Journal of Multivariate Analysis 183 (2021) 104716

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Joint mean–covariance estimation via the horseshoe

Yunfan Li [a], Jyotishka Datta [b], Bruce A. Craig [a], Anindya Bhadra [a,*]

[a] Department of Statistics, Purdue University, West Lafayette, IN 47907, USA
[b] Department of Mathematical Sciences, University of Arkansas, Fayetteville, AR 72704, USA

ARTICLE INFO

Article history:
Received 20 December 2019
Received in revised form 15 December 2020
Accepted 15 December 2020
Available online 29 December 2020

AMS 2010 subject classifications:
primary 62F15
secondary 62H12

Keywords:
Bayesian methods
eQTL analysis
Global–local priors
Seemingly unrelated regression
Shrinkage estimation

ABSTRACT

Seemingly unrelated regression is a natural framework for regressing multiple correlated responses on multiple predictors. The model is very flexible, with multiple linear regression and covariance selection models being special cases. However, its practical deployment in genomic data analysis under a Bayesian framework is limited due to both statistical and computational challenges. The statistical challenge is that one needs to infer both the mean vector and the inverse covariance matrix, a problem inherently more complex than separately estimating each. The computational challenge is due to the dimensionality of the parameter space that routinely exceeds the sample size. We propose the use of horseshoe priors on both the mean vector and the inverse covariance matrix. This prior has demonstrated excellent performance when estimating a mean vector or inverse covariance matrix separately. The current work shows these advantages are also present when addressing both simultaneously. A full Bayesian treatment is proposed, with a sampling algorithm that is linear in the number of predictors. MATLAB code implementing the algorithm is freely available from github at https://github.com/liyf1988/HS_GHS. 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

Multiple predictors–multiple responses regression, sometimes also known as the problem of multi-task learning in machine learning literature, is a common modeling framework in quantitative disciplines as diverse as finance, chemometrics and genomics. To take a concrete example from the field of genomics, this problem arises in simultaneously regressing the expression levels of multiple genes on multiple markers or regions of genetic variation, which is known as an expression quantitative trait loci (eQTL) analysis. Early studies have shown that each gene expression level is expected to be affected by only a few genomic regions [8,28] 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 that encodes conditional independence relationships [20], which, in the case of a multivariate Gaussian model, are encoded by the off-diagonal zeros in the inverse covariance matrix. Therefore, an eQTL analysis, if formulated as a multiple predictors–multiple responses regression problem, presents with non-independent error terms. In high dimensions, this necessitates regularized estimates of both the regression coefficients and the error inverse covariance matrix. Similar problems arise in econometrics, for example, in predicting the set of several correlated stock prices using a common set of covariates.

* Corresponding author.
E-mail address: bhadra@purdue.edu (A. Bhadra).

https://doi.org/10.1016/j.jmva.2020.104716
0047-259X/

A natural question then is: what is there to be gained by treating all responses jointly rather than separately regressing each response on the set of covariates, possibly adjusting for multiplicity in the responses? In multivariate regression problems with correlated error terms, early works by Zellner [31] established that joint estimation of regression coefficients improves efficiency. Zellner [31] went on to propose the seemingly unrelated regression 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. [18] adopted the seemingly unrelated regression framework in Bayesian regressions. However, these early methods in the seemingly unrelated regression framework considered a relatively modest dimension of the responses, and did not encourage sparse estimates of either the regression coefficients or the error inverse covariance matrix. Therefore, these methods cannot be applied directly to analyze modern genomic or financial data. Much more recently, both Bayesian and frequentist approaches that encourage sparsity have started to attract considerable attention in a seemingly unrelated regression framework [e.g.,1,6,9,29,30]. Precise descriptions of some of these competing approaches and understanding their strengths and limitations require some mathematical formalism. This is reserved for Section 2.

In this article, we propose a fully Bayesian method for high-dimensional seemingly unrelated regression problems with an algorithm for efficient exploration of the posterior. We impose the horseshoe prior [11] on the regression coefficients, and the graphical horseshoe prior [21] 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 [11], asymptotic optimality in testing under 0–1 loss [14], minimaxity in estimation under the ℓ_2 loss [24], and improved risk properties in linear regression [4]. The graphical horseshoe prior inherits the properties of improved Kullback–Leibler risk bounds, and nearly unbiased estimates, when applied to precision matrix estimation [21].

The beneficial theoretical and computational properties of the horseshoe and graphical horseshoe are combined in our proposed method, resulting in a prior that we term the horseshoe-graphical horseshoe or 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 [25]. 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 with a linear scaling in the number of covariates that allows arbitrary sparsity patterns in both the regression coefficients and the error precision matrix. This is at a contrast with existing Bayesian methods that require far more restrictive assumptions on the nature of associations. For example, Bhadra and Mallick [6] require that either a predictor is important to all the responses, or to none of them. The proposed method is also at a contrast with approaches that require special structures on the conditional independence relationships. For example, both Bhadra and Mallick [6] and Banterle et al. [1] require that the graphical model underlying the inverse covariance matrix is decomposable. Such assumptions are typically made for computational convenience, rather than any inherent problem-specific motivation, and the current work delineates a path forward by dispensing with them. In addition to these methodological innovations, the performance of the proposed method is compared with several competing approaches in a yeast eQTL data set and superior performances in both estimation and prediction are demonstrated.

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 $E \sim \mathrm{MN}_{n\times q}(0, I_n, \Omega_{q\times q}^{-1})$ denotes a matrix normal random variate [15] with the property that $\mathrm{vec}(E^\top) \sim \mathrm{N}_{nq}(0, I_n \otimes \Omega_{q\times q}^{-1})$, a multivariate normal, where $\mathrm{vec}(A)$ 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 network structure, which is reasonable for an eQTL analysis. The problem is then to estimate $B_{p\times q}$ and $\Omega_{q\times q}$. 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 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 $Y_i \sim \mathrm{N}(X_iB, \Omega^{-1})$, for $i \in \{1, \ldots, n\}$. The negative log likelihood function under this model, up to a constant, is

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

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

One of the earliest works in high dimensions is the multivariate regression with covariance estimation or the MRCE method [27], which adds independent ℓ_1 penalties to B and Ω, so the objective function is

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

where $\omega_{k\ell}$ are the elements of Ω, β_j are the elements of vectorized $B^\top$, and $\lambda_1, \lambda_2 > 0$ are tuning parameters. A coordinate descent algorithm is developed that iteratively solves a lasso and a graphical lasso problem to update $\hat{B}_{MRCE}$ and $\hat{\Omega}_{MRCE}$, respectively.

Cai et al. [9] developed the covariate-adjusted precision matrix estimation or CAPME procedure taking a two-stage approach and using a multivariate extension of the Dantzig selector of Candes and Tao [10]. Let $\bar{y} = n^{-1}\Sigma_{i=1}^n y_i$, $\bar{x} = n^{-1}\Sigma_{i=1}^n x_i$, $S_{xy} = n^{-1}\Sigma_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})^\top$ and $S_{xx} = n^{-1}\Sigma_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^\top$. The estimate of B in CAPME solves the optimization problem

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

where λ_n is a tuning parameter, $|A|_1$ defines the element-wise ℓ_1 norm of matrix A, and $|A|_\infty$ defines the element-wise 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 $S_{yy} = n^{-1}\Sigma_{i=1}^n (y_i - \hat{B}x_i)(y_i - \hat{B}x_i)^\top$, one estimates Ω by solving the optimization problem

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

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 [11], and at least empirically, the Dantzig selector rarely outperforms the lasso in simulations and in genomic data sets [23,32], indicating these problems might be inherited by MRCE and CAPME, respectively.

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. [17] put spike-and-slab lasso priors on the elements of B. That is, $\beta_{kj}; k \in \{1, \ldots, p\}, j \in \{1, \ldots, q\}$ 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, and the binary indicator γ_{kj} follows a priori a 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 $\omega_{\ell m}$ 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 B, and the corresponding four hyper-parameters for Ω, need to be specified in order to apply the ECM algorithm. In Deshpande et al. [17], 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 are set at predefined levels. The method does not provide samples from the full posterior.

Bhadra and Mallick [6] also consider a spike-and-slab prior on B but place Bernoulli indicators in a different way. Their priors on B and Ω^{-1} are

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

where b, c, d are fixed, positive hyper-parameters and HIW denotes the hyper-inverse Wishart distribution [16]. The vector of indicators γ selects entire rows of coefficients, depending on whether $\gamma_i = 1;\ i \in \{1, \ldots, p\}$. Similarly, the indicator G has length $q(q-1)/2$, and selects the off-diagonal elements in the precision matrix. Here $p_\gamma = \sum_{i=1}^p \gamma_i$. 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 Markov chain Monte Carlo (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 while allowing for uncertainty quantification using the full posterior. To this end, we propose to use the horseshoe prior that achieves efficient shrinkage in both sparse regression and inverse covariance estimation. We also develop an MCMC algorithm for sampling, without user-chosen tuning parameters.

3. Proposed model and estimation algorithm

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

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

where $C^+(0, 1)$ denotes the standard half-Cauchy distribution with density $p(x) \propto (1 + x^2)^{-1};\ x > 0$. This normal scale mixture on β with half-Cauchy hyperpriors on λ_j and τ is known as the horseshoe prior [11], presumably due to the

Algorithm 1 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 $\tilde{y} = \text{vec}(\Omega^{1/2}Y^\top)$, $\tilde{X} = X \otimes \Omega^{1/2}$
      %% Sample β using the horseshoe
    (2a) Sample $u \sim \text{N}_{pq}(0, \Lambda_*)$ and $\delta \sim \text{N}_{nq}(0, I_{nq})$ independently, where $\Lambda_* = \text{diag}(\lambda_j^2\tau^2)$
    (2b) Take $v = \tilde{X}u + \delta$
    (2c) Solve w from $(\tilde{X}\Lambda_*\tilde{X}^\top + I_{nq})w = \tilde{y} - v$
    (2d) Calculate $\beta = u + \Lambda_*\tilde{X}^\top w$
    (3) Sample $\lambda_j^2 \sim \text{InvGamma}(1, 1/\nu_j + \beta_j^2/(2\tau^2))$, and $\nu_j \sim \text{InvGamma}(1, 1 + 1/\lambda_j^2)$, for $j \in \{1, ..., pq\}$
    (4) Sample $\tau^2 \sim \text{InvGamma}((pq+1)/2, 1/\xi + \Sigma_{j=1}^{pq}\beta_j^2/(2\lambda_j^2))$, and $\xi \sim \text{InvGamma}(1, 1 + 1/\tau^2)$
    (5) Calculate $Y_{res} = Y - XB$ and $S = Y_{res}^\top Y_{res}$
      %% Sample Ω using the graphical horseshoe
    for $k = 1$ to q **do**
      Partition matrices Ω, 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 ω_{kk}, s_{kk}
      (6a) Sample $\gamma \sim \text{Gamma}(n/2 + 1, 2/s_{kk})$
      (6b) Sample $\upsilon \sim \text{N}(-Cs_{(-k)k}, C)$ where $C = (s_{kk}\Omega_{(-k)(-k)}^{-1} + \text{diag}(\eta_{(-k)k}\zeta^2)^{-1})^{-1}$ and $\eta_{(-k)k}$ is a vector of length $(q-1)$ with entries $\eta_{\ell k}^2$, $\ell \neq k$
      (6c) Apply transformation: $\omega_{(-k)k} = \upsilon$, $\omega_{kk} = \gamma + \upsilon^\top\Omega_{(-k)(-k)}^{-1}\upsilon$
      (7) Sample $\eta_{(-k)k} \sim \text{InvGamma}(1, 1/\rho_{(-k)k} + \omega_{(-k)k}^2/2\zeta^2)$, and $\rho_{(-k)k} \sim \text{InvGamma}(1, 1 + 1/\eta_{(-k)k})$
    end for
    (8) Sample $\zeta^2 \sim \text{InvGamma}(\binom{q}{2} + 1)/2, 1/\phi + \sum_{k,\ell:k<\ell}\omega_{k\ell}^2/2\eta_{k\ell}^2)$, and $\phi \sim \text{InvGamma}(1, 1 + 1/\zeta^2)$
    Save samples if $i >$ *burnin*
  end for
  Return MCMC samples of β and Ω
end function

shape of the induced prior on the shrinkage factor. The key motivation for this hierarchical form is that the global term τ encourages sparsity of the estimates by typically settling on a small value a posteriori, but the local heavy-tailed λ_j terms prevent the large signals from being over-shrunk. Similarly, to encourage sparsity in the off-diagonal elements of Ω, we use the graphical horseshoe (GHS) prior for Gaussian graphical models [21], i.e.,

$$\omega_{k\ell:k>\ell} \sim \text{N}(0, \eta_{k\ell}^2\zeta^2);\ \ k, \ell \in \{1, \ldots, q\},\ \ \eta_{k\ell} \sim C^+(0, 1),\ \ \zeta \sim C^+(0, 1),\ \ p(\omega_{kk}) \propto 1,$$

where $\Omega = \{\omega_{k\ell}\}$, and the prior mass is truncated to the space of $q \times q$ positive definite matrices $\mathcal{S}_q^+$. In this model, $\eta_{k\ell}$ and ζ induce shrinkage on the off-diagonal elements in Ω. The joint prior on $\Theta = (B, \Omega)$ is termed the HS-GHS prior.

MCMC samplers have been proposed for regressions using the horseshoe prior for the linear regression model with i.i.d. error terms [7,22]. However, these samplers cannot be applied to the current problem due to the correlation structure in the error. To transform the data into a model where sampling is possible, we reshape the predictors and responses. Let $\tilde{y} = \text{vec}(\Omega^{1/2}Y^\top)$, and $\tilde{X} = X \otimes \Omega^{1/2}$. Simple algebra shows that $\tilde{y} \sim \text{N}_{nq}(\tilde{X}\beta, I_{nq})$. In this way, the matrix variate normal regression problem is transformed into a multivariate normal regression problem, provided the current estimate of Ω is known. Next, given the current estimate of B, the graphical horseshoe sampler of Li et al. [21] is leveraged to estimate Ω.

A full Gibbs sampler for the above model is given in Algorithm 1. 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 notice that the conditional posterior of β is $\text{N}((\tilde{X}^\top\tilde{X} + \Lambda_*^{-1})^{-1}\tilde{X}^\top\tilde{Y}, (\tilde{X}^\top\tilde{X} + \Lambda_*^{-1})^{-1})$, where $\Lambda_* = \text{diag}(\lambda_j^2\tau^2), j \in \{1, \ldots, pq\}$. However, sampling from this normal distribution is computationally expensive because it involves computing the inverse of the $pq \times pq$ dimensional matrix $(\tilde{X}^\top\tilde{X} + \Lambda_*^{-1})$, 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. [7]. The algorithm is exact with a complexity linear in p.

To sample the precision matrix Ω conditional on B, define the residual $Y_{res} = Y - XB$, and let $S = Y_{res}^\top Y_{res}$. Since $(Y - XB) \mid \Omega \sim \text{MN}(0,\ I_n,\ \Omega^{-1})$, the problem of estimating Ω given B is exactly the zero-mean multivariate Gaussian inverse covariance estimation that the graphical horseshoe [21] solves. A detailed derivation of Algorithm 1 is given in Appendix A and 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 [9,27,30]. Second, MCMC approximation of the posterior distribution enables variable selection using posterior credible intervals. By varying the length of credible intervals, it is also possible 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 method possesses attractive theoretical properties as well.

4. Kullback–Leibler risk 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 [2] formalizes this statement. Let $D_n(\theta) = \frac{1}{n} D(p^n \| q^n(\cdot|\theta))$, where $D(\pi_1\|\pi_2) = \int \log(\pi_1/\pi_2) d\pi_1$, denotes 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 A_ϵ.

Lemma 4.1 ([2]). *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 $Y_1, \ldots, Y_n$ 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 an ϵ neighborhood of zero compared to any other prior with a bounded density at zero.

Theorem 4.2. *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 $\Omega_0 \in \mathcal{S}_q^+$ are the true regression coefficients and inverse covariance, respectively and X_i are observed covariates. Let β_{j0}, $\omega_{k\ell 0}$ and $\sigma_{k\ell 0}$ denote the jth and $k\ell$th element of $\mathrm{vec}(B_0)$, Ω_0 and $\Sigma_0 = \Omega_0^{-1}$, respectively. Suppose that $\sum_{k,\ell} \omega_{k\ell 0} \propto q$, $\sum_{k,\ell} \sigma_{k\ell 0} \propto q$, and $\sum_{i=1}^n (X_{i1} + \cdots + X_{ip})^2 \propto np^2$. Suppose that an Euclidean cube in the neighborhood of Ω_0 with $(\omega_{k\ell 0} - 2/Mn^{1/2}q, \omega_{k\ell 0} + 2/Mn^{1/2}q)$ on each dimension lies in the cone of positive definite matrices $\mathcal{S}_q^+$, where $M = \sum_{k,\ell} \sigma_{k\ell 0}/q$. 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:*

(i) *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.*

(ii) *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 4.2 is in Appendix B. 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 element-wise prior satisfying conditions in Part (1) of Theorem 4.2 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 4.2 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 [11], 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 $\omega_{k\ell 0} = 0$, 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 5.

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 consider 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 off-diagonal elements:

1. AR1. The precision matrix has an AR1 structure, with $\omega_{k,\ell} = 0.45$ for $|k - \ell| = 1$ and zero otherwise for $k \neq \ell$.

2. Cliques. The rows/columns are partitioned into disjoint groups and $\omega_{k\ell:k,\ell\in G,\ k\neq\ell}$ 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. It is important to note although these settings are used for the simulation examples, the proposed method allows arbitrary sparsity patterns in both B and Ω and is in no way dependent on these specific settings.

We generate $n \times p$ design matrix X with a Toeplitz covariance structure where $\mathrm{Cov}(X_i, X_j) = 0.7^{|i-j|}$, and $n \times q$ 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 [27], CAPME [9] and the joint high-dimensional Bayesian variable and covariance selection (BM13) [6]. The proposed HS-GHS estimator is implemented in MATLAB. The MATLAB code by Bhadra and Mallick [6] is used for BM13, and R packages 'MRCE' and 'capme' are used for MRCE and CAPME respectively. Mean squared estimation errors of the regression coefficients and the 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 is performed using the middle 75% posterior credible interval. Following Bhadra and Mallick [6], variables with posterior probability of inclusion larger than 0.5 are considered to be selected by BM13. In case the choices of these thresholds appear somewhat arbitrary, we also present receiver operating characteristic (ROC) curves for all methods to compare their overall variable selection performances as the decision threshold is varied between the two extremities, i.e., where all variables are selected and where none are selected.

Results are reported in Tables 1 and 2, along with CPU times for all methods. It is evident that the HS-GHS has the best overall statistical performance. Except for the mean squared error of Ω when $p = 200$, the HS-GHS has the best estimation, prediction, information divergence and variable selection performances in our simulations. Although the 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 the 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 genomic 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.

Fig. 1 shows the ROC curves for both B and Ω, when $p = 120$ and $q = 50$. True and false positive rates are generated by varying the width of posterior credible intervals from 0% to 100% in HS-GHS, and varying the posterior inclusion probability from 0% to 100% in BM13. In MRCE and CAPME, variables are selected by thresholding the estimated B and Ω. For each estimated β_j and $\omega_{k\ell}$, the element is considered to be a positive if its absolute value is larger than a threshold, and the threshold is varied to generate a series of variable selection results. In all four plots, the HS-GHS curves closely follow the line where the true positive rate equals one, suggesting that the credible intervals for the true nonzero parameters do not include zero. These results are consistent with the theoretical findings that horseshoe credible intervals have optimal size [25]. CAPME has the second best performance in variable selection, except when it does not generate valid ROC plots. For example, in the cliques structured precision matrix estimated by CAPME, all off-diagonal elements are estimated to be zero, so CAPME cannot generate an ROC curve in this case. Moreover, neither MRCE nor BM13 produces satisfactory ROC curves. MCMC convergence diagnostics of the HS-GHS sampler are presented in Supplementary Section S.1. Further simulation results complementing the results in this section are in Supplementary

Table 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.

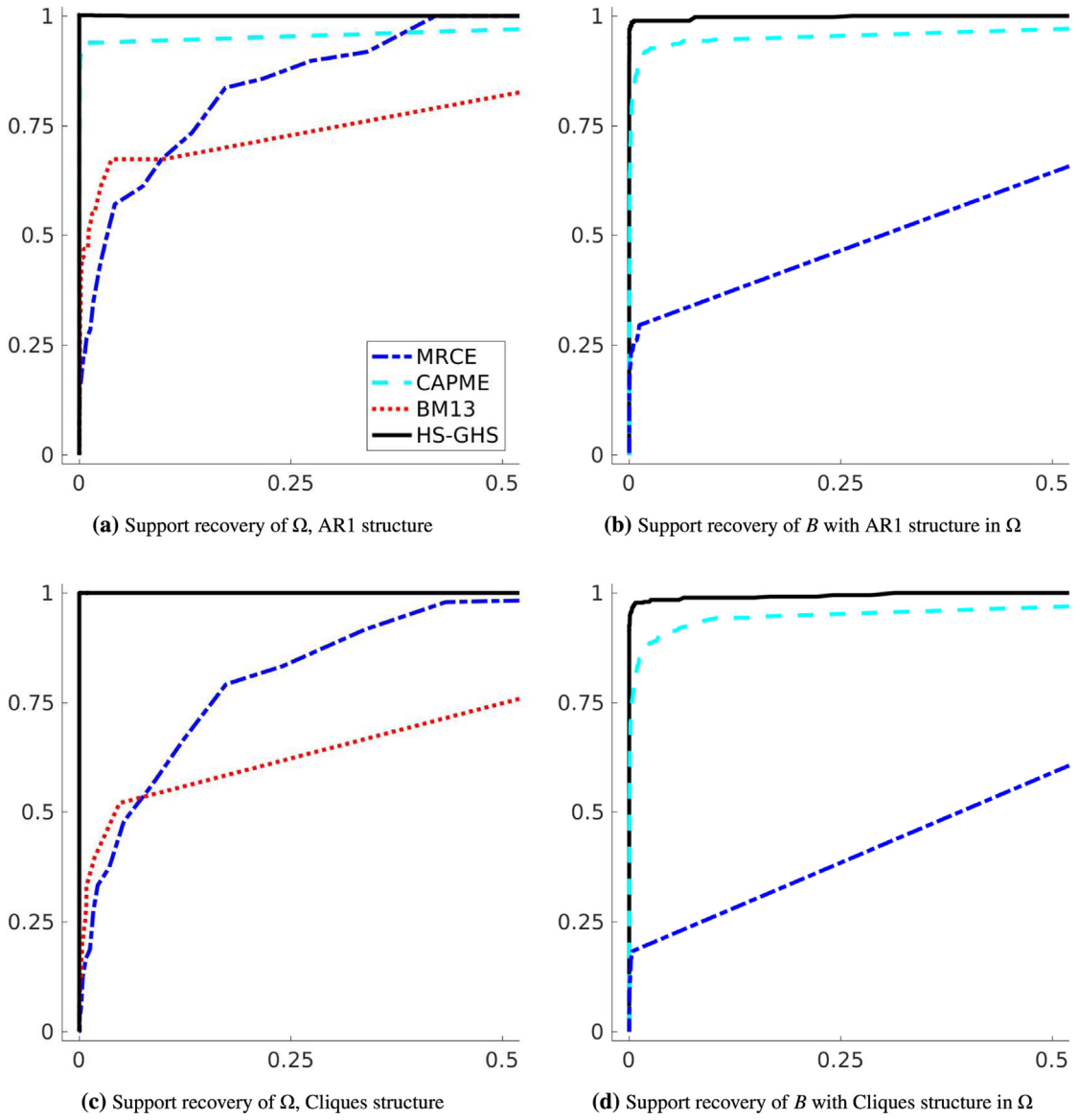
	Simulation 1: $p = 200$, $q = 25$, $n = 100$, Uniform coefficients, AR1 structure										
	MSE			Divergence	B support recovery			Ω support recovery			CPU time
Method	B	Ω	Prediction	avg KL	SEN	SPE	PRC	SEN	SPE	PRC	min.
HS-GHS	**0.0033**	0.0365	**2.6352**	**10.2075**	.9380	.9981	**.9621**	**.9658**	.9973	**.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[a]	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	–[b]	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[c]	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	–[d]	73.67
	(0.0016)	(0.0007)	(0.2930)	(0.3797)	(.0155)	(.0215)	(.0198)	(0)	(0)	–	

[a] 16 NaNs in 50 replicates.
[b] 50 NaNs. All mean and sd. calculated on non-NaN values.
[c] 23 NaNs in 50 replicates.
[d] 50 NaNs. All mean and sd. calculated on non-NaN values.

Table 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[a]	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	–[b]	81.99
	(0.0013)	(0.0004)	(0.1421)	(0.4293)	(.0122)	(.0234)	(.0251)	(0)	(0)	–	

[a] 1 NaN in 50 replicates.
[b] 50 NaNs. Mean and sd. calculated on non-NaN values.



(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

Fig. 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 rates are shown on the y-axis, and the false positive rates are shown on the x-axis.

Section S.2. Specifically, Section S.2.1 provides performance comparisons with the method of Deshpande et al. [17] for the settings of Tables 1 and 2, demonstrating superior performance for HS-GHS for most settings; Sections S.2.2 and S.2.3 provide evidence that simultaneous mean and covariance estimation indeed results in improved prediction and estimation performances compared to q separate regressions; and Section S.2.4 complements the results in Tables 1 and 2 by considering additional simulation settings.

6. Yeast eQTL data analysis

We illustrate the HS-GHS method using the yeast eQTL data analyzed by Brem and Kruglyak [8]. 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 is available in the R package `trigger` (https://www.bioconductor.org/packages/release/bioc/html/trigger.html). 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 consider 54 genes

Table 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-GHS	Gene	CAPME	MRCE	HS-GHS
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**

in the yeast mitogen-activated protein kinase (MAPK) signaling pathway in our analysis. This pathway was provided by the Kyoto Encyclopedia of Genes and Genomes database [19], and was also analyzed by Yin and Li [30] and Cai et al. [9].

Following the method described in Curtis et al. [13], we divide the genome into 316 groups based on linkage disequilibrium between the markers, and select the marker with the 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. Marginal qq-plots of residuals and other assessments of normality assumption are provided in Supplementary Section S.3. The final data set we use in our analysis contains 49 genes in the MAPK pathway and 172 markers in 110 yeast segregants.

We divide 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% posterior credible interval. Prediction and estimation results are summarized in Tables 3 and 4, respectively.

Out of 8428 regression 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-square 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 consider gene expressions that have R-square larger than 0.1 in any of these three models. Among 22 such gene expressions, CAPME has the highest R-square among the three methods in 4 gene expressions, and HS-GHS has the highest R-square in 18 gene expressions. Average prediction R-square values in these 22 genes by CAPME, MRCE and HS-GHS are 0.1327, 0.0063, 0.2771, respectively.

We also examine the 15 nonzero coefficients estimated by the HS-GHS. CAPME estimates eight of these 15 coefficients to be nonzero, and CAPME estimates have smaller absolute magnitudes than the 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. [33]. 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. [33]. All of these nonzero estimates correspond to expressions mapped far from the location of their gene of origin, and can be considered distant eQTLs. This highlights the need for a model to simultaneously accommodate expressions and markers on different genomic locations, rather than separate chromosome-specific eQTL analysis.

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 Fig. 2, while CAPME and MRCE results are in Supplementary Section S.4. 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 [12]. Fig. 2 recovers some known structures in the MAPK pathway. For instance, STE4, STE18, GPA1, STE20, CDC42, DIG1, BEM1, FUS1, STE2, STE3 and MSG5 are involved in the yeast mating process, and they are linked in the HS-GHS estimate. SLT2, SWI3, RHO1, RLM1 and MLP1 involved in the cell wall remodeling process, and YPD1, CTT1, GLO1 and SSK1 involved in the osmolyte synthesis process are also linked. It is also known that the high-osmolarity glycerol (HOG) and cell wall integrity (CWI) signaling pathways interact in yeast [26], and some genes in the HOG pathway are indeed connected to genes in the CWI pathway in the HS-GHS estimate.

Table 4
Nonzero coefficients in HS-GHS estimate, along with names and locations of the genes, locations of the markers, and CAPME estimated coefficients.

Gene	Chromosome	Within-chr. position	Marker chr.	Within-chr. marker position	HS-GHS coefficients	CAPME 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[a]
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

[a]MRCE estimate for this coefficient is 0.05 and MRCE estimates for all other coefficients in this table are 0. Thus, MRCE results are not separately presented.

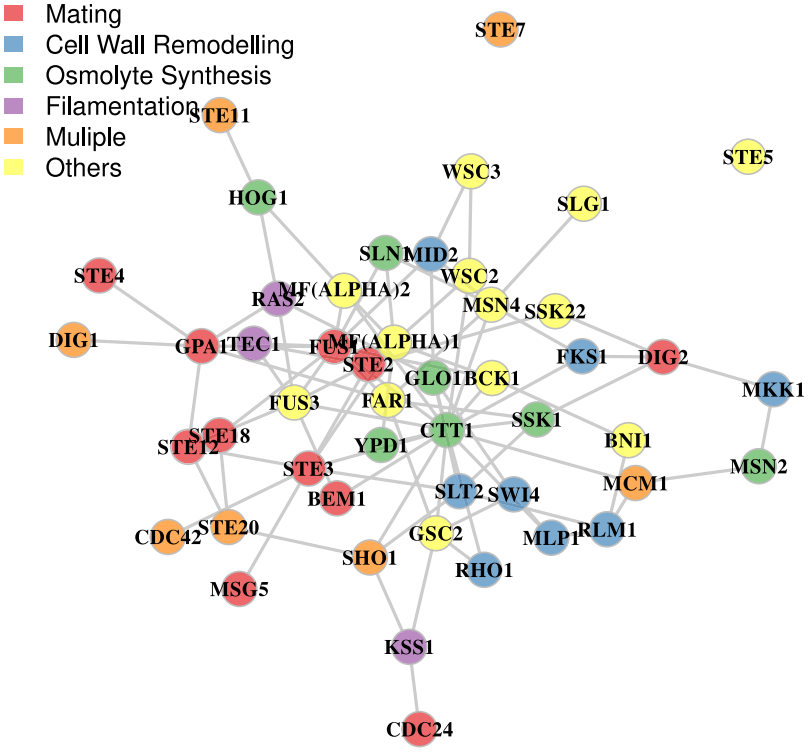


Fig. 2. The inferred graph for gene expressions in the MAPK pathway by the HS-GHS estimate. Vertex colors indicate functions of genes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

7. Conclusions

The horseshoe estimator 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 the regression coefficients and the error inverse covariance simultaneously in multiple predictors–multiple responses regressions. The proposed method allows arbitrary sparsity patterns in both 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. This fact needs to be accounted for before a naïve comparison of the scalability of SUR approaches with marginal correlation based methods for separate regression analysis. With $q = 1$, the latter approaches may scale to larger values

of p, but cannot utilize the error correlation structure as the SUR models do, consequently resulting in less statistically efficient estimates. HS-GHS also recovers the support of the regression coefficients and inverse covariance with higher precision compared to other SUR model based approaches, such as MRCE, CAPME and BM13 and outperforms these alternatives in terms of both estimation and prediction.

Computationally, the proposed sampler is the first in an SUR setting with a complexity linear in p. A major advantage of our method is samples are available from the full posterior, thereby allowing straightforward uncertainty quantification. If posterior draws are not required, it is possible to develop faster point estimation algorithms. Prominent among these possibilities is an iterated conditional modes (ICM) algorithm [3] to obtain the maximum pseudo posterior estimate. At each iteration, ICM maximizes the full conditional posteriors of all variables and converges to a deterministic solution. Since the full conditionals in the HS-GHS model are either normal, gamma or inverse gamma, the conditional modes are unique, and ICM should be easy to implement. It is also possible to include domain knowledge in designing the priors, such as pathway information for the genomic application, by coupling the local shrinkage parameters a priori. This article focuses 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+ [5], should also be explored.

Acknowledgments

The authors are grateful to two anonymous referees for helpful comments. Datta is supported by Grant No. DMS-2015460 by the US National Science Foundation. Bhadra is supported by Grants No. DMS-1613063 and DMS-2014371 by the US National Science Foundation.

Appendix A. Derivation of Algorithm 1

- *Step 2:* Since $\tilde{y} \sim \mathrm{N}_{nq}(\tilde{X}\beta, I_{nq})$ and the prior on β is horseshoe, the full conditional posterior of β is $\mathrm{N}((\tilde{X}^\top\tilde{X} + \Lambda_*^{-1})^{-1}\tilde{X}^\top\tilde{y}, (\tilde{X}^\top\tilde{X} + \Lambda_*^{-1})^{-1})$, where $\Lambda_* = \mathrm{diag}(\lambda_j^2\tau^2), j \in \{1, \ldots, pq\}$. Sampling of β is exactly the problem solved by Bhattacharya et al. [7]. Realizing that β has length pq, $\tilde{y}$ has length nq, and substituting $\tilde{X}$, $\tilde{y}$, Λ_* and β into Steps 1 to 4 in *Algorithm 1* in Bhattacharya et al. [7], yield Steps (2a)–(2d).
- *Steps 3–4:* These steps concern sampling of the shrinkage parameters λ_j for $j \in \{1, \ldots, pq\}$, and τ. Both have half Cauchy priors, which can be written as a mixture of two inverse gamma random variables. Specifically, if $x^2 \mid a \sim \mathrm{InvGamma}(1/2, 1/a)$ and $a \sim \mathrm{InvGamma}(1/2, 1)$, then Makalic and Schmidt [22] demonstrated that marginally $x \sim \mathrm{C}^+(0, 1)$. Since an inverse gamma prior is conjugate to itself and to the variance parameter in a normal model, the full conditional posteriors of λ_j^2, τ^2 and the corresponding auxiliary variables ν_j and ξ are all inverse gamma random variables. This completes Steps 3 and 4 in our Algorithm 1.
- *Steps 6–8:* Given B, if one defines $Y_{res} = Y - XB$, then sampling of Ω is the problem of sampling the precision matrix in a zero-mean multivariate normal model. Thus, Steps (6a)–(8) in Algorithm 1 follows the sampling scheme of the graphical horseshoe model for sample size n, number of features q, and scatter matrix $S = Y_{res}^\top Y_{res}$. Details for these steps can be found in Algorithm 1 of Li et al. [21].

Appendix B. Proof of Theorem 4.2

Let $A_\epsilon = \{\{B, \Omega\} : \frac{1}{n}D_n(p_{B_0,\Omega_0} \| p_{B,\Omega}) \le \epsilon\}$. 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 $(\omega_{k\ell 0} - k_2\epsilon^{1/2}/q, \omega_{k\ell 0} + k_2\epsilon^{1/2}/q)$ on each dimension. The proof is as following.

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

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

By the proof of Theorem 3.2 in Li et al. [21], $\mathrm{I} \propto n\epsilon$ when $\epsilon \to 0$. We will show that $\mathrm{II} \propto n\epsilon$ as well. The expression for II is simplified as,

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

Some algebra shows that $\text{vec}(X\mathbb{1}_{p\times q})^\top(\Omega_0\otimes I_n)\text{vec}(X\mathbb{1}_{p\times q}) = \sum_{k,\ell}\omega_{k\ell 0}\sum_i(X_{i1}+\cdots+X_{ip})^2$, and $\text{vec}(X\mathbb{1}_{p\times q})^\top(\mathbb{1}_{q\times q}\otimes I_n)\text{vec}(X\mathbb{1}_{p\times q}) = q^2\sum_i(X_{i1}+\cdots+X_{ip})^2$. Therefore,

$$\begin{aligned}\text{II} =& \frac{1}{2}\frac{\epsilon^{1/2}}{p^2q}\left\{\sum_{k,\ell}\omega_{k\ell 0}\sum_i(X_{i1}+\cdots+X_{ip})^2+\frac{\epsilon^{1/2}}{q}q^2\sum_i(X_{i1}+\cdots+X_{ip})^2\right\}\\ =& \frac{1}{2}\frac{\epsilon^{1/2}}{p^2q}(c_1np^2q+c_2\epsilon^{1/2}np^2q)=\frac{1}{2}(c_1n\epsilon^{1/2}+c_2n\epsilon).\end{aligned}$$

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, $\int_{\omega_{k\ell 0}-\epsilon^{1/2}/q}^{\omega_{k\ell 0}+\epsilon^{1/2}/q}p(\omega_{k\ell})d\omega_{k\ell}\propto\epsilon^{1/2}/q$, for any prior density $p(\omega_{k\ell})$ satisfying the conditions. Taking $\epsilon=1/n$, this gives $\log\nu(A_{1/n})$ in Part(1) of Theorem 4.2. The horseshoe prior also satisfies conditions in (1) in dimensions where $\beta_{j0}\neq 0$ and $\omega_{k\ell 0}\neq 0$, 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 $\omega_{k\ell 0}=0$. Using bounds of horseshoe prior provided in Carvalho et al. [11], it has been established by Li et al. [21] that $\int_0^{\epsilon^{1/2}/q}p(\omega_{k\ell})d\omega_{k\ell}>c_3\log(\epsilon^{-1/2}q)/(\epsilon^{-1/2}q)$. 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.

Appendix C. Supplementary material

The supplementary material contains MCMC diagnostics and additional simulation results, referenced in Section 5 and additional data analysis results, referenced in Section 6. Computer code is provided in a `.zip` archive to reproduce the simulation results in Section 5.

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jmva.2020.104716.

References

[1] M. Banterle, L. Bottolo, S. Richardson, M. Ala-Korpela, M.-R. Järvelin, A. Lewin, Sparse variable and covariance selection for high-dimensional seemingly unrelated Bayesian regression, bioRxiv, 2018, http://dx.doi.org/10.1101/467019.

[2] A.R. Barron, The Exponential Convergence of Posterior Probabilities with Implications for Bayes Estimators of Density Functions, Technical Report, Department of Statistics, University of Illinois, Champaign, IL, 1988.

[3] J. Besag, On the statistical analysis of dirty pictures, J. R. Stat. Soc. Ser. B Stat. Methodol. 48 (1986) 259–302.

[4] A. Bhadra, J. Datta, Y. Li, N.G. Polson, B. Willard, Prediction risk for the horseshoe regression, J. Mach. Learn. Res. 20 (2019) 1–39.

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

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

[7] A. Bhattacharya, A. Chakraborty, B.K. Mallick, Fast sampling with Gaussian scale mixture priors in high-dimensional regression, Biometrika 103 (2016) 985–991.

[8] R.B. Brem, L. Kruglyak, The landscape of genetic complexity across 5700 gene expression traits in yeast, Proc. Natl. Acad. Sci. USA 102 (2005) 1572–1577.

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

[10] E. Candes, T. Tao, The Dantzig selector: Statistical estimation when p is much larger than n, Ann. Statist. 35 (2007) 2313–2351.

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

[12] B. Conklin, M. Adriaens, T. Kelder, N. Salomonis, MAPK signaling pathway (*Saccharomyces cerevisiae*), 2018, https://www.wikipathways.org/index.php/Pathway:WP510, [Online; (Accessed 12 December 2018)].

[13] R.E. Curtis, S. Kim, J.L. Woolford Jr., W. Xu, E.P. Xing, 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 (2013) 196.

[14] J. Datta, J.K. Ghosh, Asymptotic properties of Bayes risk for the horseshoe prior, Bayesian Anal. 8 (2013) 111–132.

[15] A.P. Dawid, Some matrix-variate distribution theory: Notational considerations and a Bayesian application, Biometrika 68 (1981) 265–274.

[16] A.P. Dawid, S.L. Lauritzen, Hyper Markov laws in the statistical analysis of decomposable graphical models, Ann. Statist. 21 (1993) 1272–1317.

[17] S.K. Deshpande, V. Ročková, E.I. George, Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso, J. Comput. Graph. Statist. 28 (2019) 921–931.

[18] C.C. Holmes, D.G.T. Denison, B.K. Mallick, Accounting for model uncertainty in seemingly unrelated regressions, J. Comput. Graph. Statist. 11 (2002) 533–551.

[19] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, M. Hirakawa, KEGG for representation and analysis of molecular networks involving diseases and drugs, Nucleic Acids Res. 38 (2010) D355–D360.

[20] R.D. Leclerc, Survival of the sparsest: Robust gene networks are parsimonious, Mol. Syst. Biol. 4 (2008) 213.

[21] Y. Li, B.A. Craig, A. Bhadra, The graphical horseshoe estimator for inverse covariance matrices, J. Comput. Graph. Statist. 28 (2019) 747–757.

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

[23] N. Meinshausen, G. Rocha, B. Yu, Discussion: A tale of three cousins: Lasso, L2Boosting and Dantzig, Ann. Statist. 35 (2007) 2373–2384.

[24] S. van der Pas, B. Kleijn, A. van der Vaart, The horseshoe estimator: Posterior concentration around nearly black vectors, Electron. J. Stat. 8 (2014) 2585–2618.

[25] S. van der Pas, B. Szabó, A. van der Vaart, Uncertainty quantification for the horseshoe (with discussion), Bayesian Anal. 12 (2017) 1221–1274.
[26] J.M. Rodríguez-Peña, R. García, C. Nombela, J. Arroyo, The high-osmolarity glycerol (HOG) and cell wall integrity (CWI) signalling pathways interplay: A yeast dialogue between MAPK routes, Yeast 27 (2010) 495–502.
[27] A.J. Rothman, E. Levina, J. Zhu, Sparse multivariate regression with covariance estimation, J. Comput. Graph. Statist. 19 (2010) 947–962.
[28] E.E. Schadt, S.A. Monks, T.A. Drake, A.J. Lusis, N. Che, V. Colinayo, T.G. Ruff, S.B. Milligan, J.R. Lamb, G. Cavet, et al., Genetics of gene expression surveyed in maize, mouse and man, Nature 422 (2003) 297–302.
[29] A. Touloumis, J.C. Marioni, S. Tavaré, HDTD: Analyzing multi-tissue gene expression data, Bioinformatics 32 (2016) 2193–2195.
[30] J. Yin, H. Li, A sparse conditional Gaussian graphical model for analysis of genetical genomics data, Ann. Appl. Stat. 5 (2011) 2630–2650.
[31] A. Zellner, An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, J. Amer. Statist. Assoc. 57 (1962) 348–368.
[32] S. Zheng, W. Liu, An experimental comparison of gene selection by Lasso and Dantzig selector for cancer classification, Comput. Biol. Med. 41 (2011) 1033–1040.
[33] J. Zhu, B. Zhang, E.N. Smith, B. Drees, R.B. Brem, L. Kruglyak, R.E. Bumgarner, E.E. Schadt, Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks, Nature Genet. 40 (2008) 854–861.