n)$ with $p > n$. Let $X = UDW^T$ be the singular value decomposition of the design matrix. Let $\text{Rank}(D) = \min(n,p) = n$ where $D = \text{diag}(d_1, \ldots, d_n)$ with $d_1 \geq \ldots \geq d_n > 0$. Define $Z = UD$ and $\alpha = W^T\beta$. Then the regression problem can be reformulated as:

$$y = Z\alpha + \epsilon. \tag{2.2}$$

Define the ordinary least squared (OLS) estimate of α to be $\hat{\alpha} = (Z^TZ)^{-1}Z^Ty = D^{-1}U^Ty$. Following the original results by Frank and Friedman (1993), several authors have used the well-known orthogonalization technique (Polson and Scott, 2012; Clyde et al., 1996; Denison and George, 2012) to demonstrate that the estimates of many shrinkage regression methods can be expressed in terms of the posterior mean of the "orthogonalized" regression coefficients α under the following hierarchical model:

$$(\hat{\alpha}_i \mid \alpha_i, \sigma^2) \overset{ind}{\sim} \text{Normal}(\alpha_i, \sigma^2 d_i^{-2}), \tag{2.3}$$

$$(\alpha_i \mid \sigma^2, \tau^2, \lambda_i^2) \overset{ind}{\sim} \text{Normal}(0, \sigma^2\tau^2\lambda_i^2), \tag{2.4}$$

with $\sigma^2, \tau^2 > 0$. The global term τ controls the amount of shrinkage and the fixed λ_i^2 terms depend on the method at hand. Given λ_i and τ, the estimate for β under the global shrinkage prior, denoted by $\tilde{\beta}$, can be expressed in terms of the posterior mean estimate for α as follows:

$$\tilde{\alpha}_i = \frac{\tau^2\lambda_i^2 d_i^2}{1+\tau^2\lambda_i^2 d_i^2}\hat{\alpha}_i, \quad \tilde{\beta} = \sum_{i=1}^{n} \tilde{\alpha}_i w_i, \tag{2.5}$$

where $\tilde{\alpha}_i = \mathrm{E}(\alpha_i \mid \tau, \lambda_i^2, X, y)$, w_i is a $p \times 1$ vector and is the ith column of the $p \times n$ matrix W and the term $\tau^2\lambda_i^2 d_i^2/(1+\tau^2\lambda_i^2 d_i^2) \in (0,1)$ is the shrinkage factor. The expression from Equation (2.5) makes it clear that it is the orthogonalized OLS estimates $\hat{\alpha}_i$s that are shrunk and helps in interpretation. A new result that we derive in the next sections is that this orthogonalized representation is also particularly suitable for calculating the prediction risk. The reason for this is tied to the independence assumption that is now feasible in Equations (2.3) and (2.4). To give a few concrete examples, we note below that several popular shrinkage regression models fall under the framework of Equations (2.3-2.4):

1. For ridge regression, $\lambda_i^2 = 1$ for all i and we have $\tilde{\alpha}_i = \tau^2 d_i^2 \hat{\alpha}_i/(1+\tau^2 d_i^2)$.
2. For K component PCR λ_i^2 is infinite for the first K components and then 0. Thus, $\tilde{\alpha}_i = \hat{\alpha}_i$ for $i = 1, \ldots, K$ and $\tilde{\alpha}_i = 0$ for $i = K+1, \ldots, n$.
3. For regression with g-prior, $\lambda_i^2 = d_i^{-2}$ and we have $\tilde{\alpha}_i = \tau^2\hat{\alpha}_i/(1+\tau^2)$ for $i = 1, \ldots, n$.

This shows the amount of relative shrinkage $\tilde{\alpha}_i/\hat{\alpha}_i$ is constant in d_i for PCR and g-prior and is monotone in d_i for ridge regression. In none of these cases it depends on the OLS estimate $\hat{\alpha}_i$ (consequently, on y). In the next section we quantify the effect of this behavior on the prediction risk.

2.2 Stein's unbiased risk estimate for global shrinkage regression

Define the fit $\tilde{y} = X\tilde{\beta} = Z\tilde{\alpha}$, where $\tilde{\alpha}$ is the posterior mean of α. As noted by Efron (2004), the fitted risk is an underestimation of the prediction risk, and SURE for prediction is defined as:

$$R = ||y - \tilde{y}||^2 + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial \tilde{y}_i}{\partial y_i},$$

where the $\sum_{i=1}^{n}(\partial\tilde{y}_i/\partial y_i)$ term is also known as the "degrees of freedom." By Tweedie's formula (Masreliez, 1975; Pericchi and Smith, 1992) that relates the posterior mean with the marginals; we have for a Gaussian model of Equations (2.3-2.4) that: $\tilde{\alpha} = \hat{\alpha} + \sigma^2 D^{-2}\nabla_{\hat{\alpha}} \log m(\hat{\alpha})$, where $m(\hat{\alpha})$ is the marginal for $\hat{\alpha}$. Noting $y = Z\hat{\alpha}$ yields $\tilde{y} = y + \sigma^2 U D^{-1}\nabla_{\hat{\alpha}} \log m(\hat{\alpha})$. Using the independence of α_is, the formula for SURE becomes

$$R = \sigma^4 \sum_{i=1}^{n} d_i^{-2} \left\{ \frac{\partial}{\partial \hat{\alpha}_i} \log m(\hat{\alpha}_i) \right\}^2 + 2\sigma^2 \sum_{i=1}^{n} \left\{ 1 + \sigma^2 d_i^{-2} \frac{\partial^2}{\partial \hat{\alpha}_i^2} \log m(\hat{\alpha}_i) \right\}. \tag{2.6}$$

Thus, the prediction risk for shrinkage regression can be quantified in terms of the first two derivatives of the log marginal for $\hat{\alpha}$. Integrating out α_i from Equations (2.3-2.4) yields in all these cases,

$$(\hat{\alpha}_i \mid \sigma^2, \tau^2, \lambda_i^2) \overset{ind}{\sim} \text{Normal}(0, \sigma^2(d_i^{-2} + \tau^2\lambda_i^2)).$$

Thus the marginal is given by

$$m(\hat{\alpha}) \propto \prod_{i=1}^{n} \exp\left\{ - \frac{\hat{\alpha}_i^2/2}{\sigma^2(d_i^{-2} + \tau^2\lambda_i^2)} \right\},$$

which gives

$$\frac{\partial \log m(\hat{\alpha}_i)}{\partial \hat{\alpha}_i} = \frac{-\hat{\alpha}_i}{\sigma^2(d_i^{-2} + \tau^2\lambda_i^2)}; \qquad \frac{\partial^2 \log m(\hat{\alpha}_i)}{\partial \hat{\alpha}_i^2} = \frac{-1}{\sigma^2(d_i^{-2} + \tau^2\lambda_i^2)}. \tag{2.7}$$

Therefore, Equation (2.6) reduces to the following expression for SURE for shrinkage regressions $R = \sum_{i=1}^{n} R_i$, where,

$$R_i = \frac{\hat{\alpha}_i^2 d_i^2}{(1 + \tau^2\lambda_i^2 d_i^2)^2} + 2\sigma^2 \frac{\tau^2\lambda_i^2 d_i^2}{(1 + \tau^2\lambda_i^2 d_i^2)}. \tag{2.8}$$

From a computational perspective, the expression in Equation (2.8) is attractive, because it avoids costly matrix computations. For a given σ one can choose τ to minimize the prediction risk, which amounts to a one-dimensional optimization. Note that in our notation, $d_1 \geq d_2 \ldots \geq d_n > 0$. Clearly, this is the risk when when λ_is are fixed and finite (e.g., ridge regression). For K component PCR, only the first K terms appear in the sum. The d_i terms are features of the design matrix X and one may try to control the prediction risk by varying τ. When $\tau \to \infty$, $R \to 2n\sigma^2$, the risk of prediction with ordinary least squares (unbiased). When $\tau \to 0$, we get the mean-only model (zero variance): $R \to \sum_{i=1}^{n} \hat{\alpha}_i^2 d_i^2$. Regression models with $\tau \in (0, \infty)$ represent a bias-variance tradeoff. Following are the two major difficulties of global shrinkage regression.

1. Note from the first term of Equation (2.8) the risk is increased by those components for which $\hat{\alpha}_i^2 d_i^2$ is large. Choosing a large τ alleviates this problem, but at the expense of an R_i of $2\sigma^2$ even for components for which $\hat{\alpha}_i^2 d_i^2$ is small (due to the second term in Equation (2.8)). Thus, it might be beneficial to *differentially minimize* the effect of the components for which $\hat{\alpha}_i^2 d_i^2$ is large, while ensuring those for which $\hat{\alpha}_i^2 d_i^2$ is small make a contribution less than $2\sigma^2$ to risk. Yet, regression models with λ_i fixed, such as ridge, PCR, regression with g-priors, provide no mechanism for achieving this, since the relative shrinkage, defined as the ratio $\tilde{\alpha}_i / \hat{\alpha}_i$, equals $\tau^2 \lambda_i^2 d_i^2 / (1 + \tau^2 \lambda_i^2 d_i^2)$, and is solely driven by a single quantity τ.

2. Note also from Equation (2.5) that the relative shrinkage for $\hat{\alpha}_i$ is monotone in d_i; that is, those $\hat{\alpha}_i$ corresponding to a smaller d_i are necessarily shrunk more (in a relative amount). This is only sensible in the case where one has reasons to believe the low variance eigen-directions (i.e., principal components) of the design matrix are not important predictors of the response variables, an assumption that can be violated in real data (Polson and Scott, 2012).

In the light of these two problems, we proceed to demonstrate that putting a heavy-tailed prior on λ_is, in combination with a suitably small value of τ to enable global-local shrinkage can resolve both these issues. The intuition behind this is that a small value of a *global* parameter τ enables shrinkage towards zero for all the components while the heavy tails of the *local* or component-specific λ_i terms ensure the components with large values of $\hat{\alpha}_i d_i$ are not shrunk too much, and allow the λ_i terms to be learned from the data. Simultaneously ensuring both of these factors helps in controlling the prediction risk for both the noise as well as the signal terms.

2.3 Stein's unbiased risk estimate for the horseshoe regression

The global-local shrinkage regression of Polson and Scott (2012) extends the global shrinkage regression models of the previous section by putting a local (component-specific), heavy-tailed prior on the λ_i terms that allow these terms to be learned from the data, in addition to a global τ. The model equations become:

$$(\hat{\alpha}_i \mid \alpha_i, \sigma^2) \overset{ind}{\sim} \text{Normal}(\alpha_i, \sigma^2 d_i^{-2}), \tag{2.9}$$

$$(\alpha_i \mid \sigma^2, \tau^2, \lambda_i^2) \overset{ind}{\sim} \text{Normal}(0, \sigma^2\tau^2\lambda_i^2), \tag{2.10}$$

$$\lambda_i \overset{ind}{\sim} p(\lambda_i), \tag{2.11}$$

with $\sigma^2, \tau^2 > 0$. Improved mean square error over competing approaches in regression has been empirically observed by Polson and Scott (2012) with independent, standard half-Cauchy priors on λ_is. The intuitive explanation for this improved performance in a normal means model is that a heavy tailed prior on λ_i leaves the large α_i terms of Equation (2.10) un-shrunk in the posterior, whereas the global τ term provides shrinkage towards zero for all components Polson and Scott (2012); Bhadra et al. (2017); Carvalho et al. (2010). However, no explicit formulation of the prediction risk under global-local shrinkage is available so far and we explicitly demonstrate below the heavy-tailed priors λ_i terms, in addition to a global τ, can be beneficial in controlling the overall prediction risk.

Let $\lambda_i \overset{ind}{\sim} C^{+}(0,1)$, i.e., a standard half-Cauchy. Under the model of Equations (2.9-2.11), after integrating out α_i from the first two equations, we have,

$$(\hat{\alpha}_i \mid \sigma^2, \tau^2, \lambda_i^2) \overset{ind}{\sim} \text{Normal}(0, \sigma^2(d_i^{-2} + \tau^2\lambda_i^2)).$$

We have, $p(\lambda_i) \propto 1/(1+\lambda_i^2)$, the density of a standard half-Cauchy. Thus, the marginal of $\hat{\alpha}$, denoted by $m(\hat{\alpha})$, is given up to a constant of proportionality by

$$\begin{aligned} m(\hat{\alpha}) \propto & \prod_{i=1}^{n} \int_0^\infty \text{Normal}(\hat{\alpha}_i \mid 0, \sigma^2(d_i^{-2} + \tau^2\lambda_i^2)) p(\lambda_i) d\lambda_i \\ \propto & (2\pi\sigma^2)^{-n/2} \prod_{i=1}^{n} \int_0^\infty \exp\left\{ -\frac{\hat{\alpha}_i^2 d_i^2/2}{\sigma^2(1+\tau^2 d_i^2 \lambda_i^2)} \right\} \frac{d_i}{(1+\tau^2 d_i^2\lambda_i^2)^{1/2}} \frac{1}{1+\lambda_i^2} d\lambda_i. \end{aligned} \tag{2.12}$$

We now show that this integral involves the normalizing constant of a compound confluent hypergeometric distribution. We need the following result from Gordy (1998).

Proposition 2.3.1 *(Gordy, 1998). The compound confluent hypergeometric (CCH) density is given by*

$$\text{CCH}(x; p, q, r, s, \nu, \theta) = \frac{x^{p-1}(1-\nu x)^{q-1}\{\theta + (1-\theta)\nu x\}^{-r} \exp(-sx)}{B(p,q)H(p,q,r,s,\nu,\theta)}, \quad \textit{for} \quad 0 < x < 1/\nu,$$

for $p > 0$, $q > 0$, $r \in \mathbb{R}$, $s \in \mathbb{R}$, $0 \le \nu \le 1$ and $\theta > 0$, where $B(p,q)$ is the beta function and the function $H(\cdot)$ is given by

$$H(p,q,r,s,\nu,\theta) = \nu^{-p}\exp(-s/\nu)\Phi_1(q, r, p+q, s/\nu, 1-\theta),$$

where Φ_1 is the confluent hypergeometric function of two variables, given by

$$\Phi_1(\alpha, \beta, \gamma, \delta, \epsilon) = \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \frac{(\alpha)_m(\beta)_n}{(\gamma)_{m+n} m! n!} \delta^m \epsilon^n, \tag{2.13}$$

where $(a)_k$ denotes the rising factorial with $(a)_0 = 1, (a)_1 = a$ and $(a)_k = (a + k - 1)(a)_{k-1}$.

We present our first result in the following theorem and show that the marginal $m(\hat{\alpha})$ and all its derivatives lend themselves to a series representation in terms of the first and second moments of a random variable that follows a CCH distribution.

Theorem 2.3.1 *Denote $m'(\hat{\alpha}_i) = (\partial/\partial\hat{\alpha}_i)m(\hat{\alpha}_i)$ and $m''(\hat{\alpha}_i) = (\partial^2/\partial\hat{\alpha}_i^2)m(\hat{\alpha}_i)$. Then,*

1. *SURE for the global-local shrinkage regression model defined by Equations (2.9-2.11) is given by $R = \sum_{i=1}^n R_i$, where the component-wise SURE R_i is given by*

$$R_i = 2\sigma^2 - \sigma^4 d_i^{-2} \left\{ \frac{m'(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} \right\}^2 + 2\sigma^4 d_i^{-2} \frac{m''(\hat{\alpha}_i)}{m(\hat{\alpha}_i)}. \tag{2.14}$$

2. *Under independent standard half-Cauchy prior on λ_is, for the second and third terms in Equation (2.14) we have:*

$$\frac{m'(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} = -\frac{\hat{\alpha}_i d_i^2}{\sigma^2}\mathrm{E}(Z_i), \quad \textit{and} \quad \frac{m''(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} = -\frac{d_i^2}{\sigma^2}\mathrm{E}(Z_i) + \frac{\hat{\alpha}_i^2 d_i^4}{\sigma^4}\mathrm{E}(Z_i^2),$$

where $(Z_i \mid \hat{\alpha}_i, \sigma, \tau)$ follows a $\mathrm{CCH}(p = 1, q = 1/2, r = 1, s = \hat{\alpha}_i^2 d_i^2/2\sigma^2, v = 1, \theta = 1/\tau^2 d_i^2)$ distribution.

A proof is given in Appendix A.1. Theorem 2.3.1 provides a computationally tractable mechanism for calculating SURE for global-local shrinkage regressions in terms of the moments of CCH random variables. Gordy (1998) provides a simple formula for all integer moments of CCH random variables. Specifically, he shows if $X \sim \mathrm{CCH}(x; p, q, r, s, \nu, \theta)$ then

$$\mathrm{E}(X^k) = \frac{(p)_k}{(p+q)_k} \frac{H(p + k, q, r, s, \nu, \theta)}{H(p, q, r, s, \nu, \theta)}, \tag{2.15}$$

for integers $k \geq 1$. Moreover, as demonstrated by Gordy (1998), these moments can be numerically evaluated quite easily over a range of parameter values and calculations remain very stable. A consequence of this explicit formula for SURE is that the global shrinkage parameter τ can now be chosen to minimize SURE by performing a one-dimensional optimization. Perhaps more importantly, we can use the expression from Theorem 2.3.1 to understand the performance of global-local shrinkage regression for the signal and the noise terms. First we treat the case when $\hat{\alpha}_i d_i$ is large. We have the following result.

Theorem 2.3.2 *Define $s_i = \hat{\alpha}_i^2 d_i^2/2\sigma^2$. When $s_i \gg 1$, both $m''(\hat{\alpha}_i)/m(\hat{\alpha}_i)$ and $[m'(\hat{\alpha}_i)/m(\hat{\alpha}_i)]^2$ are $O(1/\hat{\alpha}_i^2)$ and therefore, the contributions of the second and the third terms to R_i in Equation (2.14) is $O(1/\hat{\alpha}_i^2 d_i^2)$. Consequently, the component-wise SURE $R_i \approx 2\sigma^2$.*

A proof is given in Appendix A.2. An intuitive explanation of this result is that component-specific shrinkage is feasible in a global-local regression model due to the heavy-tailed λ_i terms, which prevents the signal terms from getting shrunk too much and consequently, making a large contribution to the prediction risk due to a large bias. With just a global parameter τ, this component-specific shrinkage is not possible. A comparison of R_i resulting from Theorem 2.3.2 with that from Equation (2.8) demonstrates using global-local shrinkage, we can rectify a major shortcoming of global shrinkage regression, in that the terms with large s_i do not make a large contribution to the prediction risk. Next, for the case when $\hat{\alpha}_i d_i$ is small, we have the following result.

Theorem 2.3.3 *Define $s_i = \hat{\alpha}_i^2 d_i^2/2\sigma^2$. Then the following statements are true.*

1. *The component-wise SURE R_i is an increasing function of s_i in the interval $[0, 1]$ for any fixed τ.*

2. *When $s_i = 0$, the component-wise SURE R_i is a monotone increasing function of τ, and is bounded in the interval $(0, 2\sigma^2/3]$ when $\tau^2 d_i^2 \in (0, 1]$.*

A proof is given in Appendix A.3. This theorem establishes that: (i) the terms with smaller s_i in the interval $[0, 1]$ contribute less to the risk, with the minimum achieved at $s_i = 0$ (these terms can be thought of as the noise terms in an eigen-sparse regression problem) and (ii) if τ is chosen to be sufficiently small, the terms for which $s_i = 0$ has an upper bound of risk at $2\sigma^2/3$. Note that the OLS estimator has risk $2\sigma^2$ for these terms. At $s_i = 0$, the PCR risk is either 0 or $2\sigma^2$, depending on whether the term is or is not included. The ridge regression risk at $s_i = 0$ is an increasing function of the global shrinkage parameter τ and thus, it might make sense to choose a small τ if all s_i terms were small. However, in the presence of some s_i terms that are large, ridge regression cannot choose a very small τ, since the large s_i terms will then be heavily shrunk and contribute too much to the risk. This is not the case with global-local shrinkage regression methods, which can still choose a small τ to mitigate the risk from the noise terms and rely on the heavy-tailed λ_i terms to ensure large signals are not shrunk too much. Consequently, the ridge regression risk is usually larger than the global-local regression risk even for very small s_i terms, when some terms with large s_i are present along with mostly noise terms.

Theorem 2.3.3 also establishes that the maximum risk is achieved at $\hat{\alpha}_i^2 d_i^2 = 2\sigma^2$ in the region $\hat{\alpha}_i^2 d_i^2 \leq 2\sigma^2$ for global-local shrinkage regression. In fact, numerical integration using a half-Cauchy prior on λ_i shows that SURE for global-local regression is smaller than $2\sigma^2$ for any fixed τ over the entire region. We verify these assertions via simulations in the next section.

To summarize the theoretical findings, Theorem 2.3.2 together with Theorem 2.3.3 establishes that global-local shrinkage regression is effective in handling both very large and very small values of $\hat{\alpha}_i^2 d_i^2$. Specifically, Theorem 2.3.3 asserts that a small enough τ shrinks the noise terms towards zero, minimizing their contribution to risk. Whereas, according to Theorem 2.3.2, the heavy tails of the Cauchy priors for the λ_i terms ensure the large signals are not shrunk too much and ensures a risk of $2\sigma^2$ for these terms, which is an improvement over purely global methods of shrinkage.

2.4 Numerical examples

We simulate data where $n = 100$, and consider the cases $p = 100, 200, 300, 400, 500$. Let B be a $p \times k$ factor loading matrix, with all entries equal to 1. Let F_i be $k \times 1$ matrix of factor values, with all entries drawn independently from $\text{Normal}(0, 0.05)$. The ith row of the $n \times p$ design matrix X is generated by a factor model, with number of factors $k = 4$, as follows:

$$X_i = BF_i + \xi_i, \quad \xi_i \sim \text{Normal}(0, 0.1), \quad i = 1, \ldots, n.$$

Thus, the columns of X are correlated. Let $X = UDW^T$ denote the singular value decomposition of X. The observations y are generated from Equation (2.2) with $\sigma^2 = 1$, where for the true orthogonalized regression coefficients α_0, the 6, 30, 57, 67, and 96th components are randomly selected as signals, and the remaining 95 components are noise terms. Coefficients of the signals are set to be 10 or -10, and coefficients of the noise terms are 0.5 or -0.5. For the case $n = 100$ and $p = 500$, some of the true orthogonalized regression coefficients α_0, their ordinary least squared estimates $\hat{\alpha}$, and the corresponding singular values d of the design matrix, are shown in Table 2.1.

Table 2.2 lists the SURE for prediction and actual out of sample sum of squared prediction error (SSE) for the ridge, PCR, lasso and horseshoe regressions. Out of sample prediction error of the adaptive lasso is also included in the comparisons, although we are unaware of a formula for computing the SURE for adaptive lasso. SURE for ridge and PCR can be computed by an application of Equation (2.8) and SURE for the horseshoe regression is given by Theorem 2.3.1. SURE for the lasso is calculated using the result given by Tibshirani and Taylor (2012). In each case, the model is trained on 100 samples. We report the SSE on 100 testing samples, averaged over 200 testing data sets, and their standard deviations. For ridge, lasso, PCR and horseshoe regression, the global shrinkage parameters were chosen to minimize SURE for prediction. In adaptive lasso, the shrinkage parameters were chosen by cross val-

Table 2.1.: The true orthgonalized regression coefficients α_{0i}, their ordinary least square estimates $\hat{\alpha}_i$, and singular values d_i of the design matrix, for $n = 100$ and $p = 500$.

i	α_{0i}	$\hat{\alpha}_i$	d_i	$\hat{\alpha}_i d_i$
1	-0.5	-0.49	21.09	-10.29
2	-0.5	-0.48	3.16	-1.52
...	...	...	...	...
5	-0.5	-0.16	3.09	-0.49
6	10	10.72	3.01	32.31
...	...	...	...	...
29	-0.5	-0.58	2.54	-1.47
30	10	10.36	2.53	26.17
...	...	...	...	...
56	0.5	0.37	2.07	0.76
57	10	10.20	2.07	21.07
...	...	...	...	...
66	-0.5	-0.52	1.91	-0.99
67	10	9.98	1.89	18.91
...	...	...	...	...
95	-0.5	-0.17	1.41	-0.24
96	10	9.55	1.38	13.21
...	...	...	...	...
100	0.5	-0.08	1.27	-0.10

Table 2.2.: SURE and average out of sample prediction SSE (standard deviation of SSE) on one training set and 200 testing sets for $n = 100$. The competing methods are ridge regression (RR), principal components regression (PCR), the lasso regression (LASSO), the adaptive lasso (A_LASSO), and the horseshoe regression (HS). The lowest SURE in each row is in italics and the lowest average prediction SSE is in bold. A formula for SURE is unavailable for the adaptive lasso.

	RR		PCR		LASSO		A_LASSO	HS	
p	SURE	SSE	SURE	SSE	SURE	SSE	SSE	SURE	SSE
100	165.45	159.83	163.80	161.62	122.78	145.07	132.25	*116.01*	**123.07**
		(22.02)		(21.28)		(19.39)	(17.57)		(16.43)
200	185.31	176.78	190.89	195.20	137.87	151.47	156.47	*127.81*	**140.34**
		(25.05)		(26.56)		(22.32)	(22.97)		(21.42)
300	190.63	206.72	225.39	268.91	*146.81*	169.55	178.01	147.37	**162.72**
		(24.50)		(28.10)		(21.84)	(22.30)		(21.43)
400	194.69	190.20	340.57	321.11	172.05	175.99	207.60	*167.97*	**160.41**
		(22.27)		(31.39)		(21.95)	(26.00)		(20.07)
500	195.71	187.52	195.71	195.03	186.23	181.42	223.27	*174.49*	**164.03**
		(22.32)		(22.76)		(23.45)	(28.39)		(21.75)

idation due to SURE being unavailable. It can be seen that SURE in most cases are within one standard deviation of the actual out of sample prediction SSE, suggesting SURE is an accurate method for evaluating actual out of sample prediction performance. In all of the five cases, horseshoe regression has the lowest prediction SSE. The horseshoe regression also has the lowest SURE in all but one cases. Generally, SURE increases with p for all methods. The SURE for ridge regression approaches the OLS risk, which is $2n\sigma^2 = 200$ in these situations. The SURE for PCR can be larger than the OLS risk and PCR happens to be the poorest performer in most settings. Performance of the adaptive lasso also degrades compared to the lasso and the horseshoe, which remain the two best performers.

Figure 2.1 shows contribution to SURE by each component when $n = 100$ and $p = 500$, for ridge, PCR and horseshoe regressions. The components are ordered left to right on the x-axis by decreasing magnitude of d_i, and SURE for prediction on each component are shown on the y-axis. Note from Table 2.1 that the $6, 30, 57, 67$ and 96th components are the signals, meaning these terms correspond to a large α_0. For this data set, PCR selects 97 components, therefore SURE is equal to $2\sigma^2 = 2$ for the first 97 components, and is equal to $\hat{\alpha}_i^2 d_i^2$ for the last three components. Ridge SURE are large on the signal components, and decrease as the singular values d decrease on the other components. But due to the large global shrinkage parameter τ ridge must select in presence of both large signals and noise terms, its magnitude of improvement over the OLS risk $2\sigma^2$ is small. On the other hand, the horseshoe estimator does not shrink the components with large $\hat{\alpha}_i d_i$ heavily and therefore the horseshoe SURE on the signal components are almost equal to $2\sigma^2$ (according to Theorem 2.3.2) . Horseshoe SURE are also much smaller than $2\sigma^2$ on many of the noise components.

Figure 2.2 takes another look at the same results and shows component-wise SURE plotted against $\hat{\alpha}_i d_i$. Horseshoe SURE converges to $2\sigma^2$ when $\hat{\alpha}_i d_i$ is large, as expected from Theorem 2.3.2. For these components, the ridge SURE are larger than $2\sigma^2$, due to the bias introduced in estimating large signals (Carvalho et al., 2010). The upper bound of the horseshoe SURE is $2\sigma^2/3$ when $\hat{\alpha_i}^2 d_i^2 = 0$, a great improvement from the OLS risk, provided τ is chosen to be small enough. This upper bound and the other part of Theorem 2.3.3 can be verified from Figure 2.2.

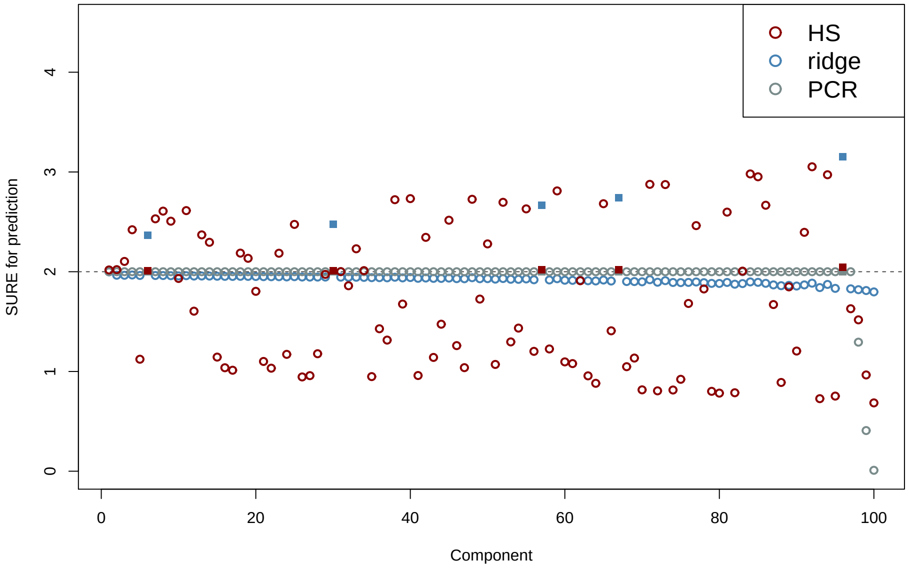


Figure 2.1.: Component-wise SURE for ridge, PCR, and horseshoe regression, when $n = 100$ and $p = 500$. Signal components (the 6, 30, 57, 67 and 96th components) are shown in solid squares and noise components shown in blank circles. Dashed horizontal line is at $2\sigma^2 = 2$.

2.5 Assessing out of sample prediction in a pharmacogenomics data set

We compare the out of sample prediction error of horseshoe regression with ridge regression, PCR, the lasso, and the adaptive = lasso on a pharmacogenomics data set. The data were originally described by Szakács et al. (2004), in which the authors studied 60 cancer cell lines in the publicly available NCI-60 database (`https://dtp.cancer.gov/discovery\_development/nci-60/`). The goal here is to predict the expression of the human ABC transporter genes (responses) using some compounds or drugs (predictors) at which 50% inhibition of cellular growth for the cell lines are induced. The NCI-60 database includes the concentration level of 1429 such compounds, out of which we use 853, which did not have any missing values, as predictors. We investigate the expression levels of transporter genes A1 to A12, (except for A11, which we omit due to missing values), and B1. Thus, in our study

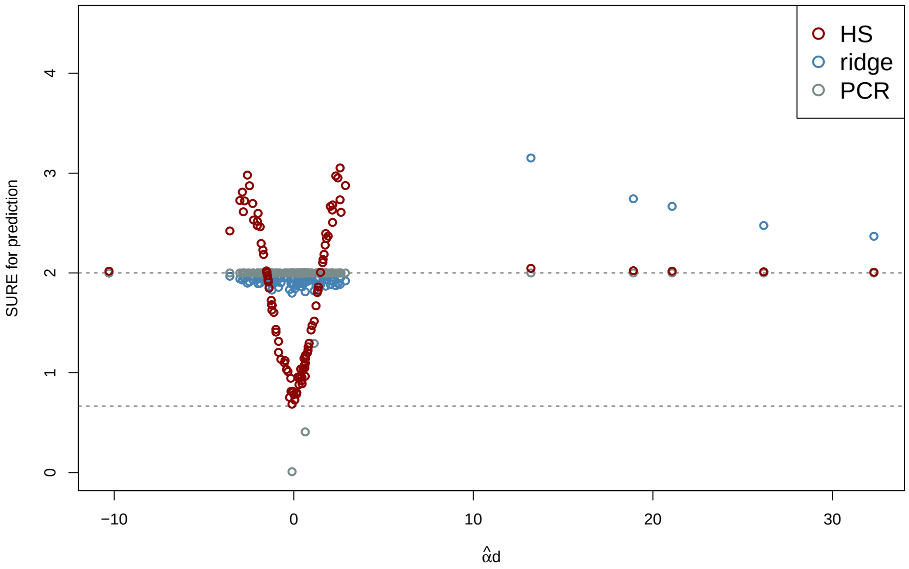


Figure 2.2.: SURE for ridge, PCR, and horseshoe regression, versus $\hat{\alpha}d$, where $\hat{\alpha}$ is the OLS estimate of the orthogonalized regression coefficient, and d is the singular value, when $n = 100$ and $p = 500$. Dashed horizontal lines are at $2\sigma^2 = 2$ and $2\sigma^2/3 = 0.67$.

X is a $n \times p$ matrix of predictors with $n = 60, p = 853$ and Y is a n-dimensional response vector for each of the 12 candidate transporter genes under consideration.

To test the performance of the methods, we split each data set into training and testing sets, with 75% (45 out of 60) of the observations in the training sets. We standardize each response by subtracting the mean and dividing by the standard deviation. We fit the model on the training data, and then calculate mean squared prediction error (prediction MSE) on the testing data. This is repeated for 20 random splits of the data into training and testing sets. The tuning parameters in ridge regression, the lasso and the adaptive lasso are chosen by five-fold cross validation on the training data. Similarly, the number of components in PCR and the global shrinkage parameter τ for horseshoe regression are chosen by cross validation as well. It is possible to use SURE to select the tuning parameters or the number of components, but one needs an estimate of the standard deviation of the errors in high-dimensional regressions. This is a problem of recent interest, as the OLS estimate of σ^2 is not well-defined in the $p > n$ case. Unfortunately, some of the existing methods we tried, such as the method of moments estimator of Dicker (2014), often resulted in unreasonable estimates for σ^2, such as negative numbers. Thus, we stick to cross validation here, as it is not necessary to estimate the residual standard deviation in that case.

The average prediction MSE over 20 random training-testing splits for the competing methods is reported in Table 2.3. Average prediction MSE for responses A1, A8 and A10 are around or larger than 1 for all of the methods. Since the responses are standardized before analysis, we might conclude that none of the methods performed well for these cases. The lasso and the adaptive lasso have the lowest prediction MSE for response A2. Among the remaining eight cases, the horseshoe regression substantially outperforms the other methods for A3, A4, A9, A12, B1, is comparable to PCR for A5 and A7, and is comparable to the adaptive lasso for A6, which are the best performers in the respective cases. Overall, the horseshoe regression performed the best in 5, the lasso in 3, the adaptive lasso in 2 and PCR in 2 cases, among the total 12 we considered.

2.6 Concluding remarks

We outlined some situations where global-local shrinkage regression is expected to perform better compared to some other commonly used "global" shrinkage or selection alternatives for high-dimensional regression. Specifically, we demonstrated that the global term helps in mitigating the prediction risk arising from the noise terms, and an appropriate choice for the tails of the local terms is crucial for controlling the risk due to the signal terms. For this article we have used the horseshoe prior as our choice for the global-local prior. However, in recent years, several other priors have been developed that fall in this class. This includes the horseshoe+ (Bhadra et al., 2017, 2016), the three-parameter beta (Armagan et al., 2011), the normal-exponential-gamma (Griffin and Brown, 2010), the generalized double Pareto (Armagan et al., 2013), the generalized shrinkage prior (Denison and George, 2012) and the

Table 2.3.: Average out of sample mean squared prediction error computed on 20 random training-testing splits (number of splits out of 20 with lowest prediction MSE), for each of the 12 human ABC transporter genes (A1-A11 and A13) in the pharmacogenomics example. Methods under consideration are ridge regression (RR), the lasso, the adaptive lasso (A_LASSO), principal components regression (PCR) and horseshoe regression (HS). Lowest prediction MSE and largest number of splits with the lowest prediction MSE for each response in bold.

Response	RR	LASSO	A_LASSO	PCR	HS
A1	1.1217	**0.9984**	0.9982	1.0982	1.2959
	(2)	**(9)**	(2)	(5)	(2)
A2	0.9957	**0.9465**	0.9339	1.0388	1.1537
	(3)	**(7)**	(6)	(1)	(3)
A3	0.7667	1.1133	0.8984	0.9109	**0.6461**
	(1)	(1)	(0)	(0)	**(18)**
A4	0.9235	0.9676	0.9579	0.9463	**0.7947**
	(2)	(2)	(2)	(1)	**(13)**
A5	0.8160	1.0569	0.8135	**0.7703**	0.7854
	(1)	(4)	(3)	**(6)**	**(6)**
A6	0.9292	0.9781	**0.8630**	0.9244	0.9462
	(4)	(4)	**(6)**	(0)	**(6)**
A7	0.9222	0.9154	0.9299	**0.8311**	0.8493
	(0)	(1)	(4)	**(8)**	(7)
A8	1.0789	1.1374	**1.0114**	1.0534	1.3360
	(6)	**(6)**	(4)	(4)	(0)
A9	0.5680	0.8116	0.6709	0.6387	**0.5452**
	(5)	(0)	(1)	(5)	**(9)**
A10	1.1846	**0.9972**	1.0090	1.0407	1.0978
	(0)	(6)	(3)	**(7)**	(4)
A12	1.0110	1.0875	1.0117	1.1204	**0.8701**
	(0)	(2)	(2)	(1)	**(15)**
B1	0.5329	0.6964	0.6270	0.5948	**0.4597**
	(2)	(4)	(2)	(0)	**(12)**

Dirichlet-Laplace prior (Bhattacharya et al., 2015). Empirical Bayes approaches have also appeared (Martin and Walker, 2014) and the spike-and-slab priors have made a resurgence due to recently developed efficient computational approaches (Ročková and George, 2014; Ročková and George, 2016). We expect the results developed in this article for horseshoe to foreshadow similar results when many of these alternatives are deployed. A particular advantage of using the horseshoe prior seems to be the tractable expression for SURE, as developed in Theorem 2.3.1. Whether this advantage carries over to some of the other global-local priors identified above is an open question. It will also be an interesting exercise to compare the performances of various global-local priors in simulations as well as in real prediction problems. Another possible direction for future investigation might be to explore the implications of our findings on the predictive density in terms of an appropriate metric, say the Kullback-Leibler loss, following the results of (George et al., 2006).

3. THE GRAPHICAL HORSESHOE ESTIMATOR FOR INVERSE COVARIANCE MATRICESO

3.1 Introduction

Estimation of the covariance, or inverse covariance matrix, of a multivariate normal vector plays a central role in numerous fields, including spatial data analysis (Cressie, 1993), variance components and longitudinal data analysis (Diggle, 2002), and the growing area of genetic data analysis (Dehmer and Emmert-Streib, 2008). Pourahmadi (2011) provides a survey of some of the most popular methods in high-dimensional covariance and inverse covariance estimation. In a penalized likelihood framework, two of the most notable methods for inverse covariance estimation are the graphical lasso (Friedman et al., 2008) and the graphical SCAD (Fan et al., 2009). Both these methods provide estimates for a high-dimensional inverse covariance matrix under an arbitrary sparsity pattern.

There has also been much recent work in covariance and inverse covariance estimation in a Bayesian framework. Banerjee and Ghosal (2014) proposed a prior distribution for estimating a banded inverse covariance matrix. Rajaratnam et al. (2008) and Xiang et al. (2015) proposed Bayesian estimators for the covariance of a decomposable Gaussian graphical model. Pati et al. (2014) considered sparse factor models for covariance matrices and induced a class of continuous shrinkage priors on the factor loadings. There are also studies that focus on the theoretical properties of these estimators, including posterior convergence rates, Bayesian minimax rates and consistency of Bayesian estimators (Banerjee and Ghosal, 2014, 2015; Xiang et al., 2015; Lee and Lee, 2017a,b). However, to our knowledge, few Bayesian estimators assume an arbitrary sparsity pattern of the true inverse covariance matrix. Under such an assumption, Banerjee and Ghosal (2015) proposed a mixture prior for graphical structure learning, and Wang (2012) developed a Bayesian version of the graphical lasso.

In this paper, we propose an alternative Bayesian estimator, which we call the graphical horseshoe estimator. This estimator works under the assumption of an arbitrary sparsity pattern in the inverse covariance matrix. We show that our estimator has better performance in adapting to sparsity in high-dimensional problems than some competing methods because of two properties of our prior: greater concentration near the origin and heavier tails. Both of these properties are inherited from the horseshoe prior of Carvalho et al. (2010) for the sparse normal means model.

Many attractive theoretical properties of the horseshoe prior have been discovered in recent years for the normal means model. These include improved Kullback–Leibler risk bounds (Carvalho et al., 2010), asymptotic optimality in testing under $0-1$ loss (Datta and Ghosh, 2013), minimaxity in estimation under the ℓ_2 loss (van der Pas et al., 2014), and improved risk properties in linear regression (Chapter 2). In this paper, we demonstrate how some of these properties translate to the estimation of the inverse covariance matrix in a multivariate Gaussian model. We discuss the implications of these properties both theoretically and empirically.

The remainder of this paper is organized as follows. The rest of Section 3.1 discusses three competing methods for sparse precision matrix estimation: the graphical lasso, the graphical SCAD, and the Bayesian graphical lasso. Section 3.2 outlines the graphical horseshoe estimator as well as a full Gibbs sampler for easy and efficient sampling. Sections 3.3 and 3.4 outline the theoretical properties of our proposed estimator along with a comparison to the graphical lasso and graphical SCAD estimators. Section 3.5 illustrates these theoretical properties through simulations. Section 3.6 applies the proposed method on a human gene expression data set to identify a sparse gene interaction network, before concluding with some discussion of possible future research topics in Section 3.7.

3.1.1 Related Works in Precision Matrix Estimation

Consider n samples from a p-dimensional multivariate normal distribution with zero mean and a $p \times p$ covariance matrix Ω^{-1}. That is,

$$\mathbf{y}_k \sim \text{Normal}(\mathbf{0}, \Omega^{-1}),$$

for $k = 1, \ldots, n$. Under this parameterization, the inverse of the covariance matrix, Ω, is referred to as the precision matrix (assumed to be positive definite). The ijth off-diagonal element in Ω is the negative of the partial covariance between features i and j, and the ith diagonal element is the inverse of the residual variance when the ith feature is regressed on all the other features (Pourahmadi, 2011). Under the multivariate normal model, zero off-diagonal elements in Ω correspond to features that are conditionally independent given the remaining features. In certain applications, estimating the precision matrix is attractive, especially in high-dimensional cases, since it is preferable to study partial correlations rather than marginal correlations (Pineda-Pardo et al., 2014).

A major challenge in precision matrix estimation is that the number of free parameters grows quadratically with the number of features. As a consequence, in high-dimensional problems, some methods assume the covariance or precision matrix has a structure, such as latent factors (Pati et al., 2014) or banding (Banerjee and Ghosal, 2014). When the structure of the true precision matrix is assumed to be arbitrary, the precision matrix is usually assumed to be sparse. In high-dimensional settings, a natural approach for estimating a sparse model is to penalize the likelihood. Friedman et al. (2008) proposed the graphical lasso, which estimates the precision matrix under the lasso penalization (Tibshirani, 1996) while maintaining the symmetry of the estimate. The graphical lasso maximizes the penalized likelihood:

$$\log(\det \Omega) - \text{tr}(S\Omega/n) - \sum_{i,j} \phi_\lambda(|\omega_{ij}|), \tag{3.1}$$

where $S = \sum_{i=1}^{n} \mathbf{y}_i\mathbf{y}_i$ is the scatter matrix, $\Omega = (\omega_{ij})$, $\phi_\lambda(|\omega_{ij}|) = \lambda|\omega_{ij}|$ is the ℓ_1 penalty, and λ is a tuning parameter. In practice, λ is often chosen by cross validation. The sum $\sum_{i,j} \phi_\lambda(|\omega_{ij}|)$ in Equation (3.1) can be taken with or without a penalty on the diagonal terms (Rothman et al., 2008; Meinshausen and Bühlmann, 2006; Yuan and Lin, 2007; Friedman et al., 2008).

A Bayesian version of graphical lasso was proposed by Wang (2012). In the Bayesian setting, the frequentist graphical lasso estimator is equivalent to the maximum a posteriori estimate of Ω under the following prior:

$$p(\Omega \mid \lambda) \propto \prod_{i<j} \{\mathrm{DE}(\omega_{ij} \mid \lambda)\} \prod_{i=1}^{p} \{\mathrm{EXP}(\omega_{ii} \mid \lambda/2)\} 1_{\Omega \in \mathcal{S}_p} ,$$

where $\mathrm{DE}(x \mid \lambda)$ represents the double exponential distribution with rate λ, $\mathrm{EXP}(x \mid \lambda)$ represents the exponential distribution with rate λ, and $\mathcal{S}_p$ is the space of $p \times p$ positive definite matrices. The tuning parameter λ, or rather the hyper-parameter in the language of Bayesian hierarchical models, can be chosen by cross-validation as in a frequentist framework (Friedman et al., 2008; Rothman et al., 2008), or by a fully Bayesian method with an appropriate hyperprior.

The smoothly clipped absolute deviation (SCAD) penalty by Fan and Li (2001) was introduced in precision matrix estimation because of its attractive asymptotic properties. The graphical SCAD maximizes the penalized likelihood in Equation (3.1) where the penalty has the first order derivative:

$$\phi'_\lambda(|x|) = \lambda \left\{ 1_{\{|x| \leq \lambda\}} + \frac{(a\lambda - |x|)_+}{(a-1)\lambda} 1_{\{|x| > \lambda\}} \right\} ,$$

with $a > 2$ and $\lambda > 0$. This penalty is linear near the origin and non-decreasing. In practice, the tuning parameter a is often fixed while λ is chosen by cross validation. The graphical SCAD estimate satisfies the oracle property given by Fan and Li (2001). The SCAD penalty does not have a Bayesian representation, although Polson and Scott (2012) provide an understanding of how priors and penalty functions are related even when some penalty functions lack Bayesian equivalents. Lam and Fan (2009) showed that under certain conditions, both frequentist graphical lasso and graphical SCAD estimates of the precision matrix converge to the true precision matrix under the Frobenius norm. However, these theoretical results depend on theoretical choices of tuning parameters, which cannot be implemented in practice. The regulatory conditions are also difficult to check in data analysis.

All methods for large sparse precision matrix estimation face the problem of accumulated estimation errors due to the large number of parameters to estimate. Furthermore, the double-exponential priors in the Bayesian lasso have been shown to possess some undesirable properties in the high-dimensional normal means problem (Carvalho et al., 2009, 2010). Although lasso and SCAD are widely-used methods with good asymptotic properties, the element-wise bias of graphical lasso estimates can be large, and graphical SCAD does not guarantee positive definite estimates (Fan et al., 2016).

To provide an alternative that remedies the accumulation of errors in high dimensions, we propose a method that obtains a sparse estimate while controlling the element-wise bias of the nonzero elements. In terms of sampling, our method follows the technique adopted in the Bayesian graphical lasso by Wang (2012). However, our method is more efficient at utilizing the sparsity of the precision matrix than the graphical lasso and the graphical SCAD, for reasons we detail in Section 3.3. Our method also guarantees positive definite and symmetric estimates.

3.2 The Graphical Horseshoe Estimator

Since an unstructured precision matrix is assumed to be sparse, a shrinkage method should be able to give a zero or very small estimate for the zero elements. Meanwhile, a method should also be able to distinguish the non-zero elements in the precision matrix and shrink them as little as possible. We propose the use of the horseshoe prior to do just this.

3.2.1 The Graphical Horseshoe Hierarchical Model

The graphical horseshoe model puts horseshoe priors on the off-diagonal elements of the precision matrix, and an uninformative prior on the diagonal elements, while respecting the constraint $\Omega \in \mathcal{S}_p$. Because the precision matrix is symmetric, we only consider the upper off-diagonal elements. The element-wise priors are specified for $i, j = 1, \dots, p$ as follows:

$$\begin{aligned} \omega_{ii} &\propto 1, \\ \omega_{ij:i<j} &\sim \text{Normal}(0, \lambda_{ij}^2 \tau^2), \\ \lambda_{ij:i<j} &\sim C^+(0,1), \\ \tau &\sim C^+(0,1), \end{aligned}$$

where $C^+(0,1)$ denotes a half-Cauchy random variable with density $p(x) \propto (1 + x^2)^{-1};\ x > 0$. The normal scale mixtures with half-Cauchy hyperpriors on the off-diagonal elements is the horseshoe prior proposed by Carvalho et al. (2010). The distinctive scale parameter λ_{ij} on each dimension is referred to as the local shrinkage parameter, and the scale parameter τ shared by all dimensions is referred to as the global shrinkage parameter. The marginal prior's peak near the origin induces efficient shrinkage of noise terms in a high-dimensional problem, and the slow decaying tail ensures that signal terms are shrunk very little (Carvalho et al., 2010).

Thus, the prior on Ω under graphical horseshoe model can be written as:

$$p(\Omega \mid \tau) \propto \prod_{i<j} \text{Normal}(\omega_{ij} \mid \lambda_{ij}^2, \tau^2) \prod_{i<j} \text{C}^+(\lambda_{ij} \mid 0, 1) 1_{\Omega \in \mathcal{S}_p},$$

where $\mathcal{S}_p$ is the space of $p \times p$ positive definite matrices. Using the properties of the horseshoe prior Carvalho et al. (2010), the induced marginal prior on ω_{ij} is proper. When $\Omega \in \mathcal{S}_p$, the diagonal elements in Ω are finite. Therefore the graphical horseshoe prior is proper. In a univariate normal case, the induced marginal prior for shrinkage has infinite mass near both 0 and 1 and is thin in between, with a "horseshoe" shape Carvalho et al. (2010).

In high-dimensional precision matrix estimation by the graphical horseshoe, the global shrinkage parameter τ adapts to the sparsity of the entire matrix Ω and shrinks the estimates of the off-diagonal elements toward zero. On the other hand, the local shrinkage parameters $\lambda_{ij:i<j}$ preserve the magnitude of non-zero off-diagonal elements, and ensure that the element-wise biases are not very large.

3.2.2 A Data-augmented Block Gibbs Sampler

Posterior samples under the graphical horseshoe hierarchical model are drawn by an augmented block Gibbs sampler, adapting the scheme proposed by Makalic and Schmidt (2016) for linear regression. Augmented variables $\nu_{ij:i<j}$ and ξ are introduced for conjugate sampling of the shrinkage parameters $\lambda_{ij:i<j}$ and τ. In each iteration, each column and row of Ω, $\Lambda = (\lambda_{ij}^2)$, and $N = (\nu_{ij})$ are partitioned from a $p \times p$ matrix of parameters and updated in a block. Then the global shrinkage parameter τ and its auxiliary variable ξ are updated.

The following part derives the posterior distribution of the precision matrix. Given data $Y_{n\times p}$ and the shrinkage parameters, the posterior of Ω under the graphical horseshoe model is

$$p(\Omega \mid Y, \Lambda, \tau) \propto |\Omega|^{\frac{n}{2}} \exp\Big\{ -\operatorname{tr}\Big(\frac{1}{2} S\Omega\Big)\Big\} \prod_{i<j} \exp\Big(- \frac{\omega_{ij}^2}{2\lambda_{ij}^2\tau^2}\Big) 1_{\Omega\in\mathcal{S}_p}.$$

It is not obvious how to sample from this distribution. Following Wang (2012), one column and row of Ω are updated at a time. Without loss of generality, the posterior distributions for the last column and the last row are derived here. First, partition the last column and row in the matrix:

$$\Omega = \begin{pmatrix} \Omega_{(-p)(-p)} & \boldsymbol{\omega}_{(-p)p} \\ \boldsymbol{\omega}'_{(-p)p} & \omega_{pp} \end{pmatrix}, \ S = \begin{pmatrix} S_{(-p)(-p)} & \mathbf{s}_{(-p)p} \\ \mathbf{s}'_{(-p)p} & s_{pp} \end{pmatrix}, \ \Lambda = \begin{pmatrix} \Lambda_{(-p)(-p)} & \boldsymbol{\lambda}_{(-p)p} \\ \boldsymbol{\lambda}'_{(-p)p} & 1 \end{pmatrix},$$

where $(-p)$ denotes the set of all indices except for p, and $\Lambda_{(-p)(-p)}$ and $\boldsymbol{\lambda}_{(-p)p}$ have entries λ_{ij}^2. Diagonal elements of $\Lambda_{(-p)(-p)}$ can be arbitrarily set to 1. Then, the full conditional of the last column of Ω is

$$\begin{aligned} p(\boldsymbol{\omega}_{(-p)p}, \omega_{pp} \,|\, \Omega_{(-p)(-p)}, Y, \Lambda, \tau) \propto & (\omega_{pp} - \boldsymbol{\omega}'_{(-p)p}\Omega^{-1}_{(-p)(-p)}\boldsymbol{\omega}_{(-p)p})^{n/2} \\ & \times \exp\{-\mathbf{s}'_{(-p)p}\boldsymbol{\omega}_{(-p)p} - s_{pp}\omega_{pp}/2 - \boldsymbol{\omega}'_{(-p)p}(\Lambda^*\tau^2)^{-1}\boldsymbol{\omega}_{(-p)p}/2\}, \end{aligned}$$

where Λ^* is a diagonal matrix with $\boldsymbol{\lambda}_{(-p)p}$ in the diagonal.

Next, a variable change is performed to obtain gamma and multivariate normal distributed variables, which can be efficiently sampled. Let $\boldsymbol{\beta} = \boldsymbol{\omega}_{(-p)p}$ and $\gamma = \omega_{pp} - \boldsymbol{\omega}'_{(-p)p}\Omega^{-1}_{(-p)(-p)}\boldsymbol{\omega}_{(-p)p}$. The Jacobian of the transformation is a constant, and the full conditional of $\boldsymbol{\beta}$ and γ is

$$\begin{aligned} p(\boldsymbol{\beta}, \gamma \,|\, \Omega_{(-p)(-p)}, Y, \Lambda, \tau) & \propto \gamma^{n/2}\exp[-\frac{1}{2}\{s_{pp}\gamma + \boldsymbol{\beta}' s_{pp}\Omega^{-1}_{(-p)(-p)}\boldsymbol{\beta} + \boldsymbol{\beta}'(\Lambda^*\tau^2)^{-1}\boldsymbol{\beta} + 2\mathbf{s}'_{(-p)p}\boldsymbol{\beta}\}] \\ & \sim \text{Gamma}(n/2+1, 2/s_{pp})\text{Normal}(-C\mathbf{s}_{(-p)p}, C), \qquad (3.2) \end{aligned}$$

where $C = \{s_{pp}\Omega^{-1}_{(-p)(-p)} + (\Lambda^*\tau^2)^{-1}\}^{-1}$.

Therefore the posterior distribution of the last row and column of Ω is obtained. All elements in the matrix Ω can be sampled by sampling one row and column at a time.

Next, the local and global shrinkage parameters λ_{ij} and τ need to be sampled. Makalic and Schmidt (2016) made the following key observation: if $x^2 \mid a \sim \text{InvGamma}(1/2, 1/a)$ and $a \sim \text{InvGamma}(1/2, 1)$, then marginally $x \sim C^+(0,1)$, where the shape–scale parameterization is used for the inverse gamma distribution. The inverse gamma distribution is conjugate for the variance parameter in a linear regression model with normal errors and to itself, which ensures all required conditionals also follow inverse gamma distribution. Thus, introduce latent ν_{ij} and write $\lambda_{ij}^2 \mid \nu_{ij} \sim \text{InvGamma}(1/2, 1/\nu_{ij})$, and $\nu_{ij} \sim \text{InvGamma}(1/2, 1)$. Since from Equation (3.2), the full conditional posterior distribution of $\boldsymbol{\beta}$ is normal, the full conditional posteriors of λ_{ij} and ν_{ij} are easily obtained as $\lambda_{ij}^2 \mid \cdot \sim \text{InvGamma}(1, 1/\nu_{ij} + \omega_{ij}^2/2\tau^2)$ and $\nu_{ij} \mid \cdot \sim \text{InvGamma}(1, 1 + 1/\lambda_{ij}^2)$, respectively. Using a similar parameterization, the full conditional posteriors for τ^2 and its auxiliary variable ξ are also inverse gamma.

Thus, combining the matrix partition and variable change for Bayesian graphical lasso proposed by Wang (2012) and the variable augmentation for the half-Cauchy prior proposed by Makalic and Schmidt (2016), the graphical horseshoe model has all conditionals in closed form and hence, admits a full Gibbs sampler. The sampler is summarized in Algorithm 1.

The constraint on $\Omega \in \mathcal{S}_p$ is maintained in every iteration as long as the starting value is positive definite, for the same reason that the positive definiteness is maintained in Bayesian graphical lasso (Wang, 2012). Suppose that at iteration t, the current sample $\Omega^{(t)}$ is positive definite. Then all of its p leading principal minors are positive. After updating the last column and row of Ω, the new sample $\Omega^{(t+1)}$ has the same leading principal minors as $\Omega^{(t)}$ except for the last one which is of order p. The last leading principal minor is $\det(\Omega^{(t+1)}) = \gamma\det(\Omega^{(t)}_{(-p)(-p)})$, and is positive since both γ and $\det(\Omega^{(t)}_{(-p)(-p)})$ are positive. Consequently, $\Omega^{(t+1)}$ after updating is positive definite.

The required full conditionals in the proposed Gibbs sampler are either multivariate normal, gamma or inverse gamma, for which efficient sampling methods exist. Full conditional posteriors of the local shrinkage parameters $\lambda_{ij:i<j}$ are mutually independent, and so are $\nu_{ij:i<j}$. This facilitates batch updating and a large number of features does not cause problems in sampling of $\boldsymbol{\lambda}$ and $\boldsymbol{\nu}$. The most computationally expensive step is the sampling of $\boldsymbol{\beta}$, where the $(p-1) \times (p-1)$ matrix $\Omega_{(-p)(-p)}$ and $\{s_{pp}\Omega^{-1}_{(-p)(-p)} + (\Lambda^*\tau^2)^{-1}\}$ need to be inverted, which has computational complexity $O(p^3)$. In Algorithm 1, $\Omega_{(-p)(-p)}$ is inverted by block form of the sampled covariance matrix, so only $\{s_{pp}\Omega^{-1}_{(-p)(-p)} + (\Lambda^*\tau^2)^{-1}\}$ needs to be inverted. MATLAB code for Algorithm 1, along with a simulation example, are freely available at *http://github.com/liyf1988/GHS.*

Algorithm 1 The Graphical Horseshoe Sampler

```
function GHS($S, n, burnin, nmc$)    ▷ Where $S = Y'Y$, n=sample size
    Set $p$ to be number of rows (or columns) in $S$
    Set initial values $\Omega = I_{p\times p}$, $\Sigma = I_{p\times p}$, $\Lambda = \nVdash$, $N = \nVdash$, $\tau = 1$, $\xi = 1$, where $\nVdash$ is a matrix with all elements equal to 1, $\Lambda$ has entries of $\lambda_{ij}^2$, $N$ has entries of $\nu_{ij}$
    for $iter = 1$ to $(burnin + nmc)$ do
        for $i = 1$ to $p$ do
            $\gamma \sim$ Gamma(shape $= n/2 + 1$, scale $= 2/s_{ii}$)    ▷ sample $\gamma$
            $\Omega^{-1}_{(-i)(-i)} = \Sigma_{(-i)(-i)} - \boldsymbol{\sigma}_{(-i)i}\boldsymbol{\sigma}'_{(-i)i}/\sigma_{ii}$
            $C = (s_{ii}\Omega^{-1}_{(-i)(-i)} + \text{diag}(\boldsymbol{\lambda}_{(-i)i}\tau^2)^{-1})^{-1}$
            $\boldsymbol{\beta} \sim$ Normal$(-C\mathbf{s}_{(-i)i}, C)$    ▷ sample $\boldsymbol{\beta}$
            $\boldsymbol{\omega}_{(-i)i} = \boldsymbol{\beta}$, $\omega_{ii} = \gamma + \boldsymbol{\beta}'\Omega^{-1}_{(-i)(-i)}\boldsymbol{\beta}$    ▷ variable transformation
            $\boldsymbol{\lambda}_{(-i)i} \sim$ InvGamma(shape $= 1$, scale $= 1/\boldsymbol{\nu}_{(-i)i} + \boldsymbol{\omega}^2_{(-i)i}/2\tau^2$)    ▷ sample $\boldsymbol{\lambda}$, where $\boldsymbol{\lambda}_{(-i)i}$ is a vector of length $(p-1)$ with entries $\lambda_{ji}^2$, $j \neq i$
            $\boldsymbol{\nu}_{(-i)i} \sim$ InvGamma$(1,\ 1 + 1/\boldsymbol{\lambda}_{(-i)i})$    ▷ sample $\boldsymbol{\nu}$
            Save updated $\Omega$
            $\Sigma_{(-i)(-i)} = \Omega^{-1}_{(-i)(-i)} + (\Omega^{-1}_{(-i)(-i)}\boldsymbol{\beta})(\Omega^{-1}_{(-i)(-i)}\boldsymbol{\beta})'/\gamma$, $\boldsymbol{\sigma}_{(-i)i} = -(\Omega^{-1}_{(-i)(-i)}\boldsymbol{\beta})/\gamma$, $\sigma_{ii} = 1/\gamma$
            Save updated $\Sigma$, $\Lambda$, $N$
        end for
        $\tau^2 \sim$ InvGamma$((\binom{p}{2} + 1)/2,\ 1/\xi + \sum_{i,j:i<j}\omega_{ij}^2/2\lambda_{ij}^2)$    ▷ sample $\tau$
        $\xi \sim$ InvGamma$(1,\ 1 + 1/\tau^2)$    ▷ sample $\xi$
    end for
    Return MC samples $\Omega$
end function
```

3.3 Kullback–Leibler Risk Bounds

In this section, we discuss the Kullback–Leibler divergence between the true sampling density and the Bayes estimator of the density function under various priors, including the graphical horseshoe prior. The Cesàro-average risk of the posterior distribution diverges for all methods when $p^2/n \to \infty$, but the upper bound of the average risk under the graphical horseshoe prior diverges slower than some other methods, as discussed below.

Suppose that there is a true sampling model. Let Ω_0 denote the true value of the precision matrix, $p_\Omega = p(y\,|\,\Omega)$ denote a sampling density with parameter Ω, and $\nu(A)$ denote the measure of some set A. Let $D(p_0||p_1) = \mathrm{E}_{p_2}\log\left(p_2/p_1\right)$ denote the

Kullback–Leibler divergence from p_1 to p_2. Then Barron (1988) proved the following lemma on the Cesàro-average risk of the Bayes posterior mean estimator of the density function.

Proposition 3.3.1 *(Barron, 1988) Let $A_\epsilon = \{\Omega : D(p_{\Omega_0}||p_\Omega) \leq \epsilon\} \subset \mathbb{R}^{p\times p}$ denote the Kullback–Leibler information neighborhood of size ϵ, centered at Ω_0. Let $\nu(d\Omega)$ be the prior measure of Ω and $\nu_n(d\Omega) \propto \prod_{i=1}^n p_\Omega(y_i)\nu(\mathrm{d}\Omega)$ be the posterior measure after observing i.i.d. $y_1, ..., y_n$ from the sampling density p_Ω. Let $\hat{p}_n = \int p_\Omega \nu_n(d\Omega)$ be the posterior mean estimator of the density function. Under the assumption that the prior measure $\nu(A_\epsilon) > 0$ for all $\epsilon > 0$, the Cesàro-average risk R_n of the estimator $\hat{p}_n$ admits the following upper bound for all $\epsilon > 0$:*

$$R_n = \frac{1}{n}\sum_{j=1}^{n} \mathrm{E}D(p_{\Omega_0}||\hat{p}_j) \leq \epsilon - \frac{1}{n}\log \nu(A_\epsilon),$$

where the expectation is with respect to the posterior predictive distribution given $y_1, ..., y_n$.

Taking $\epsilon = 1/n$, the upper bound of R_n is a function of two things: the sample size n, and the prior measure of the Kullback–Leibler information neighborhood A_ϵ of true Ω_0. Since the horseshoe prior has higher mass near the true parameter than any prior that is bounded above when the true parameter is zero, the graphical horseshoe estimator has a smaller upper bound on R_n when the true precision matrix is sparse. The result is summarized in the following theorem.

Theorem 3.3.1 *Suppose the true sampling model is $y \sim \mathrm{Normal}(\mathbf{0}, \Omega_0)$. Let σ_{ij0} denote the ijth element of the true covariance matrix Σ_0, and ω_{ij0} denote the ijth element of the true precision matrix Ω_0. Suppose that $\sum_{i,j} \sigma_{ij0} = Mp$ where M is a constant. That is, the summation of all elements in Σ_0 grows linearly with the number of features p. Suppose that an Euclidean cube in the neighborhood of Ω_0 with $(\omega_{ij0} - 2/Mn^{1/2}p, \omega_{ij0} + 2/Mn^{1/2}p)$ on each dimension lies in the cone of positive definite matrices $\mathcal{S}_p$. Then $R_n \leq \frac{1}{n} - \frac{1}{n}\log \nu(A_{1/n})$ for all n, and:*

(1) For $\hat{p}_n$ under the graphical horseshoe prior, $\log\nu(A_{1/n}) > p_0\log\{\frac{C_1}{Mn^{1/2}p}\log(2Mn^{1/2}p)\}$ $+p_1\log\frac{C_2}{n^{1/2}p}$*, where p_0 is the number of zero elements in Ω_0, p_1 is the number of nonzero elements in Ω_0, and C_1 and C_2 are constants.*

(2) Suppose $p(\omega_{ij})$ is any other prior density that is continuous, bounded above, and strictly positive on a neighborhood of the true value ω_{ij0}. Then $\log\nu(A_{1/n}) > p^2\log\frac{K_1}{n^{1/2}p}$*, where K_1 is a constant.*

Proof of Theorem 3.3.1 can be found in Appendix B. The neighborhood $A_{1/n}$ is bounded by two Euclidean cubes on $p \times p$ dimensions where the edges of these cubes have length proportional to $n^{1/2}p$ on each dimension. On these cubes, the measure of $p(\Omega)$ is obtained by the product of the measures of $p(\omega_{ij})$ on each of the p^2 dimensions of Ω. Any Bayesian estimator with a prior density bounded above near the origin gives a prior measure of order $1/(n^{1/2}p)$ on each dimension. The graphical horseshoe estimator gives a prior measure of order $\log(n^{1/2}p)/(n^{1/2}p)$ on each dimension with $\omega_{ij0} = 0$, and a measure of order $1/(n^{1/2}p)$ on each dimension with nonzero ω_{ij0}.

Some common Bayesian estimators, including the double exponential prior in Bayesian lasso, induce a prior density bounded above near the origin (Carvalho et al., 2010). Although the SCAD estimate can not be expressed as a maximum a posteriori estimate, the prior density corresponding to the SCAD penalty would be bounded by Theorem 1 in Polson and Scott (2012). Therefore, Bayesian graphical lasso has an upper bound corresponding to Part (2) of Theorem 3.3.1. Similarly, the posterior distribution of a Bayesian version of graphical SCAD would also have an upper bound corresponding to Part (2) of Theorem 3.3.1, if such a Bayesian version existed. These methods put a prior measure of order $1/(n^{1/2}p)$ near the true parameter on each dimension, regardless of whether or not the true parameter is zero. Unlike the horseshoe prior, these methods do not put dense prior mass near the origin, and do not utilize the fact (or expectation) that most of the true parameters are zero.

Theorem 1 of Rissanen (1986) gives an asymptotic lower bound on the Kullback–Leibler divergence $D(p_{\Omega_0}||\hat{p}_n)$, which is $(1/2 - \epsilon)\,k\log n$ for all $\epsilon > 0$, where k is the dimension of the parameter vector. This lower bound implies that in our problem, all methods have divergent average risk R_n when $n \to \infty$ and $p^2/n \to \infty$. Though all methods fail when dimension is large, Theorem 3.3.1 gives a non-asymptotic upper bound on R_n for any sample size n. One element where the true parameter is zero contributes $(\log n^{1/2}p - \log C)/n$ to the upper bound of R_n under a bounded prior near the origin, and $(\log Mn^{1/2}p - \log C - \log\log 2Mn^{1/2}p)/n$ to the upper bound of R_n under the graphical horseshoe prior. For each element where the true parameter is zero, the graphical horseshoe average risk has an extra $-O\{(\log\log n^{1/2}p)/n\}$ term. Consequently, when most off-diagonal elements in the true precision matrix are zero, the graphical horseshoe estimate provides a non-trivial improvement on R_n. In Section 3.5, we will compare the Kullback–Leibler divergence of graphical horseshoe estimate to graphical lasso and graphical SCAD estimates in simulations. We will show that the graphical horseshoe estimate has smaller Kullback–Leibler divergence, especially when the precision matrix is extremely sparse. However, we will discuss the bias of the graphical horseshoe estimate first.

3.4 Bias of the Graphical Horseshoe Estimator

Suppose that all diagonal elements in the precision matrix are known. Then, by the partial regression representation of the parameters (Pourahmadi, 2011), given the rest of the features, an observation of the ith feature follows a normal distribution $y_i \mid \mathbf{y}_{(-i)} \sim \text{Normal}(-\omega_{ii0}^{-1}\boldsymbol{\omega}_{i(-i)0}\mathbf{y}_{(-i)}, \omega_{ii0}^{-1})$, where y_i is an observation of the ith feature, $\mathbf{y}_{(-i)}$ is an observation of all features other than i, ω_{ii0} is the diagonal element in the true precision matrix corresponding to feature i, and $\boldsymbol{\omega}_{i(-i)0}$ is the off-diagonal elements in the true precision matrix on the ith row. Without loss of generality, the following discussion takes $i = p$. Given observations of features 1 to $p-1$, $Y_{(-p)}$, the least squares estimate of the pth column in the precision matrix is an unbiased estimate with a normal distribution

$$\hat{\boldsymbol{\omega}}_{p(-p)} \mid Y_{(-p)} \sim \text{Normal}(\boldsymbol{\omega}_{p(-p)0}, w_{pp0}(Y_{(-p)}'Y_{(-p)})^{-1}).$$

Marginally, the least squares estimate of each element $\hat{\omega}_{pj}$ in $\hat{\boldsymbol{\omega}}_{p(-p)}$ has a univariate normal distribution

$$\hat{\omega}_{pj} \,|\, Y_{(-p)} \sim \mathrm{Normal}(\omega_{pj0}, w_{pp0}(Y'_{(-p)}Y_{(-p)})^{-1}_{jj}).$$

We use this property of the least squares estimate to state our main result on the element-wise bias of the graphical horseshoe estimate.

Theorem 3.4.1 *Scale both ω_{pj} and its least squares estimate $\hat{\omega}_{pj}$ by $\{\omega_{pp0}(Y'_{(-p)}Y_{(-p)})^{-1}_{jj}\}^{-1/2}$, and denote the scaled parameter and its least squares estimate by ω'_{pj} and $\hat{\omega}'_{pj}$, respectively.*

(1) The posterior mean estimate of ω'_{pj} under the graphical horseshoe prior is $\mathrm{E}(\omega'_{pj} \,|\, Y, \tau) = (1 - \mathrm{E}(Z_{pj}))\hat{\omega}'_{pj}$*, where Z_{pj} follows a Compound Confluent Hypergeometric distribution with parameters* $(1,\, 1/2,\, 1,\, \hat{\omega}'^2_{pj}/2,\, 1,\, \omega_{pp0}(Y'_{(-p)}Y_{(-p)})^{-1}_{jj}\tau^{-2})$ *and has support between* 0 *and* 1*. Let $\theta_{pj} = \omega_{pp0}(Y'_{(-p)}Y_{(-p)})^{-1}_{jj}\tau^{-2}$, then* $\mathrm{E}(Z_{pj}) < 4(C_1 + C_2)\theta_{pj}(1 + \hat{\omega}'^2_{pj}/2)/\hat{\omega}'^4_{pj}$ *when $\hat{\omega}'^2_{pj}/2 > 1$, where $C_1 = 1 - 2e \approx 0.26$ and* $C_2 = \Gamma(1/2)\Gamma(2)/\Gamma(2.5) = 0.75$*. Consequently,* $\mathrm{E}(Z_{pj}) = O(1/\hat{\omega}'^2_{pj})$ *when $\hat{\omega}'_{pj} \to \infty$.*

(2) The posterior mean estimate of ω'_{pj} under the double-exponential prior is $\mathrm{E}(\omega'_{pj} \,|\, Y)_{lasso} = \hat{\omega}'_{pj} + \frac{\mathrm{d}}{\mathrm{d}\hat{\omega}'_{pj}}\log m_{lasso}(\hat{\omega}'_{pj})$*, where $m_{lasso}(\hat{\omega}'_{pj})$ is the marginal distribution of $\hat{\omega}'_{pj}$ under the double-exponential prior. Moreover,* $lim_{|\hat{\omega}'_{pj}| \to \infty}\frac{\mathrm{d}}{\mathrm{d}\hat{\omega}'_{pj}}\log m_{lasso}(\hat{\omega}'_{pj}) = \pm a$*, where $a = 2^{1/2}/nv$ and v is the variance of the double-exponential prior.*

(3) The squared scaled least squares estimate follows a noncentral Chi-squared distribution with one degree of freedom, i.e. $\hat{\omega}'^2_{pj} \,|\, Y_{(-p)} \sim Noncentral\chi^2(1, \omega'^2_{pj})$, and by the scaling, $\omega'^2_{pj} = \omega^2_{pj0}\omega^{-1}_{pp0}\{(Y'_{(-p)}Y_{(-p)})^{-1}_{jj}\}^{-1}$. When $n > p - 1$, $\{(Y'_{(-p)}Y_{(-p)})^{-1}_{jj}\}^{-1} \sim Gamma((n - p + 2)/2, 2(\omega_{jj0} - \omega^2_{pj0}/\omega_{pp0})^{-1})$.

Proof of Theorem 3.4.1 is in Appendix B. A very brief introduction to the Compound Confluent Hypergeometric (CCH) distribution and the upper bound of $\mathrm{E}(Z)$, where $Z \sim CCH(1, 1/2, 1, s, 1, \theta)$, can be found in Chapter 2. Part (1) of Theorem 3.4.1 states that given the data, the element-wise graphical horseshoe estimate is close to (in fact $O(1/\hat{\omega}'_{pj})$ away from) the unbiased least squares estimate when $\hat{\omega}'^2_{pj}$

is large, for any fixed global shrinkage parameter θ_{pj}. This property of the posterior mean is a consequence of the half-Cauchy distribution in the horseshoe prior. One may notice that the parameter θ_{pj} in the CCH distribution depends on the data. However, the global shrinkage parameter τ can be estimated to control θ_{pj} and $\mathrm{E}(Z_{pj})$, so that the graphical horseshoe estimate has the desired shrinkage.

On the other hand, Part (2) of Theorem 3.4.1 asserts that the posterior mean estimate of Bayesian graphical lasso does not converge to the unbiased least squares estimate for any finite n, even when $\hat{\omega}'_{pj}$ is large. In addition, the term a varies inversely with the global shrinkage parameter in the double-exponential prior and tends to be large in sparse cases (Carvalho et al., 2009). Therefore, in sparse cases, the posterior mean estimate of Bayesian graphical lasso tends to be further away from the unbiased least squares estimate.

Part (3) of Theorem 3.4.1 implies the condition that $\hat{\omega}'^2_{pj}$ is large is met with high probability when sample size is large. The parameter $\hat{\omega}'^2_{pj}$ has a noncentral χ^2 distribution with noncentrality parameter ω'^2_{pj} and 1 degree of freedom. The noncentrality parameter ω'^2_{pj} equals to a constant $\omega^2_{pj0}\omega^{-1}_{pp0}$ times a gamma distributed variable. This gamma distributed variable $\{(Y'_{(-p)}Y_{(-p)})^{-1}_{jj}\}^{-1}$ has mean proportional to $n-p+2$ and mode proportional to $n-p$. Therefore, when $\omega_{pj0} \neq 0$ and $n \gg p$, both ω'^2_{pj} and $\hat{\omega}'^2_{pj}$ are large with high probability, and the graphical horseshoe estimate of ω'_{pj} is $O(1/\hat{\omega}'_{pj})$ away from the unbiased least squares estimate.

To summarize the main implications of Theorem 3.4.1, when $n \gg p$ and the true parameter is nonzero, the graphical horseshoe estimate is close to an unbiased estimator with high probability, while the posterior mean estimate of Bayesian graphical lasso is not. When the sample size is sufficiently large, the bias of graphical horseshoe estimate is low for nonzero elements even though the method shrinks zero elements heavily. The theorem depends on the least squares estimate, which does not exist when $n < p$. However, graphical horseshoe is a shrinkage method that gives a stable estimate even when $n < p$. The bias of graphical horseshoe estimate is affected by the constant $\omega^2_{pj0}\omega^{-1}_{pp0}$, which implies that bias would be small when $\omega^2_{pj0}\omega^{-1}_{pp0}$ is large even in a $n < p$ case. In Section 3.5, we numerically demonstrate the error of graphical horseshoe estimates under various situations, with both $n > p$ and $n < p$.

3.5 Simulation Study

In this section, simulations are performed to compare the graphical horseshoe, graphical lasso, graphical SCAD, and Bayesian graphical lasso estimators. In the first example, we consider $p = 100$ features and $n = 50$ observations. The precision matrix Ω_0 is taken to be sparse with diagonal elements set to one and one of the following three patterns for off-diagonal elements (Friedman et al., 2010):

Random. Each off-diagonal element is randomly set to $\omega_{ij} < 0$ (corresponding to positive partial correlations) with probability 0.01, where the magnitude of nonzero off-diagonal elements is uniformly selected between -1 and -0.2. For these simulations, we consider 35 nonzero elements in Ω_0 with values ranging between -0.8397 and -0.2044.

Hubs. The rows/columns are partitioned into disjoint groups $\{G_k\}_1^K$. Each group has a row k where off-diagonal elements are taken to be $\omega_{ik} = 0.25$ (corresponding to negative partial correlations) for $i \in G_k$ and $\omega_{ij} = 0$ otherwise. We consider 10 groups and 10 members within each group, giving 90 nonzero off-diagonal elements in Ω_0.

Cliques. The rows/columns are partitioned into disjoint groups and $\omega_{ij:i,j \in G_k, i \neq j}$ are set to -0.45 corresponding to a positive partial correlation case and to 0.75 corresponding to a negative partial correlation case. We again consider 10 groups but only three members within each group, resulting in 30 nonzero off-diagonal elements in Ω_0.

In our second and third examples, we consider $p = 100, n = 120$ and $p = 200, n = 120$, using the sparsity structures above. The $p = 100, n = 120$ case uses the same precision matrix as the $p = 100, n = 50$ case. For the $p = 200, n = 120$ case, all settings for the precision matrix are kept the same except in the random structure where each off-diagonal ω_{ij} has a probability of 0.002 of being nonzero. The results for the three examples are summarized in Tables 3.1, 3.2 and 3.3, respectively.

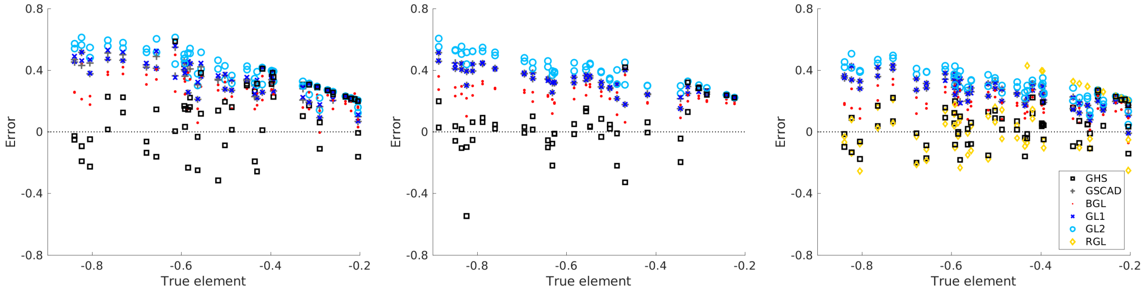
For each choice of Ω_0, 50 data sets are generated and Ω is estimated using the graphical SCAD, graphical horseshoe, frequentist graphical lasso with and without penalization on diagonal elements, and the Bayesian graphical lasso. Our graphical horseshoe estimator is implemented in MATLAB (2018). We use the posterior mean as our estimate. MATLAB code by Wang (2012) is used for the graphical SCAD and Bayesian graphical lasso. The frequentist graphical lasso is implemented using the package "glasso" (Friedman et al., 2018) in R (R Core Team, 2018). Tuning parameters in the graphical lasso and graphical SCAD are selected by five-fold cross validation using log likelihood. In the case where $p = 100$ and $n = 120$, an estimate of Ω based on the unpenalized likelihood function is feasible, and we also include a refitted graphical lasso in this comparison. For the refitted graphical lasso, the graphical lasso is first applied for variable selection, then the selected parameters in Ω are refitted using the graphical lasso algorithm, with the tuning parameter fixed at zero (i.e. no penalization). For the refitted graphical lasso, log likelihood of the final unpenalized estimate is used to calculate the cross validation score, used in selecting the tuning parameter in the variable selection step.

Stein's loss of the estimated precision matrix Ω (which equals to 2 times the Kullback–Leibler divergence of Ω from Ω_0), Frobenius norm of $\Omega - \Omega_0$, true positive rate (TPR), and false positive rate (FPR) are calculated. Since both graphical SCAD and graphical lasso provide variable selection in their estimates (i.e., some of the elements are estimated to be zero), their variable selection results are calculated using the number of nonzero estimates. Graphical horseshoe and the Bayesian graphical lasso, however, are shrinkage methods and do not estimate elements to be exactly equal to zero. For these two methods, we use the symmetric central 50% posterior credible intervals for variable selection. That is, if the 50% posterior credible interval of an off-diagonal element of Ω does not contain zero, that element is considered a discovery, and vice versa. For each statistic, we report the mean and standard deviation computed over 50 data sets. We also report the average CPU time in minutes for each method. We provide additional simulation results, a larger dimensional setting with $p = 400$ and $n = 120$, and MCMC convergence diagnostics in Appendix B. The simulations were performed on a server with 1TB of RAM, and 20 total CPU cores from a pair of Intel Xeon E5-2660 v3 CPUs at 2.60GHz, with 10 cores each.

3.5.1 Estimation

From Tables 3.1, 3.2 and 3.3, the graphical horseshoe estimate has the smallest Stein's loss and the smallest Frobenius norm (F norm) among the regularization methods considered, in eleven and ten out of twelve cases, respectively. When $p = 100$ and $n = 120$, an estimation of Ω based on the unpenalized likelihood is feasible, since $n > p$. In this case, the refitted graphical lasso, based on variable selection by graphical lasso and unpenalized estimation of the selected variables, performs well (Table 3.2). However, the graphical horseshoe performs comparably to the refitted graphical lasso, except for the hubs structured precision matrix. The graphical horseshoe is expected to perform well when the precision matrix is sparse and the absolute values of scaled nonzero elements are large. In our simulations, the hubs structure is the least sparse with small nonzero elements, and the cliques structured matrix with negative partial correlations is the sparsest with larger nonzero elements. Simulation results confirm that the advantage of graphical horseshoe is indeed larger in the cliques structure with negative partial correlations, and smaller in the hubs structure, if there is an advantage at all.

In the simulations, the graphical SCAD and frequentist graphical lasso with penalized diagonal terms are comparable in terms of Stein's loss and Frobenius norm. The frequentist graphical lasso with unpenalized diagonal terms performs somewhat worse. The Bayesian graphical lasso is by far the worst in estimation, especially in terms of Stein's loss, in accordance with the results in Section 3.3.



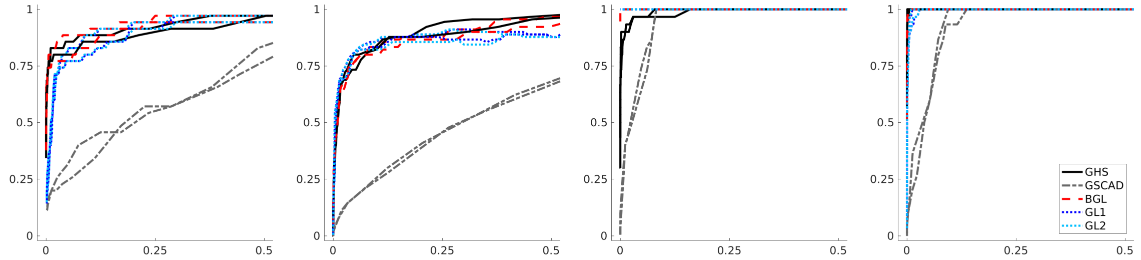
(a) $p = 100, n = 50$ (b) $p = 200, n = 120$ (c) $p = 100, n = 120$

Figure 3.1.: Errors of nonzero elements of estimated precision matrix by frequentist graphical lasso with penalized diagonal elements (GL1), frequentist graphical lasso with unpenalized diagonal elements (GL2), refitted graphical lasso (RGL), graphical SCAD (GSCAD), Bayesian graphical lasso (BGL), and graphical horseshoe (GHS). Random structure of precision matrix. Estimates using two representative data sets in simulations.

Figure 3.1 shows the estimation errors in nonzero off-diagonal elements for the random structured precision matrix. As a plot showing estimation errors using all 50 data sets will be hard to read, errors in only two representative data sets in simulations are shown in each plot. Scatterplots indicate that the errors in the graphical horseshoe estimates are randomly scattered around zero while the graphical lasso, graphical SCAD and Bayesian graphical lasso always shrink the estimates toward zero. When $p = 100$ and $n = 120$, graphical horseshoe estimates have errors comparable to the unpenalized refitted graphical lasso errors. Graphical horseshoe estimates also have smaller errors than the other estimates, especially when absolute values of true elements are large and when $n - p$ is large. These results agree with the theory and discussion in Section 3.4.

3.5.2 Variable Selection

van der Pas et al. (2017a) studied the coverage properties of marginal credible intervals under the horseshoe prior, for a sparse normal means problem. They found that the model selection procedure using credible intervals under the horseshoe prior is conservative. That is, few zero parameters in the model are falsely selected, but some of the signals are not selected. In simulations, they also discovered that the lengths of credible intervals under the horseshoe prior adapt to the signal size. In other words, parameters with larger nonzero means have wider credible intervals. In order to reduce false negatives due to wide credible intervals for large signals, we use the 50% credible interval for variable selection. By the conservative property of the procedure, false positives would be controlled under this criterion. This choice also agrees with the median probability model suggested by Barbieri and Berger (2004).



(a) Random structure (b) Hubs structure (c) Cliques structure, positive (d) Cliques structure, negative

Figure 3.2.: Receiver operating characteristic (ROC) curves of estimates by frequentist graphical lasso with penalized diagonal elements (GL1), frequentist graphical lasso with unpenalized diagonal elements (GL2), graphical SCAD (GSCAD), Bayesian graphical lasso (BGL), and graphical horseshoe (GHS), for precision matrix with random structure, hubs structure, cliques structure with positive partial correlations, and cliques structure with negative partial correlations. $p = 100$ and $n = 50$. The true positive rate is shown on the y-axis, and the false positive rate is shown on the x-axis. ROC curves of two representative data sets in simulations.

True and false positive rates are reported in Table 3.1, 3.2 and 3.3. True positive rates under the graphical horseshoe prior are indeed lower when $p = 100$ and $n = 50$. However, the true positive rate for the graphical horseshoe improves greatly when $n = 120$. The graphical horseshoe also has lower false positive rates than the other regularization methods. Figure 3.2 shows the ROC curves, plotting true positive rate against false positive rate for variable selection results, when $p = 100$ and $n = 50$. To avoid overlapping curves, ROC curves of two representative data sets were plotted in each case. The ROC curves for the graphical lasso and graphical SCAD are generated by estimating the precision matrix with a sequence of various tuning parameters. The ROC curves for the graphical horseshoe and Bayesian graphical lasso are generated by varying the length of posterior credible intervals, from 1% to 99%. Except for the graphical SCAD, which always performs worse in variable selection, the other methods have similar ROC curves. For the random and cliques structured matrix with negative partial correlations, the ROC curve of the graphical horseshoe is slightly closer to the y-axis. Although the difference is minute in terms of the false positive rate, such a difference could greatly increase precision, the rate of true positives among all discoveries, in a sparse model. When most parameters are zero, a little increase in false discovery rate greatly increases the number of false discoveries and decreases precision. In our simulations, the precision for the graphical horseshoe is almost always higher than 0.85, while the precision for other regularization methods is usually less than 0.3, making the variable selection results not very useful in applications. Additional numerical results on precision of the estimates in simulations can be found in Tables B.1, B.2 and B.3 of Appendix B.

Finally, it is worth noting that there need not to be a single variable selection result by a Bayesian model. In applications, researchers can obtain posterior samples from the graphical horseshoe or Bayesian graphical lasso, and gradually change the length of credible intervals for variable selection to have a sequence of results following the ROC curve. Such a procedure allows the researcher to start from a low false positive rate and moderate true positive rate, and gradually increase the true positive rate while having some control on precision.

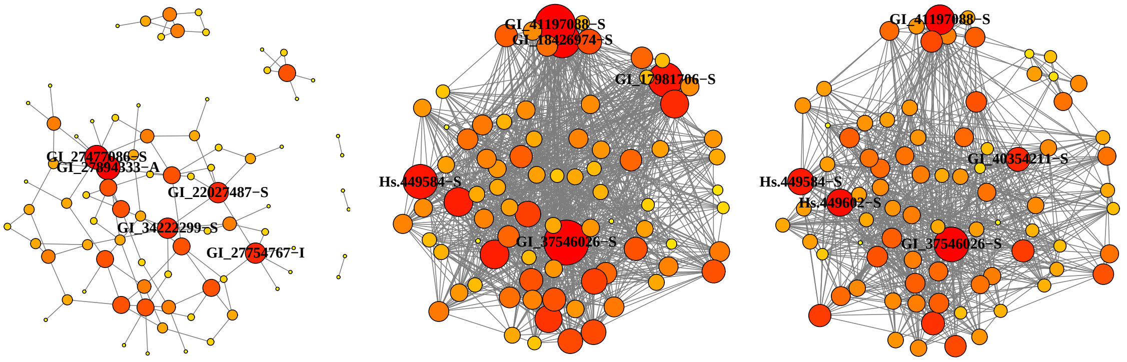
3.6 Analysis of Human Gene Expression Data

We analyze the expression of 100 genes in 60 unrelated individuals of Northern and Western European ancestry from Utah (CEU). A description of the data set can be found in Bhadra and Mallick (2013). For this analysis, we assume that the gene expressions of the individuals in this data set are identically distributed with a multivariate normal distribution. We analyze centered gene expressions using graphical horseshoe, graphical lasso with penalized diagonal elements, graphical SCAD, and Bayesian graphical lasso. Tuning parameters in graphical lasso and graphical SCAD are selected by five–fold cross validation, using log likelihood. For graphical lasso and graphical SCAD, the existence of association between a pair of genes in terms of expression is determined by whether the corresponding element in the precision matrix is estimated to be zero. For the graphical horseshoe and Bayesian graphical lasso, we used whether zero is included in the 50% posterior credible interval.

The inferred graph by graphical horseshoe, graphical lasso and Bayesian graphical lasso are shown in Figure 3.3. The graphical horseshoe estimate has 83 vertices and 109 edges. The inferred graph has 100 vertices and 1135 edges by graphical lasso estimate, and 100 vertices and 976 edges by Bayesian graphical lasso estimate. None of the graphical SCAD estimated elements in the precision matrix is zero, so the inferred graph by graphical SCAD estimate has 100 vertices and 4950 edges. The graphs by graphical lasso and Bayesian graphical lasso show similar clusters, where every gene expression is associated with at least one other gene expression, and the major clusters are densely connected as well. On the other hand, the graphical horseshoe estimate shows unconnected and much sparser clusters of gene expressions. Our resulting network using this human gene expression data can be compared with that in Bhadra and Mallick (2013), who used the same data set in a regression setting (as opposed to the zero mean setting used by us), where the gene expressions were regressed on SNPs and the resulting network on the residual terms was plotted. Comparison of these two networks should provide an insight into which edges are "robust" to the effect of being conditioned upon the SNPs.

3.7 Conclusions

The problem of precision matrix estimation in a multivariate Gaussian model poses a challenge in high-dimensional data analysis. In this paper, we proposed the



(a) GHS (b) GL (c) BGL

Figure 3.3.: The inferred graph for the CEU data, by graphical horseshoe (GHS), frequentist graphical lasso with penalized diagonal elements (GL), and Bayesian graphical lasso (BGL) estimates. Genes that are conditionally independent of all the others are not shown. Size of node is proportional to degree within each graph, the positions of nodes are comparable across graphs.

graphical horseshoe estimator with easy implementation by a full Gibbs sampler. By using a prior with high density near the origin and a Cauchy-distributed local shrinkage parameter on each dimension, the graphical horseshoe model generates estimates close to the true distribution in Kullback–Leibler divergence and with small bias for nonzero elements. Simulations confirm that the graphical horseshoe outperforms alternative methods in various situations.

We have shown when the Kullback–Leibler divergence is under consideration, all methods eventually fail in high dimensions. In addition, the difference between sample size and feature size also affects bias. This implies that efforts should be spent on variable screening prior to analysis in order to bring the feature space to a manageable size. Although some properties of variable selection by the horseshoe prior in sparse normal means problem are known, theoretical understanding of true and false discoveries under the graphical horseshoe prior are still lacking. It will also be interesting to compare the graphical horseshoe to some recently proposed methods in graphical model estimation, for instance, the spike-and-slab lasso (Deshpande et al., 2017). Use of other priors exhibiting properties similar to the horseshoe, such as the horseshoe+ (Bhadra et al., 2017) or the Dirichlet–Laplace (Bhattacharya et al., 2015) should also be explored.

Table 3.1.: Mean (sd) Stein's loss, Frobenius norm, true positive rates and false positive rates of precision matrix estimates over 50 data sets generated by multivariate normal distributions with precision matrix Ω_0, where $p = 100$ and $n = 50$. The precision matrix is estimated by frequentist graphical lasso with penalized diagonal elements (GL1), frequentist graphical lasso with unpenalized diagonal elements (GL2), graphical SCAD (GSCAD), Bayesian graphical lasso (BGL), and graphical horseshoe (GHS). The best performer in each row is shown in bold. Average CPU time is in minutes.

			Random					Hubs		
nonzero pairs			35/4950					90/4950		
nonzero elements			$\sim -\text{Unif}(0.2, 1)$					0.25		
$p = 100, n = 50$	GL1	GL2	GSCAD	BGL	GHS	GL1	GL2	GSCAD	BGL	GHS
Stein's loss	10.20	13.42	10.05	80.92	**6.44**	10.12	12.78	**10.01**	77.85	12.56
	(0.53)	(1.06)	(0.55)	(1.63)	(0.85)	(0.53)	(0.96)	(0.50)	(1.66)	(1.04)
F norm	4.33	5.30	4.31	5.58	**3.31**	3.95	4.63	**3.94**	5.97	3.96
	(0.18)	(0.20)	(0.16)	(0.26)	(0.29)	(0.13)	(0.18)	(0.14)	(0.30)	(0.27)
TPR	.8246	.7097	**.9977**	.8709	.5903	.8649	.7333	**.9987**	.8513	.2687
	(.0520)	(.0620)	(.0078)	(.0470)	(.0537)	(.0443)	(.0751)	(.0053)	(.0378)	(.0764)
FPR	.0947	.0374	.9955	.1055	**.0004**	.0919	.0281	.9976	.1189	**.0013**
	(.0141)	(.0070)	(.0102)	(.0059)	(.0003)	(.0130)	(.0086)	(.0069)	(.0058)	(.0005)
Avg CPU time	0.30	0.35	6.24	40.94	38.32	0.14	0.16	4.01	35.44	41.58
			Cliques positive					Cliques negative		
nonzero pairs			30/4950					30/4950		
nonzero elements			-0.45					0.75		
$p = 100, n = 50$	GL1	GL2	GSCAD	BGL	GHS	GL1	GL2	GSCAD	BGL	GHS
Stein's loss	9.16	14.16	8.99	81.58	**5.87**	11.00	14.37	10.90	81.27	**6.28**
	(0.55)	(1.06)	(0.52)	(2.51)	(0.93)	(0.43)	(1.02)	(0.43)	(1.98)	(1.09)
F norm	3.75	5.01	**3.71**	5.44	3.81	6.00	6.86	5.99	6.51	**3.64**
	(0.16)	(0.16)	(0.17)	(0.33)	(0.41)	(0.14)	(0.16)	(0.14)	(0.20)	(0.36)
TPR	**1**	**1**	**1**	**1**	.7487	.9993	.9880	**1**	.9993	.9733
	(0)	(0)	(0)	(0)	(.0427)	(.0047)	(.0221)	(0)	(.0047)	(.0421)
FPR	.0900	.0255	.9901	.1014	**.0003**	.0922	.0279	.9752	.1161	**.0010**
	(.0098)	(.0056)	(.0177)	(.0052)	(.0003)	(.0135)	(.0084)	(.0219)	(.0051)	(.0005)
Avg CPU time	0.24	0.28	4.52	34.45	41.65	0.18	0.20	6.91	33.88	41.05

Table 3.2.: Mean (sd) Stein's loss, Frobenius norm, true positive rates and false positive rates of precision matrix estimates over 50 data sets generated by multivariate normal distributions with precision matrix Ω_0, where $p = 100$ and $n = 120$. The precision matrix is estimated by frequentist graphical lasso with penalized diagonal elements (GL1), frequentist graphical lasso with unpenalized diagonal elements (GL2), refitted graphical lasso (RGL), graphical SCAD (GSCAD), Bayesian graphical lasso (BGL), and graphical horseshoe (GHS). The best performer in each row is shown in bold. Average CPU time is in minutes.

	Random						Hubs					
nonzero pairs	35/4950						90/4950					
nonzero elements	$\sim -\text{Unif}(0.2, 1)$						0.25					
$p = 100, n = 120$	GL1	GL2	RGL	GSCAD	BGL	GHS	GL1	GL2	RGL	GSCAD	BGL	GHS
Stein's loss	5.32	6.90	3.84	5.29	43.08	**2.15**	5.34	6.53	**3.92**	5.29	43.07	5.12
	(0.27)	(0.51)	(0.46)	(0.26)	(0.82)	(0.27)	(0.28)	(0.47)	(0.70)	(0.26)	(0.76)	(0.49)
F norm	3.37	4.12	2.26	3.36	3.94	**1.91**	3.04	3.49	**2.20**	3.02	4.28	2.54
	(0.13)	(0.16)	(0.17)	(0.13)	(0.12)	(0.14)	(0.09)	(0.12)	(0.15)	(0.09)	(0.14)	(0.12)
TPR	.9486	.8794	.6497	**.9994**	.9760	.8149	.9936	.9844	.8376	**.9998**	.9938	.8671
	(.0316)	(.0384)	(.0658)	(.0040)	(.0233)	(.0397)	(.0078)	(.0154)	(.0617)	(.0016)	(.0072)	(.0396)
FPR	.1029	.0442	.0109	.9983	.1689	**.0005**	.1029	.0431	**.0015**	.9988	.1872	.0033
	(.0150)	(.0077)	(.0029)	(.0055)	(.0066)	(.0003)	(.0161)	(.0093)	(.0009)	(.0029)	(.0066)	(.0011)
Avg CPU time	0.23	0.25	0.46	62.65	29.36	46.59	0.19	0.20	0.33	73.36	30.74	45.82
	Cliques positive						Cliques negative					
nonzero pairs	30/4950						30/4950					
nonzero elements	-0.45						0.75					
$p = 100, n = 120$	GL1	GL2	RGL	GSCAD	BGL	GHS	GL1	GL2	RGL	GSCAD	BGL	GHS
Stein's loss	4.60	7.14	**1.26**	4.57	42.69	1.89	6.01	7.49	**1.61**	5.96	44.17	1.78
	(0.25)	(0.62)	(0.20)	(0.25)	(0.94)	(0.30)	(0.21)	(0.41)	(0.49)	(0.20)	(0.81)	(0.21)
F norm	2.82	3.85	**1.63**	2.80	3.83	1.98	4.98	5.70	**1.80**	4.97	4.92	1.86
	(0.11)	(0.16)	(0.16)	(0.11)	(0.17)	(0.22)	(0.10)	(0.12)	(0.33)	(0.09)	(0.09)	(0.16)
TPR	**1**	**1**	**1**	**1**	**1**	.9840	**1**	**1**	.9947	**1**	**1**	**1**
	(0)	(0)	(0)	(0)	(0)	(.0236)	(0)	(0)	(.0170)	(0)	(0)	(0)
FPR	.0999	.0287	**.0003**	.9979	.1580	.0004	.0141	.0413	.0010	.9939	.1776	**.0008**
	(.0078)	(.0064)	(.0004)	(.0061)	(.0074)	(.0003)	(.0100)	(.0088)	(.0008)	(.0073)	(.0068)	(.0004)
Avg CPU time	0.16	0.17	0.62	4.60	32.63	37.96	0.10	0.11	0.45	10.66	32.55	37.57

Table 3.3.: Mean (sd) Stein's loss, Frobenius norm, true positive rates and false positive rates of precision matrix estimates over 50 data sets generated by multivariate normal distributions with precision matrix Ω_0, where $p = 200$ and $n = 120$. The precision matrix is estimated by frequentist graphical lasso with penalized diagonal elements (GL1), frequentist graphical lasso with unpenalized diagonal elements (GL2), graphical SCAD (GSCAD), Bayesian graphical lasso (BGL), and graphical horseshoe (GHS). The best performer in each row is shown in bold. Average CPU time is in minutes.

			Random					Hubs		
nonzero pairs			29/19900					180/19900		
nonzero elements			$\sim -\text{Unif}(0.2, 1)$					0.25		
$p = 200, n = 120$	GL1	GL2	GSCAD	BGL	GHS	GL1	GL2	GSCAD	BGL	GHS
Stein's loss	10.06	15.80	9.96	116.61	**2.94**	12.49	15.12	12.40	122.87	**11.67**
	(0.40)	(0.99)	(0.39)	(1.69)	(0.34)	(0.45)	(0.80)	(0.42)	(1.35)	(0.76)
F norm	4.49	5.97	4.45	6.76	**2.44**	4.61	5.27	4.59	7.10	**3.74**
	(0.14)	(0.17)	(0.14)	(0.18)	(0.16)	(0.10)	(0.15)	(0.08)	(0.16)	(0.14)
TPR	.9476	.8393	**1**	.9855	.8421	.9911	.9773	**1**	.9917	.7754
	(.0370)	(.0301)	(0)	(.0232)	(.0369)	(.0065)	(.0132)	(0)	(.0060)	(.0323)
FPR	.0514	.0159	.9951	.1035	**.0001**	.0657	.0257	.9997	.1197	**.0011**
	(.0065)	(.0021)	(.0095)	(.0031)	(<.0001)	(.0053)	(.0064)	(.0002)	(.0027)	(.0002)
Avg CPU time	0.03	0.04	562.13	934.53	1.44e+3	0.03	0.03	266.00	750.57	767.55
			Cliques positive					Cliques negative		
nonzero pairs			60/19900					60/19900		
nonzero elements			-0.45					0.75		
$p = 200, n = 120$	GL1	GL2	GSCAD	BGL	GHS	GL1	GL2	GSCAD	BGL	GHS
Stein's loss	11.59	17.79	11.53	124.93	**4.09**	14.56	18.12	14.50	126.44	**3.77**
	(0.37)	(0.84)	(0.34)	(1.69)	(0.39)	(0.29)	(0.80)	(0.29)	(1.45)	(0.37)
F norm	4.44	5.98	4.44	6.29	**2.97**	7.61	8.58	7.60	7.94	**2.69**
	(0.09)	(0.13)	(0.07)	(0.16)	(0.21)	(0.06)	(0.14)	(0.07)	(0.10)	(0.18)
TPR	**1**	**1**	**1**	**1**	.9633	**1**	**1**	**1**	**1**	**1**
	(0)	(0)	(0)	(0)	(.0226)	(0)	(0)	(0)	(0)	(0)
FPR	.0663	.0172	.9969	.0986	**.0002**	.0636	.0229	.9919	.1155	**.0004**
	(.0044)	(.0027)	(.0051)	(.0030)	(.0001)	(.0039)	(.0039)	(.0072)	(.0027)	(.0001)
Avg CPU time	0.02	0.03	192.46	1.64e+03	1.70e+03	0.07	0.04	466.89	1.06e+03	847.92

4. JOINT MEAN–COVARIANCE ESTIMATION VIA THE HORSESHOE WITH AN APPLICATION IN GENOMIC DATA ANALYSIS

4.1 Introduction

Multivariate regression is ubiquitous in quantitative disciplines such as finance, econometrics, and chemometrics. In recent years, multivariate regression has also been used in genomics, most notably in expression quantitative trait loci (eQTL) analysis, where the high dimensionality of the data necessitates the use of regularization methods and poses both theoretical and computational challenges. An eQTL analysis typically involves simultaneously regressing the expression levels of multiple genes on multiple markers or regions of genetic variation. Early studies have shown that each gene expression level is expected to be affected by only a few genomic regions (Schadt et al., 2003; Brem and Kruglyak, 2005) so that the regression coefficients in this application are expected to be sparse. In addition, the expression levels of multiple genes have been shown to possess a sparse network structure (Leclerc, 2008). Therefore, an eQTL analysis, if formulated as a multivariate regression problem, requires sparse estimates of both the regression coefficients and the elements of the error inverse covariance matrix.

In multivariate regression problems with correlated error matrices, joint estimation of regression coefficients are known to improve efficiency. Zellner (1962) proposed the seemingly unrelated regression (SUR) framework where the error correlation structure in multiple responses is leveraged to achieve a more efficient estimator of the regression coefficients compared to separate least squares estimators. Holmes et al. (2002) adopted the SUR framework in Bayesian regressions. However, these early methods in the SUR framework considered a relatively modest dimension of the responses, and did not encourage sparse estimates of the regression coefficients or of the error inverse covariance matrix. Therefore, these methods can not be applied directly to analyze modern genomic data. More recently, both Bayesian and frequentist approaches have also been developed for sparse, high-dimensional SUR settings. Precise descriptions of these competing approaches and understanding their strengths and limitations require some mathematical formalism. This is reserved for Section 4.2.

In this article, we propose a fully Bayesian solution for high-dimensional SUR problems with an algorithm for efficient exploration of the posterior. We impose the horseshoe prior (Carvalho et al., 2010) on the regression coefficients, and the graphical horseshoe prior (Chapter 3) on the precision matrix. In univariate normal regressions, the horseshoe prior has been shown to possess many attractive theoretical properties, including improved Kullback–Leibler risk bounds (Carvalho et al., 2010), asymptotic optimality in testing under 0-1 loss (Datta and Ghosh, 2013), minimaxity in estimation under the ℓ_2 loss (van der Pas et al., 2014), and improved risk properties in linear regression (Chapter 2). The graphical horseshoe prior inherit the properties of improved Kullback–Leibler risk bounds, and nearly unbiased estimators, when applied to precision matrix estimation (Chapter 3).

The beneficial theoretical and computational properties of horseshoe (HS) and graphical horseshoe (GHS) are combined in our proposed method, resulting in a prior that we term HS-GHS. The proposed method is fully Bayesian, so that the posterior distribution can be used for uncertainty quantification, which in the case of horseshoe is known to give good frequentist coverage (van der Pas et al., 2017a). For estimation, we derive a full Gibbs sampler, inheriting the benefits of automatic tuning and no rejection that come with it. The complexity of the proposed algorithm is linear in the number of covariates and cubic in the number of responses. To our knowledge, this is the first fully Bayesian algorithm in an SUR setting with a linear scaling in the number of covariates that allows arbitrary sparsity patterns in both the regression coefficients and the error precision matrix.

The rest of this article is organized as follows. Section 4.2 formulates the problem and describes previous works in high-dimensional settings, with brief descriptions of their respective strengths and limitations. Section 4.3 describes our proposed HS-GHS model and estimation algorithm. Section 4.4 discusses theoretical properties in terms of Kullback–Leibler divergence between the true sampling density and the marginal density under the HS-GHS prior. In Section 4.5, we evaluate the performance of our model in four simulation settings and compare them with results by competing methods described in Section 4.2. Section 4.6 describes an application in an eQTL analysis problem. We conclude by identifying some possible directions for future investigations.

4.2 Problem Formulation and Related Works in High-Dimensional Joint Mean–Covariance Modeling

Consider regressing responses $Y_{n\times q}$ on predictors $X_{n\times p}$, where n is the sample size, p is the number of features, and q is the number of possibly correlated outcomes. A reasonable parametric linear model is of the form $Y_{n\times q} = X_{n\times p}B_{p\times q} + E_{n\times q}$, where $E \sim \mathrm{MN}_{n\times q}(0, I_n, \Omega^{-1}_{q\times q})$ denotes a matrix normal random variate (Dawid, 1981) with the property that $vec(E') \sim \mathrm{N}_{nq}(0, I_n \otimes \Omega^{-1}_{q\times q})$, a multivariate normal, where $vec(A)$ converts a matrix A into a column vector by stacking the columns of A, the identity matrix of size n is denoted by I_n and $\otimes$ denotes the Kronecker product. Thus, this formulation indicates the n outcome vectors of length q are assumed uncorrelated, but within each outcome vector, the q responses share a correlation structure, which is reasonable for an eQTL analysis. The problem is then to estimate $B_{p\times q}$ and $\Omega_{q\times q}$, where both p and q can be much larger than n. We drop the subscripts denoting the dimensions henceforth when there is no ambiguity. Here Ω is also referred to as the precision matrix of the matrix variate normal, and off-diagonal zeros in it encodes a sparse conditional independence structure across the q responses, after accounting for the covariates. Of course, a consequence of the model is that one has conditionally independent (but not i.i.d.) observations of the form $Y_i \sim \mathrm{N}(X_i B, \Omega^{-1})$, for $i = 1, \ldots, n$.

The negative log likelihood function under this model, up to a constant, is

$$l(B, \Omega) = tr\{n^{-1}(Y - XB)'(Y - XB)\Omega\} - \log|\Omega|.$$

The maximum likelihood estimator for B is simply $\hat{B}^{OLS} = (X'X)^{-1}X'Y$, which does not exist when $p > n$. In addition, increasing $|\Omega|$ easily results in an infinite likelihood function. Therefore, many methods seek to regularize both B and Ω for well-behaved estimates.

One of the earliest works is the multivariate regression with covariance estimation or the MRCE method (Rothman et al., 2010), which adds independent ℓ_1 penalties to B and Ω, so the objective function is

$$(\hat{B}_{MRCE}, \hat{\Omega}_{MRCE}) = \underset{(B,\Omega)}{\operatorname{argmin}} \Big\{ l(B, \Omega) + \lambda_1 \Sigma_{k \neq l} |\omega_{kl}| + \lambda_2 \Sigma_{j=1}^{pq} |\beta_j| \Big\},$$

where ω_{kl} are the elements of Ω, β_j are the elements of vectorized B', and $\lambda_1 > 0$ and $\lambda_2 > 0$ are tuning parameters.

Cai et al. (2012) take a two-stage approach and use a multivariate extension of the Dantzig selector of Candes and Tao (2007). Let $\bar{y} = n^{-1}\Sigma_{i=1}^n y_i$, $\bar{x} = n^{-1}\Sigma_{i=1}^n x_i$, $S_{xy} = n^{-1}\Sigma_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})'$ and $S_{xx} = n^{-1}\Sigma_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})'$. The estimate of B solves the optimization problem

$$\hat{B}_{CAPME} = \underset{B}{\operatorname{argmin}} \Big\{ |B|_1 : |S_{xy} - BS_{xx}|_\infty \leq \lambda_n \Big\},$$

where λ_n is a tuning parameter, $|A|_1$ defines the elementwise ℓ_1 norm of matrix A, and $|A|_\infty$ defines the elementwise ℓ_∞ norm of A. This is equivalent to a Dantzig selector applied on the coefficients in a column-wise way. After inserting the estimator

$\hat{B}_{CAPME}$ to obtain $S_{yy} = n^{-1}\Sigma_{i=1}^{n}(y_i - \hat{B}x_i)(y_i - \hat{B}x_i)'$, one estimates Ω by the solution to the optimization problem

$$\hat{\Omega}_{CAPME} = \underset{\Omega}{\operatorname{argmin}} \Big\{ |\Omega|_1 : |I_p - S_{yy}\Omega|_\infty \leq \tau_n \Big\},$$

where τ_n is a tuning parameter. The final estimator of Ω needs to be symmetrized since no symmetry condition on Ω is imposed.

Critiques of the lasso shrinkage include that the lasso estimate is not tail robust (Carvalho et al., 2009), and at least empirically, the Dantzig selector rarely outperforms the lasso in simulations and in cancer datasets (Meinshausen et al., 2007; Zheng and Liu, 2011), indicating there is a scope of improving upon both MRCE and CAPME.

Bayesian approaches seek to implement regularization through the choice of prior, with the ultimate goal being probabilistic uncertainty quantification using the full posterior. Deshpande et al. (2017) put spike-and-slab lasso priors on the elements of B. That is, $\beta_{kj}, k = 1, \ldots, p; j = 1, \ldots, q$ is drawn *a priori* from either a 'spike' Laplace distribution with a sharp peak around zero, or a 'slab' Laplace distribution that is relatively flatter. A binary variable indicates whether a coefficient is drawn from the spike or the slab distribution. Such an element-wise prior on β_{kj} is

$$\pi(\beta_{kj}|\gamma_{kj}) \propto (\lambda_1 e^{-\lambda_1|\beta_{kj}|})^{\gamma_{kj}} (\lambda_0 e^{-\lambda_0|\beta_{kj}|})^{1-\gamma_{kj}},$$

where λ_1 and λ_0 are the parameters for the spike and slab Laplace distributions, with $\lambda_1 \gg \lambda_0$, and the binary indicator γ_{kj} follows a prior Bernoulli distribution with parameter θ, with a beta hyperprior distribution on θ with parameters a_θ and b_θ. Similarly, spike-and-slab lasso priors are put on elements ω_{lm} in Ω as well. An Expectation/Conditional Maximization (ECM) algorithm is derived for this model to obtain the posterior mode. The hyper-parameters $(\lambda_1, \lambda_0, a_\theta, b_\theta)$ for β_{kj}, and the corresponding four hyper-parameters for ω_{lm}, need to be specified in order to apply the ECM algorithm. In Deshpande et al. (2017), the Laplace distribution hyper-parameters are chosen by the trajectories of individual parameter estimates given a path of hyper-parameters, and the beta hyper-parameters were set at predefined levels. The method does not provide samples from the full posterior.

Bhadra and Mallick (2013) also consider a spike-and-slab prior on B but place Bernoulli indicators in another way. Their priors on B and Ω^{-1} are

$$B_{\gamma,G} \mid \gamma, \Omega^{-1} \sim \mathrm{MN}(0, cI_{p_\gamma}, \Omega^{-1}),$$
$$\Omega^{-1} \mid G \sim \mathrm{HIW}_G(b, dI_q),$$

where b, c, d are fixed, positive hyper-parameters and HIW denotes a hyper-inverse Wishart distribution (Dawid and Lauritzen, 1993). The indicator p_γ selects entire rows of coefficients, depending on whether $p_{\gamma_i} \neq 0,\ i = 1, \ldots, p$. Similarly, the indicator G has length $q(q-1)/2$, and selects off-diagonal elements in the covariance matrix. Elements in γ and G are independently distributed Bernoulli random variables, with hyper-parameters ω_γ and ω_G, respectively. The model allows B and Ω to be analytically integrated out to achieve fast MCMC sampling, at the expense of a somewhat restrictive assumption that a variable is selected as relevant to all of the q responses or to none of them.

Thus, it appears only a few of Bayesian shrinkage rules have been applied to joint mean and inverse covariance estimation in SUR models, and there is no fully Bayesian method that efficiently solves this problem under the assumption of arbitrary sparsity structures in B and Ω. To this effect, we propose to use the horseshoe prior which achieves efficient shrinkage in both sparse regression and inverse covariance estimation. We also propose an MCMC algorithm for sampling, without user-chosen tuning parameters.

4.3 Proposed Model and Estimation Algorithm

We define β to be the vectorized coefficient matrix, or $\beta = vec(B') = [B_{11}, ..., B_{1q}, ..., B_{p1}, ..., B_{pq}]'$. To achieve shrinkage of the regression coefficients, we put horseshoe prior on β. That is,

$$\beta_j \sim \mathrm{N}(0, \lambda_j^2 \tau^2);\ j = 1, ..., pq,$$
$$\lambda_j \sim C^+(0,1),\ \tau \sim C^+(0,1),$$

$C^+(0,1)$ denotes the standard half-Cauchy distribution with density $p(x) \propto (1 + x^2)^{-1};\ x > 0$. The normal scale mixtures on β with half-Cauchy hyperpriors λ_j and τ is known as the horseshoe prior (Carvalho et al., 2010), presumably due to the shape of the induced prior on the shrinkage factor. Similarly, to encourage sparsity in the off-diagonal elements of Ω, we use the graphical horseshoe prior for Gaussian graphical models (Chapter 3), defined as

$$\omega_{kl:k>l} \sim \mathrm{N}(0, \iota_{kl}^2 \kappa^2);\ k, l = 1, ..., q,$$
$$\iota_{kl} \sim C^+(0,1),\ \kappa \sim C^+(0,1),\ \omega_{kk} \propto \text{constant},$$

where $\Omega = \{\omega_{kl}\}$, and the prior mass is truncated to the space of $q \times q$ positive definite matrices $\mathcal{S}_q^+$. In this model, ι_{kl} and κ induce shrinkage on the off-diagonal elements in Ω.

Full Bayesian samplers have been proposed for regressions using the horseshoe prior for the linear regression model with i.i.d. error terms (Makalic and Schmidt, 2016; Bhattacharya et al., 2016). However, these samplers cannot be applied to the current problem due to the correlation in the error covariance. To transform the data into a model where sampling is possible, we reshape the predictors and responses. Let $\tilde{y} = vec(\Omega^{1/2} Y')$, and $\tilde{X} = X \otimes \Omega^{1/2}$. Simple algebra shows that $\tilde{y} \sim \mathrm{N}_{nq}(\tilde{X}\beta, I_{nq})$. In this way, the matrix variate normal regression problem is transformed into an multivariate normal regression problem, provided the current estimate Ω is known.

Next, given the current estimate of B, the graphical horseshoe sampler in Chapter 3 is leveraged to estimate Ω, using the residual term $Y_{res} = Y - XB$.

A full Gibbs sampler for the above model is given in Algorithm 2. Throughout, the shape–scale parameterization is used for all gamma and inverse gamma random variables. First, the coefficient matrix B is sampled conditional on the precision matrix Ω. We first notice that the conditional posterior of β is $\mathrm{N}((\tilde{X}'\tilde{X}+\Lambda_*^{-1})^{-1}\tilde{X}'\tilde{Y}, (\tilde{X}'\tilde{X}+\Lambda_*^{-1})^{-1})$, where $\Lambda_* = \mathrm{diag}(\lambda_j^2\tau^2), j = 1, ..., pq$. However, sampling from this normal distribution is computationally expensive because it involves computing the inverse of the $pq \times pq$ dimensional matrix $(\tilde{X}'\tilde{X} + \Lambda_*^{-1})$, with complexity $O(p^3q^3)$. Luckily, sampling β from this high-dimensional normal distribution can be solved by the fast sampling scheme proposed by Bhattacharya et al. (2016). The algorithm is exact with a complexity linear in p. Combining the fast sampling scheme for β and the variable augmentation for half-Cauchy priors using inverse gamma distributed variables (Makalic and Schmidt, 2016), we have Gibbs steps (1) to (4) in Algorithm 2. Steps (2a) to (2d) sample the coefficients $\beta = vec(B')$ using the fast sampling scheme, and Steps (3) and (4) sample the shrinkage parameters λ_j and τ, in addition to auxiliary variables, ν_j and ξ.

To sample the precision matrix Ω conditional on B, take $Y_{res} = Y - XB$ and $S = Y_{res}'Y_{res}$. Since $Y - XB \sim \mathrm{MN}(0,\ I_n,\ \Omega^{-1})$, the problem of estimating Ω given B is exactly the zero-mean multivariate Gaussian inverse covariance estimation that the graphical horseshoe solves, with details given in Algorithm 1 of Chapter 3. Therefore, Steps (6a) to (8) in Algorithm 2 follows the sampling scheme of graphical horseshoe model for sample size n, number of features q, and scatter matrix S. Steps (6a) to (6c) partitions the precision matrix and samples one column (or row) of it at a time, using a variable transformation technique first identified by Wang (2012). Then the shrinkage parameters and auxiliary variables are sampled from inverse gamma distributions in Steps (7) and (8). Chapter 3 further demonstrate that the posterior samples of Ω under the graphical horseshoe model are guaranteed to be positive definite, provided the

Algorithm 2 The HS-GHS Sampler

```
function HS-GHS($X, Y, burnin, nmc$)
    Set $n, p$ and $q$ using $\dim(X) = n \times p$ and $\dim(Y) = n \times q$.
    Initialize $\beta = \mathbf{0}_{p\times q}$ and $\Omega = I_q$.
    for $i = 1$ to $burnin + nmc$ do
        (1) Calculate $\tilde{y} = vec(\Omega^{1/2}Y')$, $\tilde{X} = X \otimes \Omega^{1/2}$
          %% Sample $\beta$ using horseshoe
        (2a) Sample $u \sim \mathrm{N}_{pq}(0, \Lambda_*)$ and $\delta \sim \mathrm{N}_{nq}(0, I_{nq})$ independently, where $\Lambda_* = \mathrm{diag}(\lambda_j^2\tau^2)$
        (2b) Take $v = \tilde{X}u + \delta$
        (2c) Solve $w$ from $(\tilde{X}\Lambda_*\tilde{X}' + I_{nq})w = \tilde{y} - v$
        (2d) Calculate $\beta = u + \Lambda_*\tilde{X}'w$
        (3) Sample $\lambda_j^2 \sim \mathrm{InvGamma}(1, 1/\nu_j + \beta_j^2/(2\tau^2))$, and $\nu_j \sim \mathrm{InvGamma}(1, 1 + 1/\lambda_j^2)$, for $j = 1, ..., pq$
        (4) Sample $\tau^2 \sim \mathrm{InvGamma}((pq + 1)/2, 1/\xi + \Sigma_{j=1}^{pq}\beta_j^2/(2\lambda_j^2))$, and $\xi \sim \mathrm{InvGamma}(1, 1 + 1/\tau^2)$
        (5) Calculate $Y_{res} = Y - XB$ and $S = Y_{res}'Y_{res}$
          %% Sample $\Omega$ using graphical horseshoe
        for $k = 1$ to $q$ do
            Partition matrices $\Omega$, $S$ to $(q-1)\times(q-1)$ upper diagonal blocks $\Omega_{(-k)(-k)}$, $S_{(-k)(-k)}$; $(q-1)\times 1$ dimensional vectors $\omega_{(-k)k}$, $s_{(-k)k}$; and scalars $\omega_{kk}$, $s_{kk}$
            (6a) Sample $\gamma \sim \mathrm{Gamma}(n/2 + 1, 2/s_{kk})$
            (6b) Sample $\upsilon \sim \mathrm{N}(-Cs_{(-k)k}, C)$ where $C = (s_{kk}\Omega_{(-k)(-k)}^{-1} + \mathrm{diag}(\iota_{(-k)k}\kappa^2)^{-1})^{-1}$ and $\iota_{(-k)k}$ is a vector of length $(q-1)$ with entries $\iota_{lk}^2, l \neq k$
            (6c) Apply transformation: $\omega_{(-k)k} = \upsilon$, $\omega_{kk} = \gamma + \upsilon'\Omega_{(-k)(-k)}^{-1}\upsilon$
            (7) Sample $\iota_{(-k)k} \sim \mathrm{InvGamma}(1, 1/\rho_{(-k)k} + \omega_{(-k)k}^2/2\kappa^2)$, and $\rho_{(-k)k} \sim \mathrm{InvGamma}(1,\ 1 + 1/\iota_{(-k)k})$
        end for
        (8) Sample $\kappa^2 \sim \mathrm{InvGamma}((\binom{q}{2}+1)/2, 1/\phi + \sum_{k,l:k<l}\omega_{kl}^2/2\iota_{kl}^2)$, and $\phi \sim \mathrm{InvGamma}(1, 1 + 1/\kappa^2)$
        Save samples if $i > burnin$
    end for
    Return MCMC samples of $\beta$ and $\Omega$
end function
```

initial value is positive definite. A MATLAB implementation, along with a simulation example, is freely available from github at *https://github.com/liyf1988/HS_GHS*.

Complexity analysis of the proposed algorithm is as follows. Once $\Omega^{1/2}$ is calculated in $O(q^3)$ time, calculating $\tilde{y}$ costs $O(nq^2)$, and calculating $\tilde{X}$ costs $O(npq^2)$. The most time consuming step is still sampling β, which is $O(n^2pq^3)$ with the fast sampling method. Nevertheless, when $n \ll p$, using the fast sampling method is considerably less computationally intensive than sampling from the multivariate normal distribution directly, which has complexity $O(p^3q^3)$. Since the complexity of the graphical horseshoe is $O(q^3)$, each iteration in our Gibbs sampler takes $O(n^2pq^3)$ time.

Although the Gibbs sampler is computation-intensive, especially compared to penalized likelihood methods, it has several advantages. First, the Gibbs sampler is automatic, and does not require cross validation or empirical Bayes methods for choosing hyperparameters. Penalized optimization methods for simultaneous estimation of mean and inverse covariance usually need two tuning parameters (Cai et al., 2012; Rothman et al., 2010; Yin and Li, 2011). Second, MCMC approximation of the the posterior distribution enables variable selection using posterior credible intervals. It is also possible to vary the length of credible intervals to assess trade-offs between false positives and false negatives in variable selection. Finally, to our knowledge this is the first fully Bayesian solution in an SUR framework with a complexity linear in p. Along with these computational advantages, we now proceed to demonstrate the proposed sampler possesses attractive theoretical properties as well.

4.4 Kullback–Leibler Divergence Bounds

Since a Bayesian method is meant to approximate an entire distribution, we provide results on Kullback–Leibler divergence between the true density (assuming there exists one) and the Bayes marginal density. Adopt the slightly non-Bayesian view that n conditionally independent observations $Y_1, \ldots, Y_n$ are available from an underlying true parametric model with parameter θ_0 and let p^n denote the *true joint density*, i.e., $p^n = \prod_{i=1}^n p(y_i; \theta_0)$. Similarly, let the marginal m^n in a Bayesian model with prior $\nu(d\theta)$ on the parameter be defined as $m^n = \int \prod_{i=1}^n q(y_i|\theta)\nu(d\theta)$, where q is the *sampling density*. If the prior on θ is such that the measure of any set according to the true density and the sampling density are not too different, then it is natural to expect p^n and m^n to merge in information as more samples are available. The following result by Barron (1988) formalizes this statement. Let $D_n(\theta) = \frac{1}{n}D(p^n||q^n(\cdot|\theta))$, where $D(\pi_1|\pi_2) = \int \log(\pi_1/\pi_2)d\pi_1$, denote the Kullback–Leibler divergence (KLD) of density π_1 with respect to π_2 and $q^n(\cdot|\theta) = \prod_{i=1}^n q(y_i|\theta)$. The set $A_\epsilon = \{\theta : D_n(\theta) < \epsilon\}$ can be thought of as a K–L information neighborhood of size ϵ, centered at θ_0. Then we have an upper bound on the KLD of p^n from m^n, in terms of the prior measure of the set D_n.

Proposition 4.4.1 *(Barron, 1988). Suppose the prior measure of the Kullback–Leibler information neighborhood is not exponentially small, i.e. for every ϵ, $r > 0$ there is an N such that for all $n > N$ one has $\nu(A_\epsilon) \geq e^{-nr}$. Then:*

$$\frac{1}{n} D(p^n || m^n) \leq \epsilon - \frac{1}{n} \log \nu(A_\epsilon).$$

The left hand side is the average Kullback–Leibler divergence between the true joint density of the samples $Y_1, ..., Y_n$ and the marginal density. The right hand side involves logarithm of the prior measure of a Kullback–Leibler information neighborhood centered at θ_0. A larger prior measure in this neighborhood of the "truth" gives a smaller upper bound for the average Kullback–Leibler divergence on the left, ensuring p^n and m^n are close in information. The following theorem shows that the HS-GHS prior, which has unbounded density at zero, achieves a smaller upper bound on the KLD when the true parameter is sparse (i.e., contains many zero elements), since it puts higher prior mass in a neighborhood of zero compared to any other prior with a bounded density function at zero.

Theorem 4.4.1 *Let $\theta_0 = (B_0, \Omega_0)$ and assume n conditionally independent observations $Y_1, \ldots, Y_n$ from the true model $Y_i \overset{ind}{\sim} \mathrm{N}(X_i B_0, \Omega_0^{-1})$, where $B_0 \in \mathbb{R}^{p \times q}$ and $\Omega_0 \in \mathcal{S}_q^+$ be the true regression coefficients and inverse covariance, respectively and X_i are observed covariates. Let β_{j0}, ω_{kl0} and σ_{kl0} denote the jth and klth element of vec(B_0), Ω_0 and $\Sigma_0 = \Omega_0^{-1}$, respectively. Suppose that $\sum_{k,l} \omega_{kl0} \propto q$, $\sum_{k,l} \sigma_{kl0} \propto q$, and $\sum_{i=1}^n (X_{i1} + \ldots + X_{ip})^2 \propto np^2$. Suppose that an Euclidean cube in the neighborhood of Ω_0 with $(\omega_{kl0} - 2/Mn^{1/2}q, \omega_{kl0} + 2/Mn^{1/2}q)$ on each dimension lies in the cone of positive definite matrices $\mathcal{S}_q^+$, where $M = \sum_{k,l} \sigma_{kl0}/q$. Then, $\frac{1}{n} D(p^n || m^n) \leq \frac{1}{n} - \frac{1}{n} \log \nu(A_{1/n})$ for all n, and:*

(1) For prior measure ν with density that is continuous, bounded above, and strictly positive in a neighborhood of zero, one obtains, $\log \nu(A_{1/n}) \propto K_1 pq \log(\frac{1}{n^{1/4}pq^{1/2}}) + K_2 q^2 \log(\frac{1}{n^{1/2}q})$, where K_1 and K_2 are constants.

(2) For prior measure ν under the HS-GHS prior, $\log \nu(A_{1/n}) > C_1(pq - |s_B|) \log\{\frac{\log(n^{1/4}pq^{1/2})}{n^{1/4}pq^{1/2}}\} + C_2|s_B|\log(\frac{1}{n^{1/4}pq^{1/2}}) + C_3(q^2 - |s_\Omega|)\log\{\frac{\log(n^{1/2}q)}{n^{1/2}q}\} + C_4|s_\Omega|\log(\frac{1}{n^{1/2}q})$, where $|s_B|$ is the number of nonzero elements in B_0, $|s_\Omega|$ is the number of nonzero elements in Ω_0, and C_1, C_2, C_3, C_4 are constants.

Proof of Theorem 4.4.1 is in Appendix B.1. Logarithm of the prior measure in the Kullback-Leibler divergence neighborhood, $\log\nu(A_{1/n})$, can be bounded by the summation of log measures in each of the $pq+q^2$ dimensions. Any Bayesian estimator with an elementwise prior satisfying conditions in Part (1) of Theorem 4.4.1 puts a prior measure proportional to $(n^{1/4}pq^{1/2})^{-1}$ in each of the pq dimensions of the regression coefficients, and a measure proportional to $(n^{1/2}q)^{-1}$ in each of the q^2 dimensions of the inverse covariance, regardless of whether the corresponding true element is zero or non-zero. Theorem 4.4.1 implies that when p and q are fixed and $n \to \infty$, the average divergence $\frac{1}{n}D(p^n||m^n)$ under any Bayesian prior converges to zero. However, when q is fixed and $p\log(n^{1/4}p)/n \to \infty$, the upper bound $n^{-1}\{1 - \log\nu(A_{1/n})\}$ diverges. Similarly, when p is fixed and $q^2\log(n^{1/2}q)/n \to \infty$, the upper bound diverges. Some common Bayesian estimators, including the double exponential prior in Bayesian lasso, induce a prior density bounded above near the origin (Carvalho et al., 2010), satisfying conditions in Part (1). Being a mixture of double exponential priors, the spike-and-slab lasso prior also satisfies conditions in Part (1).

Although the upper bound diverges when p and q are large, it can be improved by putting higher prior mass near the origin when B_0 and Ω_0 are sparse. One element where $\beta_{j0} = 0$ contributes $\log(n^{1/4}pq^{1/2})/n$ to the upper bound under a bounded prior near the origin, and $\{\log(n^{1/4}pq^{1/2}) - \log\log(n^{1/4}pq^{1/2})\}/n$ to the upper bound under the horseshoe prior. For each element where $\beta_{j0} = 0$, the HS-GHS upper bound has an extra $-O\{(\log\log n^{1/4}pq^{1/2})/n\}$ term. Similarly, for each element where $\omega_{kl0} = 0$, the HS-GHS upper bound has an extra $-O\{(\log\log n^{1/2}q)/n\}$ term. When most true coefficients and off-diagonal elements in the inverse covariance are zero, the horseshoe prior brings a non-trivial improvement on the upper bound. The theoretical findings of improved Kullback–Leibler divergence properties are extensively verified by simulations in Section 4.5.

4.5 Simulation Study

In this section, we compare the performance of the HS-GHS prior, to other multivariate normal regression methods that estimate both the regression coefficients and the precision matrix. We considered two cases, both with $p > n$. The first case has $p = 200$ and $q = 25$, and the second case has $p = 120$ and $q = 50$, and $n = 100$ in both cases. We generate a sparse $p \times q$ coefficient matrix B for each simulation setting, where 5% of the elements in B are nonzero. The nonzero elements in B follow a uniform distribution in $(-2, -0.5) \bigcup (0.5, 2)$. The precision matrix Ω is taken to be sparse with diagonal elements set to one and one of the following two patterns for the off-diagonal elements:

AR1. The precision matrix has an AR1 structure, with nonzero elements equal to 0.45.

Cliques. The rows/columns are partitioned into disjoint groups and $\omega_{kl:k,l \in G,\ k \neq l}$ are set to 0.75. When $q = 25$, we consider eight groups and three members within each group. When $q = 50$, the precision matrix contains 16 groups and each group has three members.

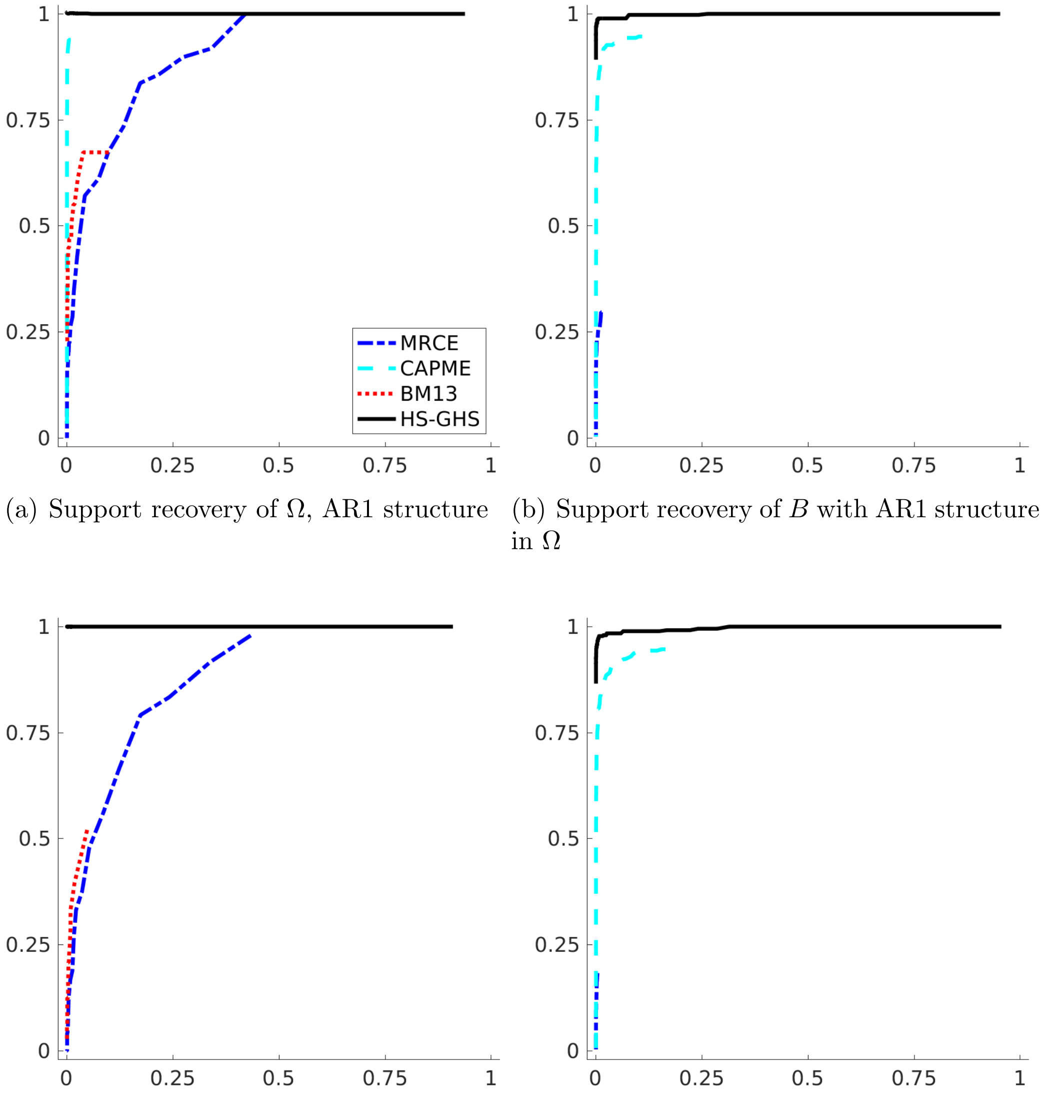
We generate an $n \times p$ design matrix X with a toeplitz covariance structure where $Cov(X_i, X_j) = 0.7^{|i-j|}$, and an $n \times q$ error matrix $E \sim \mathrm{MN}(0, I_n, \Omega^{-1})$. The $n \times q$ response matrix is set to be $Y = XB + E$. For each simulation setting, 50 data sets are generated, and B and Ω are estimated by HS-GHS, MRCE (Rothman et al., 2010), CAPME (Cai et al., 2012), and the joint high-dimensional Bayesian variable and covariance selection (BM13) by Bhadra and Mallick (2013). The proposed HS-GHS estimator was implemented in MATLAB. The MATLAB code by Bhadra and

Mallick (2013) was used for BM13, and R packages 'MRCE' and 'capme' were used for MRCE and CAPME estimates. Mean squared estimation error of regression coefficients, precision matrix; prediction mean squared error; average Kullback–Leibler divergence; and sensitivity (TP/(TP+FN)), specificity (TN/(TN+FP)), and precision (TP/(TP+FP)) in variable selection are reported. Here, TP, FP, TN and FN denote true positives, false positives, true negatives and false negatives, respectively. Variable selection for HS-GHS was performed using the 75% posterior credible interval. IN BM13, variables with posterior probability of inclusion larger than 0.5 are considered to be selected.

Results are reported in Tables 4.1 and 4.2, along with CPU times for all methods. It is evident that HS-GHS has the best overall statistical performance. Except for the mean squared error of Ω when $p = 200$, HS-GHS has the best estimation, prediction, information divergence and variable selection performance in our simulations. Although HS-GHS does not have the highest sensitivity in recovering the support of B or Ω in some cases, it has very high levels of specificity and precision. In other words, while HS-GHS may miss some true signals, it finds far fewer false positives, so that a larger proportion of true positives exists in HS-GHS findings. This property of higher precision in identifying signals is an attractive feature in applications.

In terms of the other methods, BM13 sometimes gives Ω estimate with the lowest mean squared error, but its estimate of B has higher errors, and its sensitivity for recovering the support of Ω is low. MRCE estimation of B is poor in higher dimensions, while CAPME has low mean squared errors in estimating both B and Ω. Both MRCE and CAPME are not stable in support recovery of Ω. They either tend to select every element as a positive, giving high sensitivity and low specificity, or select every element as a negative, giving zero sensitivity and high specificity.

Figure 4.1 shows the receiver operating characteristic (ROC) curves for both B and Ω, when $p = 120$ and $q = 50$. True and false positive rates were generated by varying the width of posterior credible intervals from 1% to 99% in HS-GHS, and varying the posterior inclusion probability from 1% to 99% in BM13. In MRCE and



(a) Support recovery of Ω, AR1 structure

(b) Support recovery of B with AR1 structure in Ω

(c) Support recovery of Ω, Cliques structure

(d) Support recovery of B with Cliques structure in Ω

Figure 4.1.: Receiver operating characteristic (ROC) curves of estimates by HS-GHS, joint high-dimensional Bayesian variable and covariance selection (BM13), MRCE and CAPME for $p = 120$ and $q = 50$. The true positive rate is shown on the y-axis, and the false positive rate is shown on the x-axis. One representative data set in simulations.

CAPME, variables are selected by thresholding the estimated B and Ω. For each estimated β_j and ω_{kl}, the element is considered to be a positive if its absolute value is larger than a threshold, and the threshold varies to generate a series of variable selection results. The curve for HS-GHS follows the line where true positive rate equals to one closely in all four plots. False positive rates by BM13 remains low. However, its true positive rate never exceeds 0.75. CAPME has good performance in variable selection of B, but neither CAPME or MRCE performance as well as HS-GHS in support recovery of Ω. In addition, all off-diagonal elements in Ω are estimated to be zero in the clique structured precision matrix, so CAPME cannot generate a ROC curve in this case. MCMC convergence diagnostics of the HS-GHS sampler are presented in Supplementary Section C.2 and Further simulation results complementing the results in the paper are in Supplementary Section B.4.

4.6 Yeast eQTL Data Analysis

We illustrate the HS-GHS method using the yeast eQTL data analyzed by Brem and Kruglyak (2005). The data set contains genome-wide profiling of expression levels and genotypes for 112 yeast segregants from a cross between BY4716 and RM11-1a strains of *Saccharomyces Cerevisiae*. This data set has been previously analyzed using a variety of different computational methods (Yin and Li, 2011; Cai et al., 2012; Curtis et al., 2013). The RNA was isolated and cDNA was hybridized to microarrays. The original data set contains expression values of 6216 genes assayed on each array, and genotypes at 3244 marker positions. Due to the small sample size, we only considered 54 genes in the yeast mitogen-activated protein kinase (MAPK) signalling pathway in our analysis. This pathway was provided by the Kyoto Encyclopedia of Genes and Genomes database (Kanehisa et al., 2010), and was also analyzed in Yin and Li (2011) and Cai et al. (2012).

We divide the genome into 316 groups based on linkage disequilibrium between the markers, following the method described in Curtis et al. (2013). We select the marker with largest variation within each group. Then, we apply simple screening, and find 172 markers that are marginally associated with at least one of the 54 genes with a p-value less than or equal to 0.01. We use these 172 markers as predictors and run a lasso regression on each of the 54 genes. Residuals are used to assess the normality assumption. Based on qq-plots and normality tests, we drop five genes and two yeast segregants. The final data set we use in our analysis contains 49 genes in the MAPK pathway and 172 markers in 110 yeast segregants. Marginal qq-plots of residuals and other assessments of normality assumption are provided in Supplementary Section C.4.

We divided the 110 yeast segregants into a training set containing 88 segregants, and a testing set containing 22 segregants. Coefficients of markers are estimated by HS-GHS, MRCE and CAPME using the training set, and the precision matrix of gene expressions are estimated as well. Prediction performance is measured over the testing set for each gene expression. Tuning parameters in MRCE and CAPME are selected by five-fold cross validation. Variable selection in HS-GHS are made by 75% posterial credible interval. Prediction and estimation results are summarized in Tables 4.3 and 4.4, respectively.

Out of 8428 coefficients, CAPME estimates 182 nonzero coefficients, MRCE estimates 11 nonzero coefficients, and HS-GHS estimates 15 nonzero coefficients. Prediction performance differs across these methods as well. For each gene expression, we use R-squared in the testing set, defined as (1−residual sum of squares/total sum of squares), to evaluate prediction. Many of the gene expressions cannot be predicted by any of the markers. Consequently, we only considered gene expressions that has R-squared larger than 0.1 in any of these three models. Among 22 such gene expressions, CAPME has highest R-squared among the three methods in 4 gene expressions, and HS-GHS has highest R-squared in 18 gene expressions. Average R-squared values in these 22 genes by CAPME, MRCE and HS-GHS prediction are 0.1327, 0.0063, 0.2771, respectively.

We also examined the 15 nonzero coefficients estimated by HS-GHS. CAPME estimates eight of these 15 coefficients to be nonzero, and CAPME estimates always have smaller absolute values than HS-GHS estimates. In HS-GHS estimates, the genes SWI4 and SSK2 are associated with three markers each, and FUS1 is associated with two markers. The remaining gene expressions are associated with zero or one marker. One marker on Chromosome 3, location 201166 is associated with four gene expressions (SWI4, SHO1, BCK1, SSK2), and it has the largest effect sizes among HS-GHS and CAPME estimated coefficients. This location is also identified as an eQTL hot spot by Zhu et al. (2008). In addition, a marker on Chromosome 5 and a marker on Chromosome 14 in HS-GHS nonzero estimates also correspond to two other eQTL hot spots given by Zhu et al. (2008).

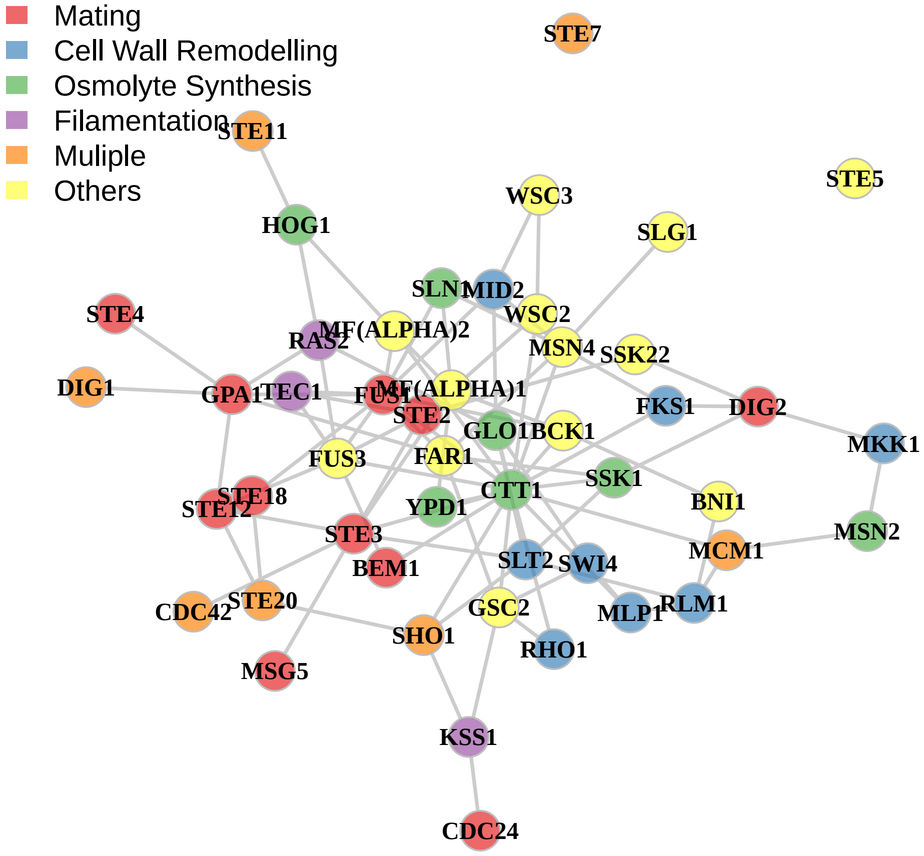


Figure 4.2.: The inferred graph for gene expressions in the MAPK pathway by the HS-GHS estimate. Vertex colors indicate functions of genes.

Out of the 1176 possible pairs among 49 genes, CAPME, MRCE, and HS-GHS estimate 702, 6, and 88 pairs to have nonzero partial covariance, respectively. We only present the HS-GHS estimated graph in Figure 4.2, while CAPME and MRCE results are in Supplementary Section C.5. Vertex colors in the graph indicate functions of genes. A current understanding of how yeast genes in the MAPK pathway respond to environmental stress and cellular signals, along with the functions of these genes, is available (Conklin et al., 2018). Figure 4.2 recovers some known structures in the MAPK pathway. For instance, STE4, STE18, GPA1, STE20, CDC42, DIG1, BEM1, FUS1, STE2, STE3 and MSG5 involved in the yeast mating process are linked in HS-GHS estimate. SLT2, SWI3, RHO1, RLM1 and MLP1 in the cell wall remodeling process, and YPD1, CTT1, GLO1 and SSK1 in the osmolyte synthesis process are also linked. It is also known that the high-osmolarity glycerol (HOG) and cell wall integrity (CWI) signalling pathways interact in yeast (Rodríguez-Peña et al., 2010), and some genes in the HOG pathway are indeed connected to genes in the CWI pathway in the HS-GHS estimate.

4.7 Conclusions

The horseshoe prior has been shown to possess many attractive theoretical properties in sparse high-dimensional regressions. In this paper, we propose the HS-GHS estimator that generates sparse estimates of regression coefficients and inverse covariance simultaneously in multivariate Gaussian regressions. We implement the estimator using a full Gibbs sampler. Simulations in high-dimensional problems confirm that HS-GHS outperforms popular alternative methods in terms of estimation of both regression coefficients and inverse covariance, and in terms of prediction. The proposed method allows arbitrary sparsity patterns B and Ω (as opposed to, say, methods based on decomposable graphs) and the number of unknown parameters inferred is $pq + q(q+1)/2$, which is indeed much larger than n in all our examples. HS-GHS also recovers the support of the coefficients and inverse covariance with higher precision. The proposed method was applied to yeast eQTL data for finding loci that explain genetic variation within the MAPK pathway, and identification of the gene network within this pathway.

The proposed method leverages and combines the beneficial properties of the horseshoe and graphical horseshoe priors, resulting in improved statistical performance. Computationally, the proposed sampler is the first in an SUR setting with a complexity linear in p, although the complexity is cubic in q. A major advantage of the proposed method is samples are available from the full posterior distribution, thereby allowing straightforward uncertainty quantification. If draws from the full posterior are not desired, it is possible faster algorithms can be developed to obtain point estimates. Prominent among these possibilities is an iterated conditional modes (ICM) algorithm (Besag, 1986) that can be used to obtain the maximum pseudo posterior estimate. At each iteration, ICM maximizes the full conditional posteriors of all variables until convergence and leads to a deterministic solution. Since the full conditionals in the HS-GHS model are either normal, gamma or inverse gamma, the modes are well defined, and ICM should be easy to implement. This article focused on the horseshoe prior, which is a member of a broader class of global-local priors, sharing a sharp peak at zero and heavy tails. Performance of other priors belonging to this family, such as the horseshoe+ (Bhadra et al., 2017), should also be explored.

Table 4.1.: Mean squared error (sd) in estimation and prediction, average Kullback-Leibler divergence, and sensitivity, specificity and precision of variable selection performance, over 50 simulated data sets, $p = 200$ and $q = 25$. The regression coefficients and precision matrix are estimated by HS-GHS, joint high-dimensional Bayesian variable and covariance selection (BM13), MRCE and CAPME. The best performer in each column is shown in bold.

	Simulation 1: $p = 200$, $q = 25$, $n = 100$, Uniform coefficients, AR1 structure										
	MSE			Divergence	B support recovery			Ω support recovery			CPU time
Method	B	Ω	Prediction	avg KL	SEN	SPE	PRC	SEN	SPE	PRC	min.
HS-GHS	**0.0033**	0.0365	**2.6352**	**10.2075**	.9380	.9981	**.9621**	**.9658**	.9973	**.9700**	788.75
	(0.0005)	(0.0123)	(0.1792)	(1.2853)	(.0155)	(.0006)	(.0122)	(.0383)	(.0039)	(.0418)	
BM13	0.0560	**0.0301**	8.4230	14.8512	-	-	-	.0200	**.9986**	.5588 [1]	54.80
	(0.0006)	(0.0005)	(0.4276)	(0.3441)	-	-	-	(.0242)	(.0019)	(.4567)	
MRCE	0.0854	0.0476	19.4201	29.9000	.0208	**.9996**	.8074	.9425	.0907	.0828	0.28
	(0.0007)	(0.0006)	(0.8754)	(0.3824)	(.0083)	(.0004)	(.1751)	(.0733)	(.0724)	(.0028)	
CAPME	0.0156	0.0417	4.0337	12.1094	**.9445**	.8187	.2167	0	1	- [2]	74.60
	(0.0014)	(0.0010)	(0.2749)	(0.4189)	(.0130)	(.0201)	(.0182)	(0)	(0)	-	
	Simulation 2: $p = 200$, $q = 25$, $n = 100$, Uniform coefficients, Cliques structure										
	MSE			Divergence	B support recovery			Ω support recovery			CPU time
Method	B	Ω	Prediction	avg KL	SEN	SPE	PRC	SEN	SPE	PRC	min.
HS-GHS	**0.0058**	**0.0371**	**3.5388**	**9.0762**	.8696	.9985	**.9693**	**.9700**	.9972	**.9687**	788.31
	(0.0010)	(0.0253)	(0.1791)	(1.3446)	(.0204)	(.0008)	(.0159)	(.0430)	(.0030)	(.0331)	
BM13	0.0570	0.0595	9.2452	14.3267	-	-	-	.0204	**.9993**	.7500 [3]	54.79
	(0.0006)	(0.0006)	(0.4789)	(0.4324)	-	-	-	(.0242)	(.0014)	(.3808)	
MRCE	0.0861	0.0756	20.1694	27.3668	.0116	**.9999**	.9370	.9507	.0788	.0825	0.16
	(0.0005)	(0.0006)	(0.9440)	(0.2892)	(.0057)	(.0001)	(.1121)	(.0581)	(.0596)	(.0041)	
CAPME	0.0188	0.0718	5.0170	11.2598	**.9266**	.8270	.2218	0	1	- [4]	73.67
	(0.0016)	(0.0007)	(0.2930)	(0.3797)	(.0155)	(.0215)	(.0198)	(0)	(0)	-	

1. 16 NaNs in 50 replicates. 3. 23 NaNs in 50 replicates. 2,4. 50 NaNs. All mean and sd. calculated on non-NaN values.

Table 4.2.: Mean squared error (sd) in estimation and prediction, average Kullback-Leibler divergence, and sensitivity, specificity and precision of variable selection performance, over 50 simulated data sets, $p = 120$ and $q = 50$. The regression coefficients and precision matrix are estimated by HS-GHS, joint high-dimensional Bayesian variable and covariance selection (BM13), MRCE and CAPME. The best performer in each column is shown in bold.

	Simulation 3: $p = 120$, $q = 50$, $n = 100$, Uniform coefficients, AR1 structure										
	MSE			Divergence	B support recovery			Ω support recovery			CPU time
Method	B	Ω	Prediction	avg KL	SEN	SPE	PRC	SEN	SPE	PRC	min.
HS-GHS	**0.0022**	**0.0041**	**2.4495**	**8.0596**	**.9709**	**.9984**	**.9696**	**.9873**	**.9995**	**.9875**	2.57e+03
	(0.0002)	(0.0009)	(0.1055)	(0.6494)	(.0087)	(.0007)	(.0120)	(.0136)	(.0007)	(.0156)	
BM13	0.0493	0.0132	5.1923	25.1810	-	-	-	.2804	.9976	.8295	217.24
	(0.0006)	(0.0006)	(0.2091)	(0.7590)	-	-	-	(.0603)	(.0015)	(.1058)	
MRCE	0.0689	0.0150	10.5162	40.3985	.2774	.9897	.5895	.9755	.1218	.0442	10.34
	(0.0022)	(0.0004)	(0.5920)	(0.8349)	(.0281)	(.0023)	(.0431)	(.0189)	(.0116)	(.0009)	
CAPME	0.0151	0.0105	3.2662	14.6163	.9462	.8887	.3122	.9514	.9795	.6705 [1]	80.69
	(0.0015)	(0.0013)	(0.1501)	(0.9668)	(.0131)	(.0184)	(.0280)	(.1390)	(.0093)	(.0782)	
	Simulation 4: $p = 120$, $q = 50$, $n = 100$, Uniform coefficients, Cliques structure										
	MSE			Divergence	B support recovery			Ω support recovery			CPU time
Method	B	Ω	Prediction	avg KL	SEN	SPE	PRC	SEN	SPE	PRC	min.
HS-GHS	**0.0032**	**0.0052**	**3.0221**	**7.8564**	.9409	**.9986**	**.9717**	**.9992**	.9990	**.9776**	2.57e+03
	(0.0004)	(0.0028)	(0.0983)	(0.8065)	(.0131)	(.0006)	(.0121)	(.0059)	(.0013)	(.0284)	
BM13	0.0506	0.0290	5.8167	24.0404	-	-	-	.0904	**.9993**	.8414	216.83
	(0.0007)	(0.0005)	(0.2225)	(0.6104)	-	-	-	(.0359)	(.0007)	(.1497)	
MRCE	0.0774	0.0298	12.0456	41.3306	.1527	.9971	.7398	.9679	.0940	.0419	8.06
	(0.0014)	(0.0010)	(0.6366)	(0.7870)	(.0192)	(.0009)	(.0625)	(.0684)	(.0780)	(.0020)	
CAPME	0.0161	0.0331	3.8324	16.9539	**.9537**	.8373	.2384	0	1	- [2]	81.99
	(0.0013)	(0.0004)	(0.1421)	(0.4293)	(.0122)	(.0234)	(.0251)	(0)	(0)	-	

1. 1 NaN in 50 replicates. 2. 50 NaNs. Mean and sd. calculated on non-NaN values.

Table 4.3.: Percentage of model explained variation in prediction of gene expressions. Model coefficients are estimated in training set ($n = 88$) and prediction performance is evaluated in testing set ($n = 22$).

Gene	CAPME	MRCE	HS-SUR	Gene	CAPME	MRCE	HS-SUR
FUS3	**15.46**	0.00	2.12	TEC1	23.08	0.00	**26.27**
FUS1	**31.78**	0.00	17.60	SSK22	21.24	0.00	**59.57**
STE2	43.78	0.00	**79.76**	MF(ALPHA)2	23.64	0.00	**48.27**
GPA1	**19.50**	0.00	1.38	FAR1	**30.66**	0.00	1.47
STE3	36.19	0.00	**76.45**	MF(ALPHA)1	39.37	0.00	**80.93**
BEM1	0.00	0.00	**16.68**	STE5	0.00	4.90	**19.60**
KSS1	2.80	0.00	**21.76**	SLN1	4.38	0.00	**10.41**
STE18	0.00	0.00	**24.88**	MLP1	0.00	0.00	**10.19**
HOG1	0.00	0.00	**19.28**	FKS1	0.00	0.00	**32.09**
MCM1	0.00	0.00	**29.96**	WSC3	0.00	0.00	**10.20**
SLG1	0.00	8.98	**10.27**	RHO1	0.00	0.00	**10.57**

Table 4.4.: Nonzero coefficients in HS-GHS estimate, along with name and location of the gene, location of the marker, and CAPME estimated coefficients.

Gene	Chromosome	Within-chr.	Marker chr.	Within-chr.	HS-GHS	CAPME
		position		marker position	coefficients	coefficients
FUS3	2	192454-193515	2	424330	0.32	0.06
BEM1	2	620867-622522	8	71742	-0.35	0.00
FUS1	3	71803-73341	4	17718	0.13	0.00
FUS1	3	71803-73341	4	527445	-0.42	-0.13
SWI4	5	382591-385872	13	361370	-0.88	0.00
SWI4	5	382591-385872	5	458085	-0.69	0.00
SWI4	5	382591-385872	3	201166	3.65	2.00
SHO1	5	397948-399051	3	201166	-1.89	-0.91
BCK1	10	247250-251686	3	201166	-4.11	-2.66
MID2	12	790676-791806	13	314816	0.29	0.06
STE11	12	849865-852018	5	109310	0.13	0.00
MFA2	14	352416-352532	14	449639	0.13	0.00
SSK2	14	680696-685435	5	395442	0.98	0.00
SSK2	14	680696-685435	13	403766	0.68	0.08
SSK2	14	680696-685435	3	201166	-3.60	-2.05

5. CONCLUSION

The horseshoe prior is a member of the global-local shrinkage methods for sparsity. Its properties under the normal means problem have been widely studied, and it has been applied to many other settings including the generalized linear model, classification, survival analysis, and low-dimensional functions of normal means. This dissertation aims at studying the prediction risk of horseshoe regression, and some novel applications of the horseshoe prior to multivariate settings.

Chapter 2 studied quadratic prediction risk of horseshoe regression, and compared Stein's unbiased risk estimator under the horseshoe prior and under some other popular global shrinkage rules. Chapter 2 shows that the local shrinkage parameters in global-local models make the procedure highly adaptive in sparse regressions. The horseshoe regression strikes a balance between variance reduction and biasedness caused by shrinkage, and often achieves lower prediction risk than global shrinkage rules.

Chapters 3 and 4 extend the horseshoe prior to precision matrix estimation and joint estimation of coefficients and precision matrix in Gaussian models. Some properties of the horseshoe prior in the normal means problem, *i.e.* reduced Kullback-Leibler divergence and tail-robustness, can be transferred to properties in precision estimation. Advancements in computational methods also make efficient Gibbs samplers possible, even in cases where number of predictors and/or number of features in the responses exceed sample size. The Gibbs algorithms sample from full Bayes conditional distributions, and enable uncertainty quantification for parameter estimates and variable selection based on posterior intervals.

Chapter 3 studied an unbiased estimator of prediction risk in a non-asymptotic setting. However, asymptotic properties of the horseshoe regression prediction risk remain a subject for future investigation. Chapters 3 and 4 showed a few properties of the horseshoe prior transfered from the normal means problem. However, other properties, such as consistency of variable selection, need to be studied in precision matrix estimation and joint coefficient and precision estimation. Another possible direction for future development is whether the methods in Chapters 3 and 4 are robust to non-Gaussian data. Many data sets in applications may not be Gaussian distributed, and the horseshoe prior can be successful in a lot more applications if it allows some diversion from the normal assumption.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Armagan, A., Clyde, M., and Dunson, D. B. (2011). Generalized beta mixtures of Gaussians. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P., Pereira, F. C. N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 523–531.

Armagan, A., Dunson, D. B., and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica*, 23(1):119–143.

Banerjee, S. and Ghosal, S. (2014). Posterior convergence rates for estimating large precision matrices using graphical models. *Electronic Journal of Statistics*, 8(2):2111–2137.

Banerjee, S. and Ghosal, S. (2015). Bayesian structure learning in graphical models. *Journal of Multivariate Analysis*, 136:147–162.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897.

Barron, A. R. (1988). *The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions.* Technical report, Department of Statistics, University of Illinois, Champaign, IL.

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 259–302.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016). Default bayesian analysis with global-local shrinkage priors. *Biometrika*, 103(4):955–969.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12(4):1105–1131.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. T. (2019). Lasso meets horseshoe: a survey. *Statistical Science*. to appear.

Bhadra, A. and Mallick, B. K. (2013). Joint high-dimensional Bayesian variable and covariance selection with an application to eQTL analysis. *Biometrics*, 69(2):447–457.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4):985–991.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490.

Brem, R. B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences*, 102(5):1572–1577.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2012). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, 100(1):139–156.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351.

Carlin, B. P. and Polson, N. G. (1991). Inference for nonconjugate bayesian models using the gibbs sampler. *Canadian Journal of statistics*, 19(4):399–405.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *International Conference on Artificial Intelligence and Statistics*, pages 73–80.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Chun, H. and Keles, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25.

Clyde, M., Desimone, H., and Parmigiani, G. (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association*, 91(435):1197–1208.

Conklin, B., Adriaens, M., Kelder, T., and Salomonis, N. (2018). Mapk signaling pathway (saccharomyces cerevisiae). `https://www.wikipathways.org/index.php/Pathway:WP510`. [Online; accessed 12-December-2018].

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403.

Cressie, N. (1993). *Statistics for Spatial Data.* John Wiley & Sons, Hoboken, NJ.

Curtis, R. E., Kim, S., Woolford Jr, J. L., Xu, W., and Xing, E. P. (2013). Structured association analysis leads to insight into saccharomyces cerevisiae gene regulation by finding multiple contributing eqtl hotspots associated with functional gene modules. *BMC genomics*, 14(1):196.

Datta, J. and Ghosh, J. K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Analysis*, 8(1):111–132.

Dawid, A. P. (1981). Some matrix-variate distribution theory: notational considerations and a bayesian application. *Biometrika*, 68(1):265–274.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317.

Dehmer, M. and Emmert-Streib, F. (2008). *Analysis of Microarray Data: A Network-Based Approach.* Wiley-VCH, Weinheim, Chichester.

Denison, D. G. T. and George, E. I. (2012). *Bayesian prediction with adaptive ridge estimators*, volume 8 of *IMS Collections*, pages 215–234. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Deshpande, S. K., Rockova, V., and George, E. I. (2017). Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *arXiv preprint arXiv:1708.08911*.

Dicker, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika*, 101(2):269–284.

Diggle, P. (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

Efron, B. (1975). Biased versus unbiased estimation. *Advances in Mathematics*, 16(3):259–277.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation (with discussion). *Journal of the American Statistical Association*, 99(467):619–642.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive lasso and SCAD penalties. *The Annals of Applied Statistics*, 3(2):521–541.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J., Liao, Y., and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22:1947–1975.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Applications of the lasso and grouped lasso to the estimation of sparse graphical models. Technical report, Stanford Univ.

Friedman, J., Hastie, T., and Tibshirani, R. (2018). *glasso: Graphical Lasso: Estimation of Gaussian Graphical Models*. R package version 1.10.

George, E. I., Liang, F., and Xu, X. (2006). Improved minimax predictive densities under Kullback-Leibler loss. *The Annals of Statistics*, 34(1):78–91.

Ghosh, P., Tang, X., Ghosh, M., Chakrabarti, A., et al. (2016). Asymptotic properties of bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. *Bayesian Analysis*, 11(3):753–796.

Gordy, M. B. (1998). A generalization of generalized beta distributions. In *Finance and Economics Discussion Series*. Division of Research and Statistics, Division of Monetary Affairs, Federal Reserve Board.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference and prediction*. Springer, New York, 2nd edition.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Holmes, C. C., Denison, D. T., and Mallick, B. K. (2002). Accounting for model uncertainty in seemingly unrelated regressions. *Journal of Computational and Graphical Statistics*, 11(3):533–551.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379, Berkeley, Calif. University of California Press.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(3):300–303.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic acids research*, 38(suppl_1):D355–D360.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics*, 37(6B):4254–4278.

Leclerc, R. D. (2008). Survival of the sparsest: robust gene networks are parsimonious. *Molecular Systems Biology*, 4(1):213.

Lee, K. and Lee, J. (2017a). Estimating large precision matrices via modified Cholesky decomposition. *arXiv preprint arXiv:1707.01143*.

Lee, K. and Lee, J. (2017b). Optimal Bayesian minimax rates for unconstrained large covariance matrices. *arXiv preprint arXiv:1702.07448*.

Magnusson, M., Jonsson, L., and Villani, M. (2016). Dolda-a regularized supervised topic model for high-dimensional multi-class regression. *arXiv preprint arXiv:1602.00260*.

Makalic, E. and Schmidt, D. F. (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15(4):661–675.

Martin, R. and Walker, S. G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electronic Journal of Statistics*, 8(2):2188–2206.

Masreliez, C. (1975). Approximate non-Gaussian filtering with linear state and observation relations. *IEEE Transactions on Automatic Control*, 20(1):107–110.

MATLAB (2018). *9.4.0.813654 (R2018a)*. The MathWorks Inc., Natick, Massachusetts.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Meinshausen, N., Rocha, G., and Yu, B. (2007). Discussion: A tale of three cousins: Lasso, L2Boosting and Dantzig. *The Annals of Statistics*, 35(6):2373–2384.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pati, D., Bhattacharya, A., Pillai, N. S., and Dunson, D. (2014). Posterior contraction in sparse bayesian factor models for massive covariance matrices. *The Annals of Statistics*, 42(3):1102–1130.

Peltola, T., Havulinna, A. S., Salomaa, V., and Vehtari, A. (2014). Hierarchical bayesian survival analysis and projective covariate selection in cardiovascular event risk prediction. In *BMA@ UAI*, pages 79–88. Citeseer.

Pericchi, L. R. and Smith, A. F. M. (1992). Exact and approximate posterior moments for a normal location parameter. *Journal of the Royal Statistical Society. Series B*, 54:793–804.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051.

Pineda-Pardo, J. A., Bruña, R., Woolrich, M., Marcos, A., Nobre, A. C., Maestú, F., and Vidaurre, D. (2014). Guiding functional connectivity estimation by structural connectivity in meg: an application to discrimination of conditions of mild cognitive impairment. *Neuroimage*, 101:765–777.

Polson, N. G. and Scott, J. G. (2012). Local shrinkage rules, Lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B*, 74(2):287–311.

Pourahmadi, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statistical Science*, 26(3):369–387.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rajaratnam, B., Massam, H., and Carvalho, C. M. (2008). Flexible covariance estimation in graphical gaussian models. *The Annals of Statistics*, 36(6):2818–2849.

Rissanen, J. (1986). Stochastic complexity and modeling. *The Annals of Statistics*, 14:1080–1100.

Ročková, V. and George, E. I. (2016). The spike-and-slab lasso. *Journal of the American Statistical Association*, to appear.

Rodríguez-Peña, J. M., García, R., Nombela, C., and Arroyo, J. (2010). The high-osmolarity glycerol (hog) and cell wall integrity (cwi) signalling pathways interplay: a yeast dialogue between mapk routes. *Yeast*, 27(8):495–502.

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962.

Ročková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G., Milligan, S. B., Lamb, J. R., Cavet, G., et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. *Nature*, 422(6929):297–302.

Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7):2144–2162.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Szakács, G., Annereau, J.-P., Lababidi, S., Shankavaram, U., Arciello, A., Bussey, K. J., Reinhold, W., Guo, Y., Kruh, G. D., Reimers, M., Weinstein, J. N., and Gottesman, M. M. (2004). Predicting drug sensitivity and resistance: Profiling ABC transporter genes in cancer cells. *Cancer Cell*, 6(2):129 – 137.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58:267–288.

Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232.

van der Pas, S., Kleijn, B., and van der Vaart, A. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2):2585–2618.

van der Pas, S., Szabó, B., and van der Vaart, A. (2017a). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis*, 12(4):1221–1274.

van der Pas, S., Szabó, B., van der Vaart, A., et al. (2017b). Adaptive posterior contraction rates for the horseshoe. *Electronic Journal of Statistics*, 11(2):3196–3225.

Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886.

Wei, R. (2017). Bayesian variable selection using continuous shrinkage priors for nonparametric models and non-gaussian data.

Xiang, R., Khare, K., and Ghosh, M. (2015). High dimensional posterior convergence rates for decomposable graphical models. *Electronic Journal of Statistics*, 9(2):2828–2854.

Yin, J. and Li, H. (2011). A sparse conditional gaussian graphical model for analysis of genetical genomics data. *The annals of applied statistics*, 5(4):2630.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.

Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American statistical Association*, 57(298):348–368.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In Goel, P. K. and Zellner, A., editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. Elsevier, North-Holland.

Zheng, S. and Liu, W. (2011). An experimental comparison of gene selection by lasso and dantzig selector for cancer classification. *Computers in Biology and Medicine*, 41(11):1033–1040.

Zhu, J., Zhang, B., Smith, E. N., Drees, B., Brem, R. B., Kruglyak, L., Bumgarner, R. E., and Schadt, E. E. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature genetics*, 40(7):854.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

APPENDICES

A. SUPPLEMENTARY MATERIAL TO CHAPTER 2

A.1 Proof of Theorem 2.3.1

Part A follows from Equation (2.6) with standard algebraic manipulations. To prove part B, define $Z_i = 1/(1+\tau^2\lambda_i^2 d_i^2)$. Then, from Equation (2.12)

$$\begin{aligned} m(\hat{\alpha}) =& (2\pi\sigma^2)^{-n/2}\prod_{i=1}^n \int_0^1 \exp(-z_i\hat{\alpha}_i^2 d_i^2/2\sigma^2) d_i z_i^{1/2}\left(\frac{z_i\tau^2 d_i^2}{1-z_i+z_i\tau^2 d_i^2}\right)\frac{1}{\tau d_i}(1-z_i)^{-1/2} z_i^{-3/2} dz_i \\ =& (2\pi\sigma^2)^{-n/2}\prod_{i=1}^n \int_0^1 \exp(-z_i\hat{\alpha}_i^2 d_i^2/2\sigma^2)(1-z_i)^{-1/2}\left\{\frac{1}{\tau^2 d_i^2}+\left(1-\frac{1}{\tau^2 d_i^2}\right)z_i\right\}^{-1} dz_i. \end{aligned}$$

From the definition of the compound confluent hypergeometric (CCH) density in Gordy (1998), the result of the integral is proportional to the normalizing constant of the CCH density and we have from Proposition 2.3.1 that,

$$m(\hat{\alpha}) \propto (2\pi\sigma^2)^{-n/2}\prod_{i=1}^n H\left(1, \frac{1}{2}, 1, \frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2}, 1, \frac{1}{\tau^2 d_i^2}\right).$$

In addition, the random variable $(Z_i \mid \hat{\alpha}_i, \sigma, \tau)$ follows a $\mathrm{CCH}(1, 1/2, 1, \hat{\alpha}_i^2 d_i^2/2\sigma^2, 1, 1/\tau^2 d_i^2)$ distribution. Lemma 3 of Gordy (1998) gives,

$$\frac{d}{ds}H(p,q,r,s,\nu,\theta) = -\frac{p}{p+q}H(p+1,q,r,s,\nu,\theta).$$

This yields after some algebra that,

$$\begin{aligned}
\frac{m'(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} =& -\frac{2}{3}\frac{H\left(2, \frac{1}{2}, 1, \frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2}, 1, \frac{1}{\tau^2 d_i^2}\right)}{H\left(1, \frac{1}{2}, 1, \frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2}, 1, \frac{1}{\tau^2 d_i^2}\right)}\frac{\hat{\alpha}_i d_i^2}{\sigma^2}, \\
\frac{m''(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} =& \frac{-\frac{2}{3}H\left(2, \frac{1}{2}, 1, \frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2}, 1, \frac{1}{\tau^2 d_i^2}\right)\frac{d_i^2}{\sigma^2} + \frac{8}{15}H\left(3, \frac{1}{2}, 1, \frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2}, 1, \frac{1}{\tau^2 d_i^2}\right)\frac{\hat{\alpha}_i^2 d_i^4}{\sigma^4}}{H\left(1, \frac{1}{2}, 1, \frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2}, 1, \frac{1}{\tau^2 d_i^2}\right)}.
\end{aligned}$$

The correctness of the assertion

$$\frac{m'(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} = -\frac{\hat{\alpha}_i d_i^2}{\sigma^2}\mathrm{E}(Z_i), \quad \text{and} \quad \frac{m''(\hat{\alpha}_i)}{m(\hat{\alpha}_i)} = -\frac{d_i^2}{\sigma^2}\mathrm{E}(Z_i) + \frac{\hat{\alpha}_i^2 d_i^4}{\sigma^4}\mathrm{E}(Z_i^2),$$

can then be verified using Equation (2.15), completing the proof.

A.2 Proof of Theorem 2.3.2

Define $s_i = \hat{\alpha}_i^2 d_i^2 / 2\sigma^2$ and $\theta_i = (\tau^2 d_i^2)^{-1}$. From Theorem 2.3.1 the component-wise prediction risk is

$$\begin{aligned}
R_i =& 2\sigma^2 - 2\sigma^2\mathrm{E}(Z_i) - \hat{\alpha}_i^2 d_i^2\{\mathrm{E}(Z_i)\}^2 + 2\hat{\alpha}_i{}^2 d_i^2\mathrm{E}(Z_i^2) \\
=& 2\sigma^2[1 - \mathrm{E}(Z_i) + 2s_i\mathrm{E}(Z_i^2) - s_i\{\mathrm{E}(Z_i)\}^2], \qquad \text{(A.1)}
\end{aligned}$$

where $Z_i = 1/(1+\tau^2\lambda_i^2 d_i^2)$. Recall that $(Z_i \mid \hat{\alpha}_i, \sigma, \tau)$ follows a $\mathrm{CCH}(1, 1/2, 1, \hat{\alpha}_i^2 d_i^2/2\sigma^2, 1, 1/\tau^2 d_i^2)$ distribution. Thus, its first and second moments are

$$\begin{aligned}
\mathrm{E}(Z_i) =& \frac{\int_0^1 z_i(1-z_i)^{-\frac{1}{2}}\{\theta_i + (1-\theta_i)z_i\}^{-1}\exp(-s_i z_i)dz_i}{\int_0^1 (1-z_i)^{-\frac{1}{2}}\{\theta_i + (1-\theta_i)z_i\}^{-1}\exp(-s_i z_i)dz_i}, \\
\mathrm{E}(Z_i^2) =& \frac{\int_0^1 z_i^2(1-z_i)^{-\frac{1}{2}}\{\theta_i + (1-\theta_i)z_i\}^{-1}\exp(-s_i z_i)dz_i}{\int_0^1 (1-z_i)^{-\frac{1}{2}}\{\theta_i + (1-\theta_i)z_i\}^{-1}\exp(-s_i z_i)dz_i},
\end{aligned}$$

where the conditioning variables under the expectation notations have been omitted throughout for the sake of brevity. We assume $\max(d_i) < \infty$ and $0 < \theta_i < \infty$. First consider $\theta_i \in (0, 1)$. Then, clearly, $1 < \{\theta_i + (1-\theta_i)z_i\}^{-1} < \theta_i^{-1}$ when $0 \leq z_i \leq 1$.

Define $a = \log(s_i^2)/s_i$. It is easy to verify that $a \in (0,1)$ when $s_i \geq 1$. Now, the numerator of $\mathrm{E}(Z_i)$ is

$$\begin{aligned}
&\int_0^1 z_i(1-z_i)^{-\frac{1}{2}}\{\theta_i + (1-\theta_i)z_i\}^{-1}\exp(-s_i z_i)dz_i \\
<&\theta_i^{-1}\int_0^1 z_i(1-z_i)^{-\frac{1}{2}}\exp(-s_i z_i)dz_i \\
<&\theta_i^{-1}\int_0^a z_i(1-a)^{-\frac{1}{2}}\exp(-s_i z_i)dz_i + \theta_i^{-1}\int_a^1 z_i(1-z_i)^{-\frac{1}{2}}\exp(-as_i)dz_i \qquad (\text{for } 0 < a < 1) \\
=&\theta_i^{-1}(1-a)^{-\frac{1}{2}}\left\{-\frac{a}{s_i}\exp(-as_i) - \frac{1}{s_i^2}\exp(-as_i) + \frac{1}{s_i^2}\right\} + \theta_i^{-1}\exp(-as_i)\int_a^1 z_i(1-z_i)^{-\frac{1}{2}}dz_i \\
=&\theta_i^{-1}\left(1-\frac{1}{s_i}\log s_i^2\right)^{-\frac{1}{2}}\left(-\frac{1}{s_i^4}\log s_i^2 - \frac{1}{s_i^4} + \frac{1}{s_i^2}\right) + \frac{1}{\theta_i s_i^2}\int_a^1 z_i(1-z_i)^{-\frac{1}{2}}dz_i,
\end{aligned}$$

which is $O(1/s_i^2)$. Similarly, divide the integration between $(0,a)$ and $(a,1)$, and take $a = \log(s_i^3)/s_i$. It is easy to verify $a \in (0,1)$ for $s_i \geq 5$. Then, an upper bound of the numerator of $\mathrm{E}(Z_i^2)$ is found to be $O(1/s_i^3)$ by similar calculations. The denominator of $\mathrm{E}(Z_i)$ and $\mathrm{E}(Z_i^2)$ is

$$\begin{aligned}
&\int_0^1 (1-z_i)^{-\frac{1}{2}}\{\theta_i + (1-\theta_i)z_i\}^{-1}\exp(-s_i z_i)dz_i \\
>&\int_0^1 (1-z_i)^{-\frac{1}{2}}\exp(-s_i z_i)dz_i \\
>&\int_0^1 \exp(-s_i z_i)dz_i \\
=&\frac{1}{s_i} - \frac{1}{s_i}\exp(-s_i).
\end{aligned}$$

Combining the upper bounds on the numerators and the lower bounds on the denominators show $\mathrm{E}(Z_i)$ to be $O(1/s_i)$ and $\mathrm{E}(Z_i^2)$ to be $O(1/s_i^2)$ for large s_i. Using this in Equation (A.1) completes the proof when $\theta_i \in (0,1)$. The result in the case where $\theta_i \in [1,\infty)$ follows from similar calculations and we omit the proof.

A.3 Proof of Theorem 2.3.3

The proof of Theorem 2.3.3 makes use of Propositions A.3.1-A.3.3, stated below, with proofs given in Appendix A.4.

Proposition A.3.1 *If* $Z \sim \mathrm{CCH}(p,q,r,s,\nu,\theta)$*, then* $(\partial/\partial s)\mathrm{E}(Z^k) = \mathrm{E}(Z)\mathrm{E}(Z^k) - \mathrm{E}(Z^{k+1})$.

Proposition A.3.2 *If* $Z \sim \mathrm{CCH}(p,q,r,s,\nu,\theta)$*, then* $(\partial^2/\partial^2 s)\mathrm{E}(Z) = \mathrm{E}\{(Z-\mu)^3\}$*, where* $\mu = \mathrm{E}(Z)$.

Proposition A.3.3 *If* $Z \sim \mathrm{CCH}(p,q,r,s,\nu,\theta)$*, then* $(\partial/\partial\theta)\mathrm{E}(Z) = -\mathrm{Cov}(Z,W)$*, for* $W = (1-\nu Z)\{\theta + (1-\theta)\nu Z\}^{-1}$*. If* $0 < \theta \le 1$ *then* $(\partial/\partial\theta)\mathrm{E}(Z) > 0$.

Recall from Appendix A.1 that if we define $Z_i = 1/(1+\tau^2\lambda_i^2 d_i^2)$ then the density of Z_i is given by

$$(Z_i \mid \hat{\alpha}_i, d_i, \tau, \sigma^2) \sim \mathrm{CCH}\left(Z_i \mid 1, \frac{1}{2}, 1, \frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2}, 1, \frac{1}{\tau^2 d_i^2}\right). \tag{A.2}$$

The risk is $R = \sum_{i=1}^n R_i$ with

$$\begin{aligned} R_i =& 2\sigma^2[1 - \mathrm{E}(Z_i) + 2s_i\mathrm{E}(Z_i^2) - s_i\{\mathrm{E}(Z_i)\}^2] \\ =& 2\sigma^2[1 - \mathrm{E}(Z_i) + s_i\mathrm{E}(Z_i^2) + s_i\mathrm{Var}(Z_i)], \end{aligned} \tag{A.3}$$

where $s_i = \hat{\alpha}_i^2 d_i^2/2\sigma^2$. Thus,

$$\begin{aligned} \frac{\partial R_i}{\partial s_i} =& -2\sigma^2\frac{\partial \mathrm{E}(Z_i)}{\partial s_i} + 2\sigma^2\frac{\partial}{\partial s_i}\{s_i\mathrm{E}(Z_i^2)\} + 2\sigma^2\frac{\partial}{\partial s_i}\{s_i\mathrm{Var}(Z_i)\} \\ :=& \mathrm{I} + \mathrm{II} + \mathrm{III}. \end{aligned} \tag{A.4}$$

Now, as a corollary to Lemma A.3.1, $(\partial/\partial s_i)\mathrm{E}(Z_i) = \{\mathrm{E}(Z_i)\}^2 - \mathrm{E}(Z_i^2) = -\mathrm{Var}(Z_i) < 0$, giving $\mathrm{I} > 0$. The strict inequality follows from the fact that Z_i is not almost surely a constant for any $s_i \in \mathbb{R}$ and $(\partial/\partial s_i)\mathrm{E}(Z_i)$ is continuous at $s_i = 0$. Next, consider II. Define $\theta_i = (\tau^2 d_i^2)^{-1}$ and let $0 \leq s_i \leq 1$. Then,

$$\begin{aligned}\frac{\partial}{\partial s_i}\{s_i \mathrm{E}(Z_i^2)\} =& \mathrm{E}(Z_i^2) + s_i \frac{\partial}{\partial s_i}\mathrm{E}(Z_i^2) \\ =& \mathrm{E}(Z_i^2) + s_i\{\mathrm{E}(Z_i)\mathrm{E}(Z_i^2) - \mathrm{E}(Z_i^3)\} \qquad \text{(by Lemma A.3.1)} \\ =& s_i \mathrm{E}(Z_i)\mathrm{E}(Z_i^2) + \{\mathrm{E}(Z_i^2) - s_i \mathrm{E}(Z_i^3)\}.\end{aligned}$$

Now, clearly, the first term, $s_i \mathrm{E}(Z_i)\mathrm{E}(Z_i^2) \geq 0$. We also have $Z_i^2 - s_i Z_i^3 = Z_i^2(1 - s_i Z_i) \geq 0$ a.s. when $0 \leq Z_i \leq 1$ a.s. and $0 \leq s_i \leq 1$. Thus, the second term $\mathrm{E}(Z_i^2) - s_i \mathrm{E}(Z_i^3) \geq 0$. Putting the terms together gives II ≥ 0. Finally, consider III. Denote $\mathrm{E}(Z_i) = \mu_i$. Then,

$$\begin{aligned}\frac{\partial}{\partial s_i}\{s_i \mathrm{Var}(Z_i)\} =& \mathrm{Var}(Z_i) + \frac{\partial}{\partial s_i}\{\mathrm{Var}(Z_i)\} \\ =& \mathrm{Var}(Z_i) - s_i \frac{\partial^2 \mathrm{E}(Z_i)}{\partial s_i^2} \\ =& \mathrm{E}\{(Z_i - \mu_i)^2\} - s_i \mathrm{E}\{(Z_i - \mu_i)^3\} \qquad \text{(by Lemma A.3.2)} \\ =& \mathrm{E}[(Z_i - \mu_i)^2\{1 - s_i(Z_i - \mu_i)\}].\end{aligned}$$

Now, $(Z_i - \mu_i)^2\{1 - s_i(Z_i - \mu_i)\} \geq 0$ a.s. when $0 \leq Z_i \leq 1$ a.s. and $0 \leq s_i \leq 1$ and thus, III ≥ 0. Using I, II and III in Equation (A.4) yields R_i is an increasing function of s_i when $0 \leq s_i \leq 1$, completing the proof of Part A.

To prove Part B, we need to derive an upper bound on the risk when $s_i = 0$. First, consider $s_i = 0$ and $0 < \theta_i \leq 1$. we have from Equation (A.3) that $R_i = 2\sigma^2(1 - \mathrm{E}Z_i)$. By Lemma A.3.3, $(\partial/\partial\theta_i)\mathrm{E}(Z_i) > 0$ and R_i is a monotone decreasing function of θ_i, where $\theta_i = (\tau^2 d_i^2)^{-1}$. Next consider the case where $s_i = 0$ and $\theta_i \in (1, \infty)$. Define $\tilde{Z}_i = 1 - Z_i \in (0, 1)$ when $Z_i \in (0, 1)$. Then, by Equation (A.2) and a formula on Page 9 of Gordy (1998), we have that $\tilde{Z}_i$ also follows a CCH distribution. Specifically,

$$(\tilde{Z}_i \mid \hat{\alpha}_i, d_i, \tau, \sigma^2) \sim \mathrm{CCH}\left(\tilde{Z}_i \mid \frac{1}{2}, 1, 1, -\frac{\hat{\alpha}_i^2 d_i^2}{2\sigma^2}, 1, \tau^2 d_i^2\right),$$

and we have $R_i = 2\sigma^2 \mathrm{E}(\tilde{Z}_i)$. Define $\tilde{\theta}_i = \theta_i^{-1} = \tau^2 d_i^2$. Then by Lemma A.3.3, $(\partial/\partial\tilde{\theta}_i)\mathrm{E}(\tilde{Z}_i) = -\mathrm{Cov}(\tilde{Z}_i, \tilde{W}_i) > 0$ on $0 < \tilde{\theta}_i < 1$. Therefore, R_i is a monotone increasing function of $\tilde{\theta}_i$ on $0 < \tilde{\theta}_i < 1$, or equivalently a monotone decreasing function of θ_i on $\theta_i \in (1, \infty)$.

Thus, combining the two cases above, we get that the risk at $s_i = 0$ is a monotone decreasing function of θ_i for any $\theta_i \in (0, \infty)$, or equivalently, an increasing function of $\tau^2 d_i^2$. Since $0 \leq \tilde{Z}_i \leq 1$ almost surely, a natural upper bound on R_i is $2\sigma^2$. However, it is possible to do better provided τ is chosen sufficiently small. Assume that $\tau^2 \leq d_i^{-2}$. Then, since R_i is monotone increasing in θ_i, the upper bound of the risk is achieved when $\theta_i = (\tau^2 d_i^2)^{-1} = 1$. In this case, $\mathrm{E}(Z_i)$ has a particularly simple expression, given by

$$\begin{aligned}\mathrm{E}(Z_i) =& \frac{\int_0^1 z_i(1-z_i)^{-\frac{1}{2}}\{\theta_i + (1-\theta_i)z_i\}^{-1} dz_i}{\int_0^1 (1-z_i)^{-\frac{1}{2}}\{\theta_i + (1-\theta_i)z_i\}^{-1} dz_i} \\ =& \frac{\int_0^1 z_i(1-z_i)^{-\frac{1}{2}} dz_i}{\int_0^1 (1-z_i)^{-\frac{1}{2}} dz_i} = \frac{2}{3}.\end{aligned}$$

Thus, $\sup R_i = 2\sigma^2(1 - \mathrm{E}Z_i) = 2\sigma^2/3$ when $\tau^2 \leq d_i^{-2}$, completing the proof.

A.4 Proofs of propositions

A.4.1 Proof of Proposition A.3.1

Let, $Z \sim \mathrm{CCH}(p, q, r, s, \nu, \theta)$. Then for any integer k

$$\mathrm{E}(Z^k) = \frac{\int_0^{1/\nu} z^{k+p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r} \exp(-sz) dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta + (1-\theta)\nu z\}^{-r} \exp(-sz) dz}.$$

Thus,

$$\begin{aligned}\frac{\partial}{\partial s}\mathrm{E}(Z^k) =& \frac{\int_0^{1/\nu} -z^{k+p}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz} \\ &- \left[\frac{\int_0^{1/\nu} z^{k+p-1}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right. \\ &\quad \left.\times \frac{\int_0^{1/\nu} -z^{p}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right] \\ =& -\mathrm{E}(Z^{k+1}) + \mathrm{E}(Z)\mathrm{E}(Z^k).\end{aligned}$$

For an alternative proof directly using the $H(\cdot)$ functions, see Appendix D of Gordy (1998).

A.4.2 Proof of Proposition A.3.2

Let, $Z \sim \mathrm{CCH}(p,q,r,s,\nu,\theta)$. From Lemma A.3.1, $(\partial/\partial s)\mathrm{E}(Z) = -\mathrm{E}(Z^2) + \{\mathrm{E}(Z)\}^2 = -\mathrm{Var}(Z)$. Let $\mu = \mathrm{E}(Z)$. Then,

$$
\begin{aligned}
\frac{\partial^2}{\partial s^2}\mathrm{E}(Z) =& -\frac{\partial}{\partial s}\mathrm{Var}(Z)\\
=& -\frac{\partial}{\partial s}\left[\frac{\int_0^{1/\nu}(z-\mu)^2 z^{p-1}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right]\\
=&\frac{\int_0^{1/\nu}(z-\mu)^2 z^{p}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}\\
&-\left[\frac{\int_0^{1/\nu}(z-\mu)^2 z^{p-1}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right.\\
&\quad\left.\times\frac{\int_0^{1/\nu} z^{p}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right]\\
=&\mathrm{Cov}(Z,(Z-\mu)^2)\\
=&\mathrm{E}[(Z-\mu)\{(Z-\mu)^2-\mathrm{E}(Z-\mu)^2\}]\\
=&\mathrm{E}\{(Z-\mu)^3\}-\mathrm{Var}(Z)\mathrm{E}(Z-\mu)=\mathrm{E}\{(Z-\mu)^3\}.
\end{aligned}
$$

A.4.3 Proof of Proposition A.3.3

Let $Z \sim \mathrm{CCH}(p,q,r,s,\nu,\theta)$ and $W=(1-\nu Z)\{\theta+(1-\theta)\nu Z\}^{-1}$. Then,

$$
\begin{aligned}
\frac{\partial}{\partial\theta}\mathrm{E}(Z) =& -\frac{\int_0^{1/\nu} z^{p}(1-\nu z)^{q}\{\theta+(1-\theta)\nu z\}^{-(r+1)}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}\\
&+\left[\frac{\int_0^{1/\nu} z^{p}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right.\\
&\quad\left.\times\frac{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q}\{\theta+(1-\theta)\nu z\}^{-(r+1)}\exp(-sz)dz}{\int_0^{1/\nu} z^{p-1}(1-\nu z)^{q-1}\{\theta+(1-\theta)\nu z\}^{-r}\exp(-sz)dz}\right]\\
=& -\mathrm{E}(ZW)+\mathrm{E}(Z)\mathrm{E}(W)=-\mathrm{Cov}(Z,W).
\end{aligned}
$$

When $0<\theta\le 1$, it is obvious that Z and W are negatively correlated, and thus $-\mathrm{Cov}(Z,W)>0$.

B. SUPPLEMENTARY MATERIAL TO CHAPTER 3

B.1 Proof of Theorem 3.3.1

We claim that an Euclidean cube of p^2 dimensions with $(\omega_{ij0}, \omega_{ij0} + \sqrt{\epsilon}/2Mp)$ on each dimension lies inside A_ϵ and an Euclidean cube with p^2 dimensions with $(\omega_{ij0} - 2\sqrt{\epsilon}/Mp, \omega_{ij0} + 2\sqrt{\epsilon}/Mp)$ on each dimension contains A_ϵ. The proof is as following.

$D(p_{\Omega_0}||p_\Omega) = \frac{1}{2}\{\log|\Omega^{-1}\Omega_0| + \mathrm{tr}(\Omega\Omega_0^{-1}) - p\}$. Take $\Omega = \Omega_0 + (\delta/Mp)\nVdash$ where $\nVdash$ is a matrix with all elements equal to 1, then

$$\begin{aligned}
\mathrm{tr}(\Omega_0^{-1}\Omega) - p =& \mathrm{tr}(\Omega_0^{-1}(\Omega_0 + (\delta/Mp)\nVdash)) - \mathrm{tr}(\Omega_0^{-1}\Omega_0) \\
=& \mathrm{tr}(\Omega_0^{-1}(\delta/Mp)\nVdash) \\
=& \sum_{i,j} \sigma_{ij0} * (\delta/Mp) = \delta, \\
\log|\Omega^{-1}\Omega_0| =& \log|(\Omega_0 + (\delta/Mp)\nVdash)^{-1}\Omega_0| \\
=& -\log|\Omega_0^{-1}(\Omega_0 + (\delta/Mp)\nVdash)| \\
=& -\log|I + \Omega_0^{-1}(\delta/Mp)\nVdash)| \\
=& -\log\prod_{i=1}^{p}(1 + \lambda_i) \\
=& -\sum_{i=1}^{p}\log(1 + \lambda_i),
\end{aligned}$$

where λ_i are the eigenvalues of the matrix $\Omega_0^{-1}(\delta/Mp)\nVdash$. The matrix $\Omega_0^{-1}(\delta/Mp)\nVdash$ has column rank 1, and its only non-zero eigenvalue is equal to $\sum_{i,j}\sigma_{ij0} * (\delta/Mp) = \delta$. Therefore $D(p_{\Omega_0}||p_\Omega) = \delta - \log(1 + \delta)$. The function $x - \log(1 + x)$ has expansion $x^2/2 + O(x^3)$ at $x = 0$. Take $\delta = \sqrt{\epsilon}/2$ and $2\sqrt{\epsilon}$, and it can be verified that the claim at the beginning of this proof is true when $\sqrt{\epsilon} \to 0$.

Now that we find cubes that lies in and contains A_ϵ, we can bound $\nu(A_\epsilon)$ by the product of prior measures on each dimension of these cubes. For any prior $p(\omega_{ij})$ satisfying the conditions stated in the part (2) of the theorem, $\int_{\omega_{ij0}-\sqrt{\epsilon}/Mp}^{\omega_{ij0}+\sqrt{\epsilon}/Mp} p(\omega_{ij})d\omega_{ij} \propto \frac{\sqrt{\epsilon}}{Mp}$ since the density is bounded above. The horseshoe prior also satisfies these conditions when $\omega_{ij0} \neq 0$ Carvalho et al. (2010), so the same formula holds for graphical horseshoe prior when $\omega_{ij0} \neq 0$. Taking $\epsilon = 1/n$ and summing over p^2 dimensions completes the proof of Part (2) of Theorem 3.3.1.

By Theorem 1 in Carvalho et al. (2010), the horseshoe prior has tight bounds when $\tau = 1$. Using these bounds, $K/2 \int_0^{\sqrt{\epsilon}/Mp} \log(1 + 4/\omega_{ij}^2)d\omega_{ij} < \int_0^{\sqrt{\epsilon}/Mp} p(\omega_{ij})d\omega_{ij}$ when $\omega_{ij0} = 0$, $K = 1/\sqrt{2\pi^3}$. Using the variable change in the proof of Theorem 4 in Carvalho et al. (2010), let $u = 4/\omega_{ij}^2$, then integrate by parts

$$\begin{aligned}&\int_0^{\sqrt{\epsilon}/Mp} \log(1 + 4/\omega_{ij}^2)d\omega_{ij}\\ =&\int_{4M^2p^2/\epsilon}^{\infty} \log(1 + u)u^{-3/2}\mathrm{d}u\\ =&\frac{2\sqrt{\epsilon}}{Mp}\log\left(1 + \frac{4M^2p^2}{\epsilon}\right) + 4\left(\frac{\pi}{2} - \arctan\sqrt{\frac{4M^2p^2}{\epsilon}}\right).\end{aligned}$$

After some algebra and taking $\epsilon = 1/n$, the final expression is $\frac{2}{M\sqrt{n}p}\log(1+4M^2np^2) + \frac{2}{M\sqrt{n}p} - O\{(\frac{1}{4M^2np^2})^{3/2}\} > \frac{4}{M\sqrt{n}p}\log(2M\sqrt{n}p)$. Having fixed values of τ other than 1 does not change the rate of this integration with respect to $\sqrt{\epsilon}/Mp$. Part (1) of Theorem 3.3.1 is derived.

B.2 Proof of Theorem 3.4.1

First, consider the posterior mean estimate under the graphical horseshoe prior. It is obvious that $\hat{\omega}'_{pj} \,|\, Y_{(-p)} \sim \text{Normal}(\omega'_{pj0}, 1)$ and $\hat{\omega}'^2_{pj} \,|\, Y_{(-p)} \sim \text{Noncentral}\,\chi^2(1, \omega'^2_{pj0})$. From the horseshoe prior, $\omega_{pj} \sim \text{Normal}(0, \lambda_{pj}^2\tau^2)$ and $\omega'_{pj} \sim \text{Normal}(0, \lambda_{pj}^2\tau^2\omega_{pp}^{-1}m^{-1})$

where $m = \{(Y'_{(-p)}Y_{(-p)})^{-1}_{jj}\}^{-1}$. We use ω_{pp} and ω_{pp0} interchangeably in the proof since all the diagonal elements are assumed known. Then

$$\hat{\omega}'_{pj} \,|\, Y_{(-p)}, \lambda^2_{pj}, \tau^2, \sim \text{Normal}(0, 1 + \lambda^2_{pj}\tau^2\omega^{-1}_{pp}m^{-1}).$$

To get the marginal distribution of $\hat{\omega}'_{pj}$, integrate the local shrinkage parameter λ_{pj},

$$m(\hat{\omega}'^2_{pj}) \propto \int_0^\infty (2\pi)^{-1/2}\pi^{-1}(1+\lambda^2_{pj}\tau^2\omega^{-1}_{pp}m^{-1})^{-1/2}\exp\left\{-\frac{\hat{\omega}'^2_{pj}}{2(1+\lambda^2_{pj}\tau^2\omega^{-1}_{pp}m^{-1})}\right\}\frac{1}{1+\lambda^2_{pj}}\mathrm{d}\lambda_{pj}.$$

Let $Z_{pj} = 1/(1+\lambda^2_{pj}\tau^2\omega^{-1}_{pp}m^{-1})$ so that the Jaobian is $\frac{\partial\lambda_{pj}}{\partial Z_{pj}} = -\frac{1}{2}\left\{\frac{1}{\omega^{-1}_{pp}m^{-1}\tau^2}(\frac{1}{Z_{pj}}-1)\right\}^{-1/2}$ $\frac{Z^{-2}_{pj}}{\omega^{-1}_{pp}m^{-1}\tau^2}$, then

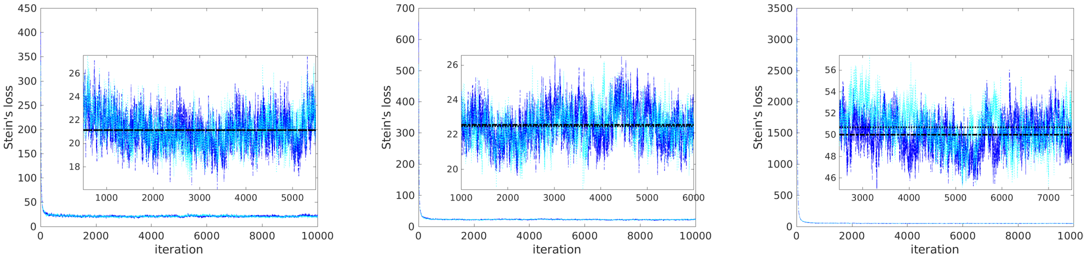
$$m(\hat{\omega}'^2_{pj}) \propto \int_0^1 \exp\left(-\frac{Z_{pj}\hat{\omega}'^2_{pj}}{2}\right)(1-Z_{pj})^{-1/2}\left\{\frac{1}{\omega^{-1}_{pp}m^{-1}Z^2_{pj}} + (1-\frac{1}{\omega^{-1}_{pp}m^{-1}\tau^2})Z_{pj}\right\}^{-1}\mathrm{d}Z_{pj}.$$

This expression differs only by a scale $\omega^{-1}_{pp}m^{-1}$ from expressions leading to Proposition 4.1 in Chapter 2. Using proof of Theorem 4.1 in Appendix A and Theorem 2 in Carvalho et al. (2010), the posterior mean under the horseshoe prior is $\mathrm{E}(\omega'_{pj} \,|\, Y, \tau) = (1 - \mathrm{E}(Z_{pj}))\hat{\omega}'_{pj}$, where $Z_{pj} \sim \text{CCH}(1, 1/2, 1, \hat{\omega}'^2_{pj}/2, 1, 1/\omega^{-1}_{pp}m^{-1}\tau^2)$. Let $\theta_{pj} = 1/(\omega^{-1}_{pp}m^{-1}\tau^2)$, then an upper bound for $\mathrm{E}(Z_{pj})$ is $4(C_1 + C_2)\theta_{pj}(1 + \hat{\omega}'^2_{pj}/2)/\hat{\omega}'^4_{pj}$ when $\hat{\omega}'^2_{pj}/2 > 1$, where $C_1 = 1 - 2e$ and $C_2 = \Gamma(1/2)\Gamma(2)/\Gamma(2.5)$ by Theorem 4.2 in Chapter 2. Consequently, $\mathrm{E}(Z_{pj})$ is $O(1/\hat{\omega}'^2_{pj})$ when $\hat{\omega}'_{pj} \to \infty$, completing the proof of Part (1).

Now consider the posterior mean estimate under the double-exponential prior in Part (2). Since double-exponential distribution is a scale mixture of normals Park and Casella (2008), the posterior mean estimate has expression $\mathrm{E}(\omega'_{pj} \,|\, Y)_{lasso} = \hat{\omega}'_{pj} + \frac{\mathrm{d}}{\mathrm{d}\hat{\omega}'_{pj}}\log m_{lasso}(\hat{\omega}'_{pj})$ by Theorem 2 in Carvalho et al. (2010). Equation (5) in Carvalho et al. (2009) states that $lim_{|\hat{\omega}'_{pj}|\to\infty}\frac{\mathrm{d}}{\mathrm{d}\hat{\omega}'_{pj}}\log m_{lasso}(\hat{\omega}'_{pj}) = \pm a$, where a varies inversely with the global shrinkage parameter in the prior. The proof of this statement is in Pericchi and Smith (1992).

Now consider the condition that $\hat{\omega}'^2_{pj}$ is large. $(Y'_{(-p)}Y_{(-p)})^{-1}$ follows an inverse Wishart distribution with scale matrix $\Sigma^{-1}_{(-p)}$ and n degrees of freedom, where $\Sigma_{(-p)}$ is the covariance matrix without the pth column and pth row. Consequently, $(Y'_{(-p)}Y_{(-p)})^{-1}_{jj}$ follows a one-dimensional inverse Wishart distribution with scale $\Sigma^{-1}_{(-p)jj}$ and $n-p+2$ degrees of freedom, and its inverse $\{(Y'_{(-p)}Y_{(-p)})^{-1}_{jj}\}^{-1}$ follows a Wishart distribution with scale $\{\Sigma^{-1}_{(-p)jj}\}^{-1}$ and $n-p+2$ degrees of freedom, or equivalently a gamma distribution with shape parameter $(n-p+2)/2$ and scale parameter $2\{\Sigma^{-1}_{(-p)jj}\}^{-1}$. By matrix inversion in blocked form, $\Sigma^{-1}_{(-p)} = \Omega_{(-p)} - \boldsymbol{\omega}_{(-p)p}\boldsymbol{\omega}_{p(-p)}/\omega_{pp}$, so that the scale parameter of the gamma distribution is $2(\omega_{jj0} - \omega^2_{jp0}/\omega_{pp0})^{-1}$, as claimed in Part (3).

B.3 MCMC Convergence Diagnostics



(a) $p = 100$, $n = 50$

(b) $p = 200$, $n = 120$

(c) $p = 400$, $n = 120$

Figure B.1.: Stein's loss of the sampled Ω at each iteration using Algorithm 1 for graphical horseshoe, under (a) hubs structure, $p = 100$, $n = 50$, and (b) hubs structure, $p = 200$, $n = 120$, (c) hubs structure, $p = 400$, $n = 120$. The first data set in the corresponding simulations are used. The dashed line and dotted line show Stein's loss of samples from two chains with different starting values, a $p \times p$ identity matrix and a random $p \times p$ positive definite symmetric matrix. The enlarged area show Stein's loss in a shorter range of iterations, and the horizontal lines show the average Stein's loss of iterations within that range.

We evaluate the convergence and mixing of the proposed graphical horseshoe Gibbs sampler by plotting Stein's loss of sampled Ω across iterations (i.e., a trace plot), using different starting values for each chain. It is shown that when $p = 100$ and $p = 200$, MCMC samples using Algorithm 1 converge within 500 iterations, and mix reasonably well. When $p = 400$, the algorithm takes longer to converge. In the simulations, we use more burn-in samples when the dimension is higher. We use 500 burn-in samples when $p = 100$; 1000 burn-in samples when $p = 200$; and 2500 burn-in samples when $p = 400$. We use 5000 iterations in all cases, for both graphical horseshoe and Bayesian graphical lasso. Figure B.1 shows the plots used for MCMC diagnostics. Formal tests such as Gelman–Rubin diagnostics could be carried out using the MCMC output, if desired.

B.4 Additional Simulation Results

We provide additional simulation results in this section. The purpose is two-fold:

1. Tables B.1, B.2 and B.3 provide estimates of sensitivity, specificity, precision and accuracy for the same settings used in Section 3.5, complementing the TPR and FPR presented in Tables 3.1, 3.2 and 3.3.

2. Following requests by the referees, Table B.4 provides results on a larger simulation setting, with $p = 400, n = 120$, for the hubs and cliques negative structures.

Table B.1.: Sensitivity (true positive/(true positive+false negative)=$TP/(TP+FN)$), specificity (true negative/(true negative+false positive)=$TN/(TN+FP)$), precision ($TP/(TP+FP)$), and accuracy ($(TP+TN)/(TP+TN+FP+FN)$) of precision matrix estimates over 50 data sets generated by multivariate normal distributions with precision matrix Ω_0, where $p = 100$ and $n = 50$. The precision matrix is estimated by frequentist graphical lasso with penalized diagonal elements (GL1), frequentist graphical lasso with unpenalized diagonal elements (GL2), graphical SCAD (GSCAD), Bayesian graphical lasso (BGL), and graphical horseshoe (GHS). The best performer in each row is shown in bold.

			Random					Hubs		
nonzero pairs			35/4950					90/4950		
nonzero elements			$\sim -\text{Unif}(0.2, 1)$					0.25		
$p = 100, n = 50$	GL1	GL2	GSCAD	BGL	GHS	GL1	GL2	GSCAD	BGL	GHS
SEN	.8246	.7097	**.9977**	.8709	.5903	.8649	.7333	**.9987**	.8513	.2687
	(.0520)	(.0620)	(.0078)	(.0470)	(.0537)	(.0443)	(.0751)	(.0053)	(.0378)	(.0764)
SPE	.9053	.9626	.0045	.8945	**.9996**	.9081	.9719	.0024	.8811	**.9987**
	(.0141)	(.0070)	(.0102)	(.0059)	(.0004)	(.0130)	(.0086)	(.0069)	(.0058)	(.0006)
PREC	.0593	.1213	.0071	.0556	**.9134**	.1503	.3378	.0182	.1172	**.8031**
	(.0073)	(.0166)	(<.0001)	(.0039)	(.0626)	(.0166)	(.0559)	(<.0001)	(.0057)	(.0677)
ACC	.9048	.9608	.0116	.8943	**.9967**	.9074	.9676	.0205	.8806	**.9855**
	(.0138)	(.0067)	(.0101)	(.0058)	(.0005)	(.0123)	(.0075)	(.0067)	(.0055)	(.0012)
			Cliques positive					Cliques negative		
nonzero pairs			30/4950					30/4950		
nonzero elements			-0.45					0.75		
$p = 100, n = 50$	GL1	GL2	GSCAD	BGL	GHS	GL1	GL2	GSCAD	BGL	GHS
SEN	**1**	**1**	**1**	**1**	.7487	.9993	.9880	**1**	.9993	.9733
	(0)	(0)	(0)	(0)	(.0427)	(.0047)	(.0221)	(0)	(.0047)	(.0421)
SPE	.9100	.9745	.0099	.8986	**.9997**	.9078	.9721	.0248	.8839	**.9990**
	(.0098)	(.0056)	(.0177)	(.0052)	(.0003)	(.0135)	(.0084)	(.0219)	(.0051)	(.0005)
PREC	.0641	.1991	.0061	.0569	**.9352**	.0632	.1881	.0062	.0500	**.8611**
	(.0067)	(.0365)	(.0001)	(.0027)	(.0541)	(.0090)	(.0448)	(.0001)	(.0021)	(.0615)
ACC	.9106	.9747	.0159	.8992	**.9981**	.9084	.9722	.0307	.8846	**.9988**
	(.0097)	(.0055)	(.0176)	(.0052)	(.0004)	(.0135)	(.0083)	(.0218)	(.0051)	(.0005)

Table B.2.: Sensitivity (true positive/(true positive+false negative)=$TP/(TP+FN)$), specificity (true negative/(true negative+false positive)=$TN/(TN+FP)$), precision ($TP/(TP+FP)$), and accuracy ($(TP+TN)/(TP+TN+FP+FN)$) of precision matrix estimates over 50 data sets generated by multivariate normal distributions with precision matrix Ω_0, where $p = 100$ and $n = 120$. The precision matrix is estimated by frequentist graphical lasso with penalized diagonal elements (GL1), frequentist graphical lasso with unpenalized diagonal elements (GL2), refitted graphical lasso (RGL), graphical SCAD (GSCAD), Bayesian graphical lasso (BGL), and graphical horseshoe (GHS). The best performer in each row is shown in bold.

	Random						Hubs					
nonzero pairs	35/4950						90/4950					
nonzero elements	$\sim -\text{Unif}(0.2, 1)$						0.25					
$p = 100, n = 120$	GL1	GL2	RGL	GSCAD	BGL	GHS	GL1	GL2	RGL	GSCAD	BGL	GHS
SEN	.9486	.8794	.6497	**.9994**	.9760	.8149	.9936	.9844	.8376	**.9998**	.9938	.8671
	(.0316)	(.0384)	(.0658)	(.0040)	(.0233)	(.0397)	(.0078)	(.0154)	(.0617)	(.0016)	(.0072)	(.0396)
SPE	.8971	.9558	.9891	.0017	.8322	**.9995**	.8971	.9569	**.9985**	.0012	.8128	.9967
	(.0150)	(.0077)	(.0029)	(.0055)	(.0066)	(.0002)	(.0161)	(.0093)	(.0009)	(.0029)	(.0066)	(.0011)
PREC	.0627	.1265	.3075	.0071	.0398	**.9248**	.1541	.3044	**.9131**	.0182	.0896	.8334
	(.0084)	(.0170)	(.0521)	(<.0001)	(.0018)	(.0392)	(.0191)	(.0490)	(.0408)	(<.0001)	(.0029)	(.0476)
ACC	.8975	.9552	.9867	.0088	.8332	**.9982**	.8988	.9574	**.9955**	.0193	.8161	.9944
	(.0149)	(.0075)	(.0026)	(.0054)	(.0065)	(.0004)	(.0158)	(.0091)	(.0010)	(.0029)	(.0064)	(.0010)
	Cliques positive						Cliques negative					
nonzero pairs	30/4950						30/4950					
nonzero elements	-0.45						0.75					
$p = 100, n = 120$	GL1	GL2	RGL	GSCAD	BGL	GHS	GL1	GL2	RGL	GSCAD	BGL	GHS
SEN	**1**	**1**	**1**	**1**	**1**	.9840	**1**	**1**	.9947	**1**	**1**	**1**
	(0)	(0)	(0)	(0)	(0)	(.0236)	(0)	(0)	(.0170)	(0)	(0)	(0)
SPE	.9001	.9713	**.9997**	.0021	.8420	.9996	.8959	.9587	.9990	.0061	.8224	**.9992**
	(.0078)	(.0064)	(.0004)	(.0061)	(.0074)	(.0002)	(.0100)	(.0088)	(.0008)	(.0073)	(.0068)	(.0004)
PREC	.0579	.1812	**.9574**	.0061	.0372	.9459	.0558	.1331	.8728	.0061	.0332	**.8882**
	(.0046)	(.0335)	(.0571)	(<.0001)	(.0017)	(.0388)	(.0053)	(.0241)	(.0927)	(<.0001)	(.0012)	(.0537)
ACC	.9007	.9715	**.9997**	.0082	.8430	.9996	.8966	.9590	.9990	.0121	.8234	**.9992**
	(.0078)	(.0063)	(.0004)	(.0060)	(.0073)	(.0002)	(.0099)	(.0088)	(.0008)	(.0072)	(.0067)	(.0004)

Table B.3.: Sensitivity (true positive/(true positive+false negative)=$TP/(TP+FN)$), specificity (true negative/(true negative+false positive)=$TN/(TN+FP)$), precision ($TP/(TP+FP)$), and accuracy ($(TP+TN)/(TP+TN+FP+FN)$) of precision matrix estimates over 50 data sets generated by multivariate normal distributions with precision matrix Ω_0, where $p = 200$ and $n = 120$. The precision matrix is estimated by frequentist graphical lasso with penalized diagonal elements (GL1), frequentist graphical lasso with unpenalized diagonal elements (GL2), graphical SCAD (GSCAD), Bayesian graphical lasso (BGL), and graphical horseshoe (GHS). The best performer in each row is shown in bold.

			Random					Hubs		
nonzero pairs			29/19900					180/19900		
nonzero elements			$\sim -\text{Unif}(0.2, 1)$					0.25		
$p = 200, n = 120$	GL1	GL2	GSCAD	BGL	GHS	GL1	GL2	GSCAD	BGL	GHS
SEN	.9476	.8393	**1**	.9855	.8421	.9911	.9773	**1**	.9917	.7754
	(.0370)	(.0301)	(0)	(.0232)	(.0369)	(.0065)	(.0132)	(0)	(.0060)	(.0323)
SPE	.9486	.9841	.0049	.8965	**.9999**	.9343	.9743	.0002	.8803	**.9989**
	(.0065)	(.0021)	(.0095)	(.0031)	(<.0001)	(.0053)	(.0064)	(.0002)	(.0027)	(.0002)
PREC	.0265	.0722	.0015	.0137	**.9334**	.1218	.2662	.0090	.0703	**.8693**
	(.0031)	(.0077)	(<.0001)	(<.0001)	(.0463)	(.0105)	(.0467)	(<.0001)	(.0014)	(.0273)
ACC	.9486	.9838	.0064	.8967	**.9997**	.9348	.9743	.0093	.8813	**.9969**
	(.0065)	(.0021)	(.0094)	(.0030)	(<.0001)	(.0052)	(.0063)	(.0002)	(.0027)	(.0003)
			Cliques positive					Cliques negative		
nonzero pairs			60/19900					60/19900		
nonzero elements			-0.45					0.75		
$p = 200, n = 120$	GL1	GL2	GSCAD	BGL	GHS	GL1	GL2	GSCAD	BGL	GHS
SEN	**1**	**1**	**1**	**1**	.9633	**1**	**1**	**1**	**1**	**1**
	(0)	(0)	(0)	(0)	(.0226)	(0)	(0)	(0)	(0)	(0)
SPE	.9337	.9828	.0031	.9014	**.9998**	.9364	.9771	.0081	.8845	**.9996**
	(.0044)	(.0027)	(.0051)	(.0030)	(.0001)	(.0039)	(.0039)	(.0072)	(.0027)	(.0001)
PREC	.0438	.1516	.0030	.0298	**.9465**	.0455	.1198	.0030	.0255	**.8973**
	(.0027)	(.0180)	(<.001)	(.0008)	(.0312)	(.0023)	(.0209)	(<.0001)	(.0005)	(.0367)
ACC	.9339	.9828	.0061	.9017	**.9997**	.9366	.9772	.0111	.8849	**.9996**
	(.0044)	(.0027)	(.0051)	(.0030)	(.0001)	(.0038)	(.0039)	(.0072)	(.0027)	(.0001)

Table B.4.: Mean (sd) Stein's loss, Frobenius norm,, sensitivity (true positive/(true positive+false negative)=$TP/(TP+FN)$), specificity (true negative/(true negative+false positive)=$TN/(TN+FP)$), precision ($TP/(TP+FP)$), and accuracy ($(TP+TN)/(TP+TN+FP+FN)$) of precision matrix estimates over 20 data sets generated by multivariate normal distributions with precision matrix Ω_0, where $p = 400$ and $n = 120$. The precision matrix is estimated by frequentist graphical lasso with penalized diagonal elements (GL1), frequentist graphical lasso with unpenalized diagonal elements (GL2), graphical SCAD (GSCAD), Bayesian graphical lasso (BGL), and graphical horseshoe (GHS). The best performer in each row is shown in bold.

			Hubs					Cliques negative		
nonzero pairs			360/79800					120/79800		
nonzero elements			0.25					0.75		
$p = 400, n = 120$	GL1	GL2	GSCAD	BGL	GHS	GL1	GL2	GSCAD	BGL	GHS
Stein's loss	29.27	35.53	28.75	333.98	**26.92**	33.90	42.91	33.71	352.22	**8.36**
	(0.90)	(0.55)	(1.00)	(2.34)	(0.97)	(0.74)	(1.04)	(0.73)	(3.03)	(0.62)
F norm	7.07	8.05	6.96	12.09	**5.60**	11.30	12.80	11.28	12.92	**4.00**
	(0.19)	(0.05)	(0.21)	(0.19)	(0.11)	(0.23)	(0.10)	(0.22)	(0.11)	(0.16)
SEN	.9826	.9569	**1**	.9871	.6844	**1**	**1**	**1**	**1**	**1**
	(.0092)	(.0130)	(0)	(.0063)	(.0302)	(0)	(0)	(0)	(0)	(0)
SPE	.9651	.9882	.0001	.9378	**.9996**	.9580	.9884	.0102	.9359	**.9998**
	(.0065)	(.0006)	(<.0001)	(.0011)	(<.0001)	(.0059)	(.0015)	(.0119)	(.0010)	(<.0001)
PREC	.1157	.2690	.0045	.0671	**.8884**	.0353	.1163	.0015	.0230	**.9052**
	(.0166)	(.0111)	(<.0001)	(.0013)	(.0221)	(.0052)	(.0103)	(<.0001)	(.0003)	(.0228)
ACC	.9652	.9880	.0046	.9380	**.9982**	.9581	.9885	.0117	.9360	**.9998**
	(.0065)	(.0006)	(<.0001)	(.0011)	(.0001)	(.0059)	(.0015)	(.0119)	(.0010)	(<.0001)
Avg CPU time	0.22	0.25	2.03e+3	5.48e+3	6.63e+3	0.30	0.33	2.01e+3	5.81e+3	5.52e+3

C. SUPPLEMENTARY MATERIAL TO CHAPTER 4

C.1 Proof of Theorem 3.3.1

Let $A_\epsilon = \{\{B, \Omega\} : \frac{1}{n} D_n(p_{B_0,\Omega_0} || p_{B,\Omega}) \leq \epsilon\}$. We claim that $A_\epsilon \subset \mathbb{R}^{p\times q} \times \mathbb{R}^{q\times q}$ is bounded by an Euclidean cube of $pq + q^2$ dimensions with $(\beta_{j0} - k_1\epsilon^{1/4}/pq^{1/2}, \beta_{j0} + k_1\epsilon^{1/4}/pq^{1/2})$, and $(\omega_{kl0} - k_2\epsilon^{1/2}/q, \omega_{kl0} + k_2\epsilon^{1/2}/q)$ on each dimension. The proof is as following.

Let $B = B_0 + (\epsilon^{1/4}/pq^{1/2})\nVdash_{p\times q}$, $\Omega = \Omega_0 + (\epsilon^{1/2}/q)\nVdash_{q\times q}$, where $\nVdash_{m\times n}$ denotes a $m \times n$ matrix with all elements equal to 1. Then,

$$\begin{aligned} D_n(p_{B_0,\Omega_0}||p_{B,\Omega}) =& \frac{n}{2}\{\log|\Omega^{-1}\Omega_0| + tr(\Omega\Omega_0^{-1}) - q\} + \frac{1}{2} vec(XB - XB_0)'(\Omega \otimes I_n) vec(XB - XB_0) \\ :=& \mathrm{I} + \mathrm{II}. \end{aligned}$$

By the proof of Theorem 3.2 in Chapter 3, I $\propto n\epsilon$ when $\epsilon \to 0$. We will show that II $\propto n\epsilon$ as well. The expression for II is simplified as,

$$\begin{aligned} \mathrm{II} =& \frac{1}{2} vec(XB - XB_0)'(\Omega \otimes I_n) vec(XB - XB_0) \\ =& \frac{1}{2}\frac{\epsilon^{1/4}}{pq^{1/2}} vec(X\nVdash_{p\times q})' \left\{\left(\Omega_0 + \frac{\epsilon^{1/2}}{q}\nVdash_{q\times q}\right) \otimes I_n\right\} \frac{\epsilon^{1/4}}{pq^{1/2}} vec(X\nVdash_{p\times q}) \\ =& \frac{1}{2}\frac{\epsilon^{1/2}}{p^2 q} vec(X\nVdash_{p\times q})' \left\{\Omega_0 \otimes I_n + \left(\frac{\epsilon^{1/2}}{q}\nVdash_{q\times q}\right) \otimes I_n\right\} vec(X\nVdash_{p\times q}). \end{aligned}$$

Some algebra shows that $vec(X\mathbb{1}_{p\times q})'(\Omega_0 \otimes I_n)vec(X\mathbb{1}_{p\times q}) = \sum_{k,l}\omega_{kl0}\sum_i(X_{i1}+\ldots+X_{ip})^2$ and $vec(X\mathbb{1}_{p\times q})'(\mathbb{1}_{q\times q}\otimes I_n)vec(X\mathbb{1}_{p\times q}) = q^2\sum_i(X_{i1}+\ldots+X_{ip})^2$. Therefore,

$$\begin{aligned}
\text{II} =& \frac{1}{2}\frac{\epsilon^{1/2}}{p^2q}\left\{\sum_{k,l}\omega_{kl0}\sum_i(X_{i1}+\ldots+X_{ip})^2 + \frac{\epsilon^{1/2}}{q}q^2\sum_i(X_{i1}+\ldots+X_{ip})^2\right\}\\
=&\frac{1}{2}\frac{\epsilon^{1/2}}{p^2q}(c_1np^2q + c_2\epsilon^{1/2}np^2q)\\
=&\frac{1}{2}(c_1n\epsilon^{1/2} + c_2n\epsilon).
\end{aligned}$$

Combining I and II, $\frac{1}{n}D_n(p_{B_0,\Omega_0}||p_{B,\Omega}) \propto \epsilon$ when $\epsilon \to 0$. We have proved that A_ϵ is bounded by cubes of $pq+q^2$ dimensions described above. Now that we find cubes that bound A_ϵ, we will bound $\nu(A_\epsilon)$ by the product of prior measures on each dimension of these cubes. For any prior measure with density $p(\beta_j)$ that is continuous, bounded above, and strictly positive on a neighborhood of the true β_{j0}, one has $\int_{\beta_{j0}-\epsilon^{1/4}/(pq^{1/2})}^{\beta_{j0}+\epsilon^{1/4}/(pq^{1/2})} p(\beta_j)d\beta_j \propto \epsilon^{1/4}/(pq^{1/2})$, since the density is bounded above. Similarly, $\int_{\omega_{kl0}-\epsilon^{1/2}/q}^{\omega_{kl0}+\epsilon^{1/2}/q} p(\omega_{kl})d\omega_{kl} \propto \epsilon^{1/2}/q$, for any prior density $p(\omega_{kl})$ satisfying the conditions. Taking $\epsilon = 1/n$, this gives $\log\nu(A_{1/n})$ in Part(1) of Theorem 3.3.1. The horseshoe prior also satisfies conditions in (1) in dimensions where $\beta_{j0} \neq 0$ and $\omega_{kl0} \neq 0$, so the same measures hold for HS-GHS in nonzero dimensions.

Now we need prior measure of horseshoe prior on dimensions where $\beta_{j0} = 0$ and $\omega_{kl0} = 0$. Using bounds of horseshoe prior provided in Carvalho et al. (2010), it has been established by Chapter 3 that $\int_0^{\epsilon^{1/2}/q} p(\omega_{kl})d\omega_{kl} > c_3\log(\epsilon^{-1/2}q)/(\epsilon^{-1/2}q)$. Similar calculations show that $\int_0^{\epsilon^{1/4}pq^{1/2}} p(\beta_j)d\beta_j > c_4\log(\epsilon^{-1/4}pq^{1/2})/(\epsilon^{-1/4}pq^{1/2})$. Taking $\epsilon = 1/n$, this gives Part (2) of the theorem and completes the proof.

C.2 MCMC Convergence Diagnostics

Figure C.1 shows the trace plots of the log likelihood over 6,000 MCMC iterations and the inside panel in each plot shows the trace plot after discarding the first 1,000 draws as burn-in samples. The plots indicate quick mixing. Formal MCMC diag-

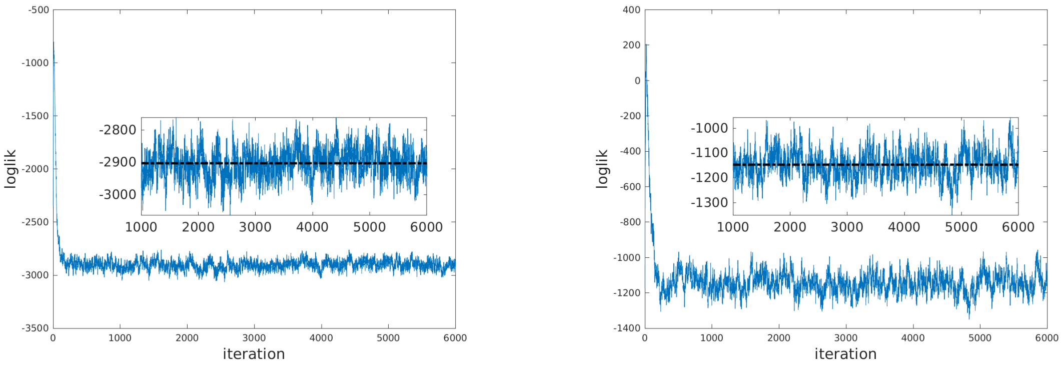


Figure C.1.: Loglikelihood at each iteration using Algorithm 2 for HS-GHS, under (a) AR1 structured inverse covariance matrix, $p = 120$, $q = 50$, and (b) AR1 structured inverse covariance matrix, $p = 200$, $q = 25$. Horizontal lines show log likelihood averaged in interation 1000 to 6000. The first data set in the corresponding simulations are used.

nostics, such as Gelman–Rubin test could be performed using the MCMC output, if desired.

C.3 Additional Simulation Results

We provide additional simulation results, complementing those in Section 4.5. Tables C.1 and C.2 provide results when $p = 100$, $q = 25$. Tables C.3 and C.4 supplement Tables 4.1 and 4.2 with more simulation settings. In the star structured inverse covariance matrix, $\omega_{1k} = 0.25$, $k = 2, ..., q$, all diagonal elements equal to 1, and the rest of the elements all equal to 0. In the case of large coefficients, all nonzero coefficients equal to 5. Other structures of the inverse covariance matrix and uniform distributed coefficients are described in Section 4.5. One fifth of the coefficients are nonzero when $p = 100$ and $q = 25$, and 1/20 of the coefficients are nonzero in the other dimensions.

Table C.1.: Mean squared error (sd) in estimation and prediction, average Kullback-Leibler divergence, and sensitivity, specificity and precision of variable selection performance, over 50 simulated data sets. The regression coefficients and precision matrix are estimated by HS-GHS, joint high-dimensional Bayesian variable and covariance selection (BM13), MRCE and CAPME. The best performer in each column is shown in bold.

	Simulation 1: $p = 100$, $q = 25$, $n = 100$, Uniform coefficients, AR1 structure										
	MSE			Divergence	B support recovery			Ω support recovery			CPU time
Method	B	Ω	Prediction	avg KL	SEN	SPE	PRC	SEN	SPE	PRC	min.
HS-GHS	**0.0166**	0.0047	**2.6235**	**4.6663**	.9674	**.9695**	**.8885**	.9942	**.9959**	**.9565**	96.53
	(0.0017)	(0.0015)	(0.1740)	(0.4478)	(.0136)	(.0060)	(.0196)	(.0146)	(.0034)	(.0356)	
BM13	0.1396	0.0313	4.9680	12.7152	-	-	-	.5533	.9903	.8363	6.17
	(0.0035)	(0.0012)	(0.3073)	(0.3328)	-	-	-	(.0758)	(.0051)	(.0760)	
MRCE	0.0230	**0.0034**	2.7459	5.0754	**.9952**	.6373	.4076	**.9992**	.8249	.3399	24.27
	(0.0022)	(0.0011)	(0.1851)	(0.4761)	(.0045)	(.0267)	(.0178)	(.0059)	(.0416)	(.0543)	
CAPME	0.0460	0.0253	3.2043	8.4143	.9761	.5775	.3704	.5075	.9801	.7184^1	40.06
	(0.0061)	(0.0100)	(0.2287)	(1.3079)	(.0141)	(.0747)	(.0382)	(.4931)	(.0294)	(.1354)	
	Simulation 2: $p = 100$, $q = 25$, Uniform coefficients, Star structure										
	MSE			Divergence	B support recovery			Ω support recovery			CPU time
Method	B	Ω	Prediction	avg KL	SEN	SPE	PRC	SEN	SPE	PRC	min.
HS-GHS	**0.0138**	**0.0058**	**1.5459**	**4.6722**	.9789	**.9630**	**.8693**	.5089	.9955	**.8882**	96.62
	(0.0014)	(0.0018)	(0.0856)	(0.4392)	(.0100)	(.0069)	(.0211)	(.1540)	(.0051)	(.1153)	
BM13	0.1362	0.0188	3.8594	12.2304	-	-	-	.0289	.9943	.2307^2	4.91
	(0.0034)	(0.0004)	(0.2270)	(0.1689)	-	-	-	(.0359)	(.0037)	(.2708)	
MRCE	0.0193	0.0109	1.6357	6.2894	.9938	.6270	.4004	**.9167**	.8575	.3356^3	21.17
	(0.0021)	(0.0033)	(0.0863)	(0.6595)	(.0051)	(.0252)	(.0159)	(.1761)	(.0731)	(.1439)	
CAPME	0.0255	0.0143	1.8071	5.6583	**.9954**	.5099	.3377	0	**1**	-^{4}	40.33
	(0.0026)	(0.0012)	(0.1016)	(0.2677)	(.0043)	(.0379)	(.0174)	(0)	(0)	-	

1.23 NaNs in 50 replicates. 2. 5 NaNs in 50 replicates. 3. 1 NaN in 50 replicates. 4. 50 NaNs. All mean and sd. calculated on non-NaN values.

Table C.2.: Mean squared error (sd) in estimation and prediction, average Kullback-Leibler divergence, and sensitivity, specificity and precision of variable selection performance, over 50 simulated data sets. The regression coefficients and precision matrix are estimated by HS-GHS, joint high-dimensional Bayesian variable and covariance selection (BM13), MRCE and CAPME. The best performer in each column is shown in bold.

	Simulation 3: $p = 100$, $q = 25$, Uniform coefficients, Cliques structure										
	MSE			Divergence	B support recovery			Ω support recovery			CPU time
Method	B	Ω	Prediction	avg KL	SEN	SPE	PRC	SEN	SPE	PRC	min.
HS-GHS	**0.0261**	**0.0044**	**3.3854**	**4.5777**	.9148	**.9701**	**.8846**	**1**	.9952	.9499	96.53
	(0.0026)	(0.0018)	(0.1745)	(0.4076)	(.0188)	(.0057)	(.0194)	(0)	(.0044)	(.0437)	
BM13	0.1417	0.0601	5.5897	11.4533	-	-	-	.4567	.9988	**.9674**	4.80
	(0.0038)	(0.0021)	(0.3192)	(0.3326)	-	-	-	(.1163)	(.0019)	(.0611)	
MRCE	0.0363	0.0147	3.5770	6.4222	**.9763**	.6443	.4079	**1**	.6924	.2293	24.55
	(.0036)	(.0057)	(.1968)	(0.7096)	(.0110)	(.0300)	(.0198)	(0)	(.0815)	(.0457)	
CAPME	0.0534	0.0668	3.9208	8.8355	.9697	.5535	.3538	0	**1**	-[1]	40.78
	(0.0053)	(0.0009)	(0.2257)	(0.2505)	(.0119)	(.0530)	(.0260)	(0)	(0)	-	

	Simulation 4: $p = 100$, $q = 25$, Coefficients=5, AR1 structure										
	MSE			Divergence	B support recovery			Ω support recovery			CPU time
Method	B	Ω	Prediction	avg KL	SEN	SPE	PRC	SEN	SPE	PRC	min.
HS-GHS	**0.0125**	**0.0057**	**2.5508**	**4.5945**	**1**	**.9669**	**.8836**	**.9950**	.9954	**.9514**	97.66
	(0.0013)	(0.0018)	(0.1705)	(0.5065)	(0)	(.0074)	(.0233)	(.0137)	(.0039)	(.0401)	
BM13	1.5774	0.0521	35.2494	34.3200	-	-	-	.2133	.9659	.3533	4.85
	(0.0325)	(<0.0001)	(2.5042)	(0.2394)	-	-	-	(.0732)	(.0086)	(.1029)	
MRCE	0.0550	0.0113	3.3325	11.9130	**1**	.1830	.2346	.9900	.8510	.3965	27.45
	(0.0094)	(0.0066)	(0.2599)	(2.1085)	(0)	(.0396)	(.0090)	(.0332)	(.0648)	(.1161)	
CAPME	0.0638	0.0377	4.5765	14.4007	**1**	.5498	.3588	0	**1**	-[2]	38.93
	(0.0086)	(0.0013)	(0.5602)	(1.4711)	(0)	(.0491)	(.0256)	(0)	(0)	-	

1,2. 50 NaNs. All mean and sd. calculated on non-NaN values.

Table C.3.: Mean squared error (sd) in estimation and prediction, average Kullback-Leibler divergence, and sensitivity, specificity and precision of variable selection performance, over 50 simulated data sets, $p = 200$ and $q = 25$. The regression coefficients and precision matrix are estimated by HS-GHS, joint high-dimensional Bayesian variable and covariance selection (BM13), MRCE and CAPME. The best performer in each column is shown in bold.

	Simulation 5: $p = 200$, $q = 25$, $n = 100$, Uniform coefficients, Star structure										
	MSE			Divergence	B support recovery			Ω support recovery			CPU time
Method	B	Ω	Prediction	avg KL	SEN	SPE	PRC	SEN	SPE	PRC	min.
HS-GHS	**0.0027**	0.0341	**1.6015**	10.3918	.9557	.9975	**.9525**	.3856	.9953	**.8523**	1.01e+03
	(0.0003)	(0.0178)	(0.0686)	(1.4390)	(.0115)	(.0008)	(.0145)	(.1277)	(.0041)	(.1130)	
BM13	0.0543	**0.0150**	7.1188	11.0194	-	-	-	.0011	.9979	.0385 [1]	54.67
	(0.0006)	(0.0004)	(0.3606)	(0.2732)	-	-	-	(.0079)	(.0025)	(.1961)	
MRCE	0.0865	0.0362	18.7449	32.2416	.0050	**1.0000**	.9932 [2]	**.9256**	.0825	.0607	0.10
	(0.0003)	(0.0004)	(0.8318)	(0.3412)	(.0043)	(<.0001)	(.0411)	(.0783)	(.0673)	(.0050)	
CAPME	0.0096	0.0221	2.3653	**8.2904**	**.9770**	.8098	.2132	0	**1**	- [3]	74.11
	(0.0009)	(0.0012)	(0.1280)	(0.4024)	(.0083)	(.0101)	(.0094)	(0)	(0)	-	
	Simulation 6: $p = 200$, $q = 25$, $n = 100$, Coefficients=5, AR1 structure										
	MSE			Divergence	B support recovery			Ω support recovery			CPU time
Method	B	Ω	Prediction	avg KL	SEN	SPE	PRC	SEN	SPE	PRC	min.
HS-GHS	**0.0017**	**0.0400**	**2.4693**	**9.5349**	**1**	**.9986**	**.9737**	.9817	.9970	**.9672**	770.02
	(0.0002)	(0.0120)	(0.1647)	(1.2234)	(0)	(.0006)	(.0108)	(.0281)	(.0039)	(.0396)	
BM13	0.7306	0.0520	80.2711	30.2897	-	-	-	0	.9898	0	103.75
	(0.0077)	(<0.0001)	(4.5903)	(0.2125)	-	-	-	(0)	(.0050)	(0)	
MRCE	1.2326	0.1333	297.9516	66.9764	.0187	.9903	.5295 [4]	**.9902**	.0079	.0799	1.05
	(0.0159)	(0.2715)	(18.2395)	(18.3250)	(.0136)	(.0166)	(.4218)	(.0249)	(.0146)	(.0018)	
CAPME	0.0202	0.0426	5.1766	14.9741	**1**	.8070	.2146	0	**1**	- [5]	66.87
	(0.0022)	(0.0010)	(0.4076)	(0.5903)	(0)	(.0097)	(.0083)	(0)	(0)	-	

1. 24 NaNs in 50 replicates. 2. 13 NaNs in 50 replicates. 3,5. 50 NaNs. 4. 5 NaNs in 50 replicates. All mean and sd. calculated on non-NaN values.

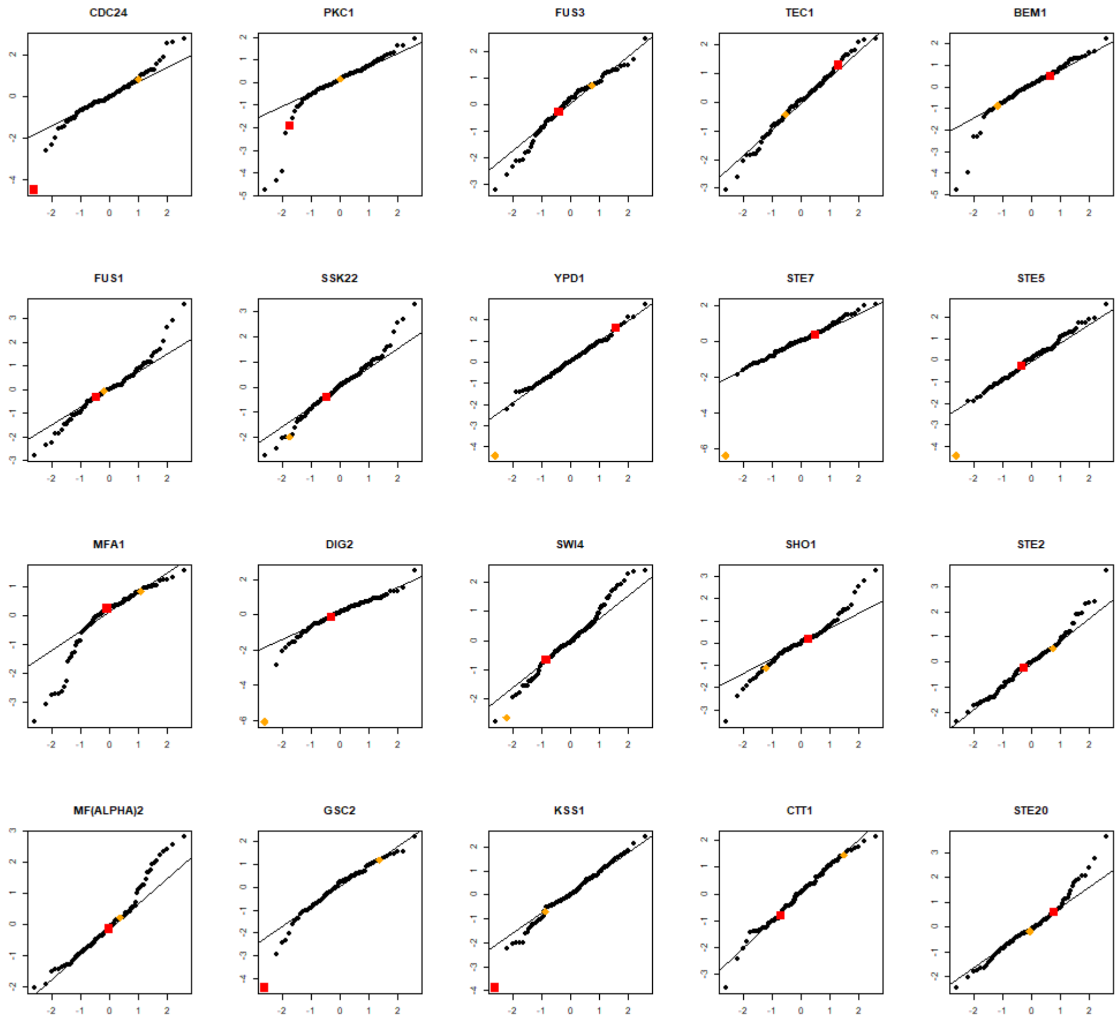
Table C.4.: Mean squared error (sd) in estimation and prediction, average Kullback-Leibler divergence, and sensitivity, specificity and precision of variable selection performance, over 50 simulated data sets, $p = 120$ and $q = 50$. The regression coefficients and precision matrix are estimated by HS-GHS, joint high-dimensional Bayesian variable and covariance selection (BM13), MRCE and CAPME. The best performer in each column is shown in bold.

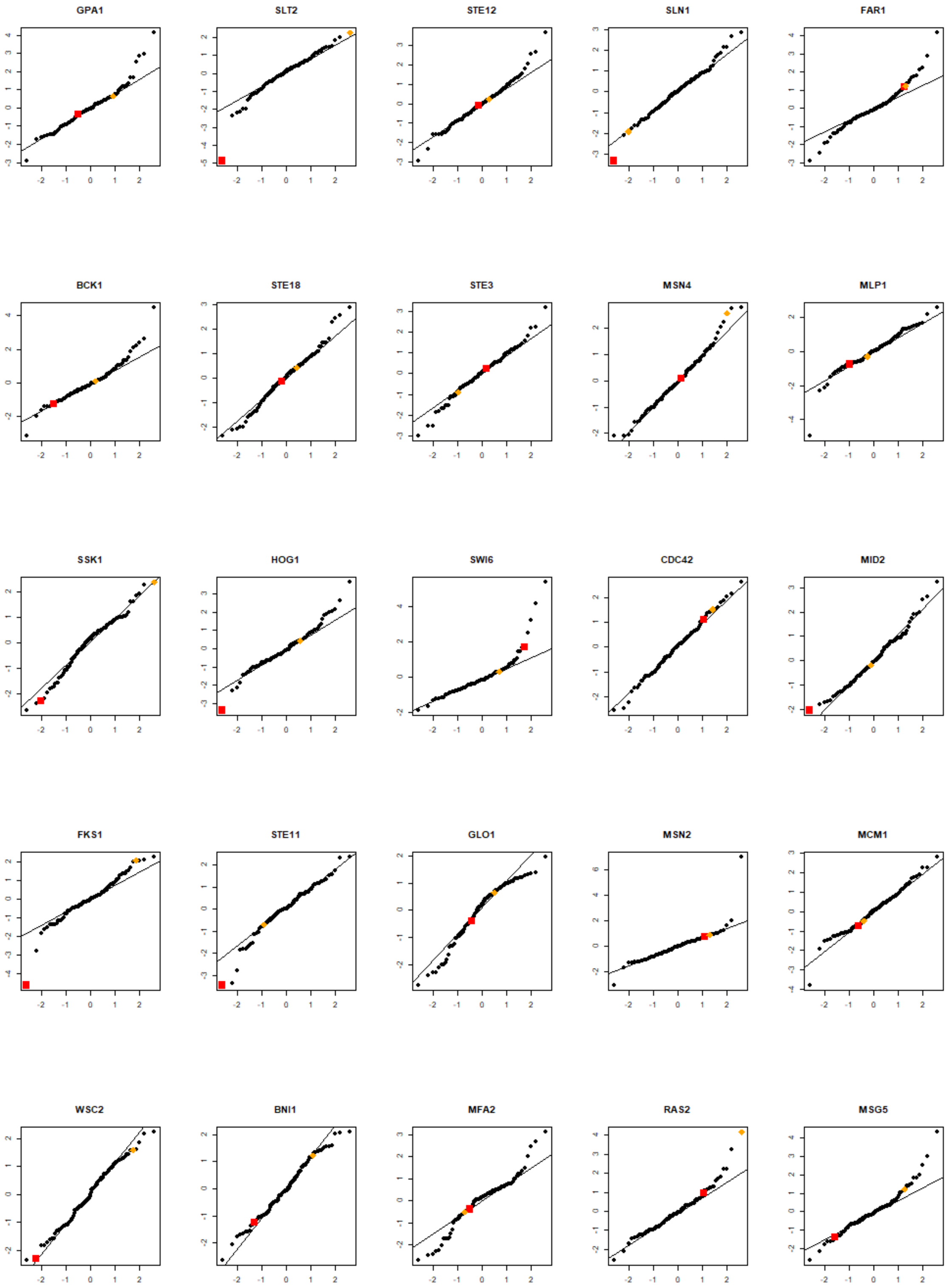
	Simulation 3: $p = 120$, $q = 50$, $n = 100$, Uniform coefficients, Star structure										
	MSE			Divergence	B support recovery			Ω support recovery			CPU time
Method	B	Ω	Prediction	avg KL	SEN	SPE	PRC	SEN	SPE	PRC	min.
HS-GHS	**0.0018**	**0.0036**	**1.2768**	**7.3906**	.9810	.9980	**.9628**	.4378	.9995	**.9380**	2.58e+03
	(0.0002)	(0.0008)	(0.0454)	(0.7136)	(.0069)	(.0006)	(.0114)	(.1445)	(.0007)	(.0980)	
BM13	0.0463	0.0046	3.7041	18.4162	-	-	-	.0044	.9962	.0190	220.54
	(0.0006)	(0.0002)	(0.1603)	(0.3480)	-	-	-	(.0152)	(.0018)	(.0654)	
MRCE	0.0856	0.0128	11.4955	47.3913	.0227	**.9995**	.8614	**.7800**	.2495	.0156	3.70
	(0.0021)	(0.0017)	(0.5913)	(1.2672)	(.0280)	(.0024)	(.1810)	(.2386)	(.2473)	(.0023)	
CAPME	0.0072	0.0048	1.6072	9.0566	**.9893**	.8195	.2250	0	**1**	-[1]	81.38
	(0.0007)	(0.0006)	(0.0659)	(0.4222)	(.0050)	(.0168)	(.0166)	(0)	(0)	-	
	Simulation 4: $p = 120$, $q = 50$, $n = 100$, Coefficients, AR1 structure										
	MSE			Divergence	B support recovery			Ω support recovery			CPU time
Method	B	Ω	Prediction	avg KL	SEN	SPE	PRC	SEN	SPE	PRC	min.
HS-GHS	**0.0014**	**0.0044**	**2.3994**	**7.5001**	**1**	**.9987**	**.9757**	.9902	.9995	**.9880**	2.56e+03
	(0.0001)	(0.0010)	(0.0997)	(0.6304)	(0)	(.0005)	(.0098)	(.0125)	(.0007)	(.0165)	
BM13	0.6005	0.0253	34.9632	54.9328	-	-	-	0	.9919	0	217.87
	(0.0084)	(<0.0001)	(2.0061)	(0.3322)	-	-	-	(0)	(.0019)	(0)	
MRCE	1.2349	0.0259	176.9700	95.2552	.0207	.9984	.4862 [2]	**.9906**	.0105	.0400	9.76
	(0.0113)	(0.0003)	(11.3600)	(0.9369)	(.0167)	(.0024)	(.2114)	(.0211)	(.0169)	(.0005)	
CAPME	0.0178	0.0188	3.8546	23.8957	**1**	.8206	.2288	.0184	**.9999**	.8491 [3]	77.02
	(0.0022)	(0.0011)	(0.2719)	(1.1530)	(0)	(.0200)	(.0227)	(.1299)	(.0010)	-	

1. 50 NaNs. 2. 1 NaN in 50 replicates. 3. 49 NaNs in 50 replicates. Mean and sd. calculated on non-NaN values.

C.4 Assessment of normality assumption for eQTL analysis

Figure C.2 shows normal qq-plots of residual gene expression in 54 MAPK pathway genes. The expressions were regressed on the 172 markers using lasso regression, and residuals were calculated. Residuals of PKC1, MFA1, SWI6, MFA2 and SSK2 failed univariate Kolmogorov-Smirnov normality test at significance level 0.05, and these genes were removed from the data set for analysis. Yeast segregants shown as red and orange squares were removed from the data set for analysis.




GPA1
SLT2
STE12
SLN1
FAR1
BCK1
STE18
STE3
MSN4
MLP1
SSK1
HOG1
SWI6
CDC42
MID2
FKS1
STE11
GLO1
MSN2
MCM1
WSC2
BNI1
MFA2
RAS2
MSG5
-2
-1
0
1
2

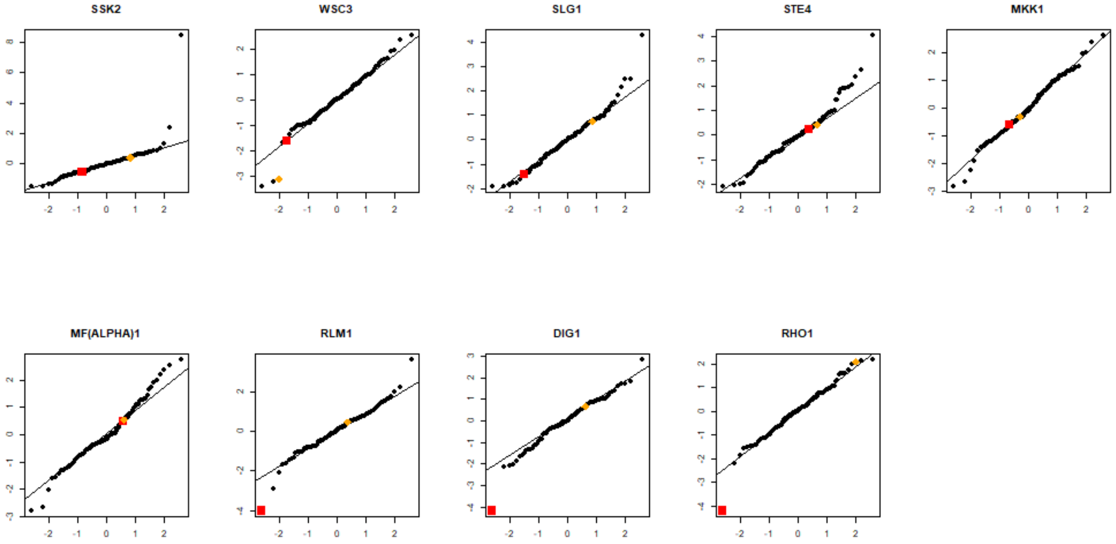
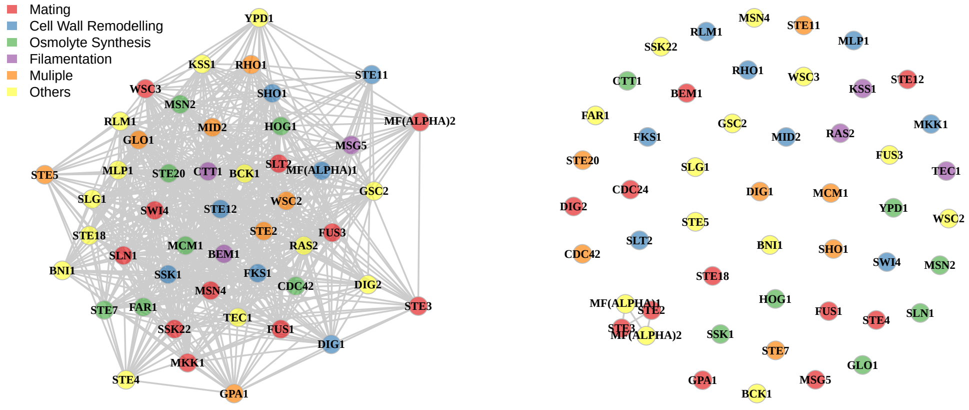


Figure C.2.: Normal q-q plots of gene expressions.

C.5 Additional eQTL analysis results

Figure C.3 shows the inferred graphs by CAPME and MRCE estimates, complementing the result presented in Figure 4.2 for the proposed HS-GHS estimate.



(a) Inferred graph by CAPME estimate

(b) Inferred graph by MRCE estimate

Figure C.3.: The inferred graph for the yeast eQTL data, estimated by (a) CAPME, and (b) MRCE.