Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Simultaneous Variable and Covariance Selection With the Multivariate Spike-and-Slab LASSO

Sameer K. Deshpande, Veronika Ročková & Edward I. George

To cite this article: Sameer K. Deshpande, Veronika Ročková & Edward I. George (2019): Simultaneous Variable and Covariance Selection With the Multivariate Spike-and-Slab LASSO, Journal of Computational and Graphical Statistics, DOI: 10.1080/10618600.2019.1593179

To link to this article: https://doi.org/10.1080/10618600.2019.1593179

View supplementary material

Accepted author version posted online: 20 Mar 2019.
Published online: 17 May 2019.

Submit your article to this journal

Article views: 122

View Crossmark data

Citing articles: 1 View citing articles

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
https://doi.org/10.1080/10618600.2019.1593179

Taylor & Francis
Taylor & Francis Group

Simultaneous Variable and Covariance Selection With the Multivariate Spike-and-Slab LASSO

Sameer K. Deshpande[a], Veronika Ročková[b], and Edward I. George[c]

[a]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA; [b]Department of Econometrics and Statistics at Booth School of Business, University of Chicago, Chicago, IL; [c]Department of Statistics, University of Pennsylvania, Philadelphia, PA

ABSTRACT
We propose a Bayesian procedure for simultaneous variable and covariance selection using continuous spike-and-slab priors in multivariate linear regression models where q possibly correlated responses are regressed onto p predictors. Rather than relying on a stochastic search through the high-dimensional model space, we develop an ECM algorithm similar to the EMVS procedure of Ročková and George targeting modal estimates of the matrix of regression coefficients and residual precision matrix. Varying the scale of the continuous spike densities facilitates dynamic posterior exploration and allows us to filter out negligible regression coefficients and partial covariances gradually. Our method is seen to substantially outperform regularization competitors on simulated data. We demonstrate our method with a re-examination of data from a recent observational study of the effect of playing high school football on several later-life cognition, psychological, and socio-economic outcomes. An R package, scripts for replicating examples in this article, and results from further simulation studies are provided in the supplementary materials available online.

ARTICLE HISTORY
Received August 2017
Revised February 2019

KEYWORDS
Bayesian shrinkage; EM algorithm; Gaussian graphical modeling; Multivariate regression; Nonconvex optimization

1. Introduction

We consider the multivariate Gaussian linear regression model, in which one simultaneously regresses $q > 1$ possibly correlated responses onto a common set of p covariates. In this setting, one observes n independent pairs of data $(\mathbf{x}_i, \mathbf{y}_i)$ where $\mathbf{y}_i \in \mathbb{R}^q$ contains the q outcomes and $\mathbf{x}_i \in \mathbb{R}^p$ contains measurements of the covariates. One then models $\mathbf{y}_i = \mathbf{x}_i'B + \varepsilon_i$, with $\varepsilon_1, \ldots, \varepsilon_n \sim \mathrm{N}\left(\mathbf{0}_q, \Omega^{-1}\right)$, independently, where $B = \left(\beta_{j,k}\right)_{j,k}$ and $\Omega = \left(\omega_{k,k'}\right)_{k,k'}$ are unknown $p \times q$ and $q \times q$ matrices, respectively. The main thrust of this article is to propose a new methodology for the simultaneous identification of the regression coefficient matrix B and the residual precision matrix Ω. Our framework additionally includes estimation of B when Ω is known and estimation of Ω when B is known as important special cases.

The identification and estimation of a sparse set of regression coefficients have been extensively explored in the univariate linear regression model, often through a penalized likelihood framework (see, e.g., Tibshirani 1996; Zou 2006; Fan and Li 2001; Zhang 2010). When moving to the multivariate setting, it is very tempting to deploy one's favorite univariate procedure to each of the q responses separately, thereby assembling an estimate of B column-by-column. Such an approach fails to account for the correlations between responses and may lead to poor predictive performance (see, e.g., Breiman and Friedman 1997). In many applied settings one may reasonably believe that some groups of covariates are simultaneously "relevant" to many responses. A response-by-response approach to variable selection fails to investigate or leverage such structural assumptions. While block-structured regularization approaches like Turlach, Venables, and Wright (2005), Obozinski, Wainwright, and Jordan (2011), and Peng et al. (2010) account for these assumptions, they do not explicitly model the residual correlation structure, essentially assuming $\Omega = I$.

Estimation of a sparse precision matrix and Gaussian graphical model G from multivariate Gaussian data has a similarly rich history, dating back to Dempster (1972), who coined the phrase *covariance selection* to describe this problem. The vertices of the graph G correspond to the coordinates of the multivariate Gaussian vector and an edge between vertices k and k' signifies that the corresponding coordinates are conditionally dependent. These conditional dependency relations are encoded in the support of Ω. A particularly popular approach to estimating Ω is the graphical LASSO (GLASSO), which adds an ℓ_1 penalty to the negative log-likelihood of Ω (see, e.g., Yuan and Lin 2007; Banerjee, Ghaoui, and d'Aspremont 2008; Friedman, Hastie, and Tibshirani 2008).

While variable selection and covariance selection each have long, rich histories, joint variable and covariance selection has only recently attracted attention. To the best of our knowledge, Rothman, Levina, and Zhu (2010) were among the first to consider the simultaneous sparse estimation of B and Ω, solving the penalized likelihood problem

CONTACT Sameer K. Deshpande ✉ sameerd@alum.mit.edu Massachusetts Institute of Technology, 32 Vassar St., Bldg 32-G451, Cambridge, MA 02139-4307.
Supplementary materials for this article are available online. Please go to www.tanfonline.com/r/JCGS.

$$\arg\min_{B,\Omega} \left\{ -\frac{n}{2}\log|\Omega| + \frac{1}{2}\mathrm{tr}\left((\mathbf{Y}-\mathbf{X}B)\,\Omega\,(\mathbf{Y}-\mathbf{X}B)'\right) + \lambda\sum_{j,k}|\beta_{j,k}| + \xi\sum_{k\neq k'}|\omega_{k,k'}| \right\}. \quad (1)$$

Their procedure, called MRCE for "multivariate regression with covariance estimation," induces sparsity in B and Ω with separate ℓ_1 penalties and can be viewed as an elaboration of both the LASSO and GLASSO. Following Rothman, Levina, and Zhu (2010), several authors have proposed solving problems similar to that in Equation (1): Yin and Li (2011) considered nearly the same objective but with adaptive LASSO penalties, Lee and Liu (2012) proposed weighting each $|\beta_{j,k}|$ and $|\omega_{k,k'}|$ individually, and Abegaz and Wit (2013) replaced the ℓ_1 penalties with SCAD penalties. Though the ensuing joint optimization problem can be numerically unstable in high-dimensions, all of these authors report relatively good performance in estimating B and Ω. Cai et al. (2013) took a somewhat different approach, first estimating B in a column-by-column fashion with a separate Dantzig selector for each response and then estimating Ω by solving a constrained ℓ_1 optimization problem. Under mild conditions, they established the asymptotic consistency of their two-step procedure, called CAPME for "covariate-adjusted precision matrix estimation."

Bayesians too have considered variable and covariance selection. A workhorse of sparse Bayesian modeling is the spike-and-slab prior, in which one models parameters as being drawn *a priori* from either a point-mass at zero (the "spike") or a much more diffuse continuous distribution (the "slab") (Mitchell and Beauchamp 1988). George and McCulloch (1993) relaxed this formulation slightly by taking the spike and slab distributions to be zero-mean Gaussians, with the spike distribution very tightly concentrated around zero. More recently, Ročková and George (2018) took both the spike and slab distributions to be Laplacian, which led to posterior distributions with exactly sparse modes. Under mild conditions, their "spike-and-slab LASSO" prior produces posterior distributions that concentrate asymptotically around the true regression coefficients at nearly the minimax rate.

An important Bayesian approach to covariance selection begins by specifying a prior over the underlying graph G and a hyper-inverse Wishart prior (Dawid and Lauritzen 1993) on $\Sigma|G$. This prior is constrained to the set of symmetric positive-definite matrices such that off-diagonal entry $\omega_{k,k'}$ of $\Sigma^{-1} = \Omega$ is nonzero if and only if there is an edge between vertices k and k' in G. Though it is conceptually appealing, using the hyper-inverse Wishart prior is computationally prohibitive in high dimensions, requiring repeated approximation of the marginal posterior probability of G. See Carvalho, Massam, and West (2007) and Jones et al. (2005) for computational details and Uhler, Lenkoski, and Richards (2018) for recent expressions of this probability. Rather than using the hyper-inverse Wishart, Wang (2015) and Banerjee and Ghosal (2015) placed spike-and-slab priors on the off-diagonal elements of Ω, using a Laplacian slab and a point-mass spike at zero, with the latter authors establishing posterior consistency results.

Despite their conceptual elegance, spike-and-slab priors result in highly multimodal posteriors that can slow the mixing of MCMC simulations. This is exacerbated in the multivariate regression setting, especially when p and q are moderate-to-large relative to n. To overcome slow mixing, Brown, Vannucci, and Fearn (1998) and Bhadra and Mallick (2013) restricted attention to multivariate linear regression models in which a variable was selected as "relevant" to either all or none of the outcomes. Despite the computational tractability, this "all-or-nothing" approach may be unrealistic and overly restrictive. Richardson, Bottolo, and Rosenthal (2010) overcame this with an evolutionary MCMC simulation, but made the equally restrictive assumption that Ω was diagonal.

In this article, we extend the deterministic EMVS framework of Ročková and George (2014) and spike-and-slab LASSO framework of Ročková and George (2018) to the multivariate linear regression setting without imposing any restrictions on the supports of B and Ω. Like EMVS, our proposed procedure targets posterior modes and reduces to solving a series of penalized likelihood problems with adapting penalties. Our prior model of the uncertainty about which free parameters are large and which are essentially negligible allows us to perform selective shrinkage, leading to vastly superior support recovery and estimation performance compared to nonadaptive procedures like MRCE and CAPME. Moreover, we have found our joint treatment of B and Ω, which embraces the residual correlation structure from the outset, can sometimes detect smaller covariate effects than two-step procedures that first estimate B either column-wise or by assuming $\Omega = I$ and then estimate Ω.

The rest of this article is organized as follows. We formally introduce our model and algorithm in Section 2. In Section 3, we embed this algorithm within a path-following scheme that facilitates *dynamic posterior exploration*, identifying putative modes of B and Ω over a range of different posterior distributions indexed by the "tightness" of the prior spike distributions. We present the results of several simulation studies in Section 3.2. In Section 4, we reanalyze the data of Deshpande et al. (2017), a recent observational study on the effects of playing high school football on a range of cognitive, behavioral, psychological, and socio-economic outcomes later in life. We conclude with a discussion in Section 5.

2. Model and Algorithm

We begin with some notation. For any matrix of covariates effects B, we let $\mathbf{R}(B) = \mathbf{Y}-\mathbf{X}B$ denote the residual matrix whose kth column is denoted $\mathbf{r}_k(B)$. Finally, let $S(B) = n^{-1}\mathbf{R}(B)'\mathbf{R}(B)$ be the residual covariance matrix. In what follows, we will usually suppress the dependence of $\mathbf{R}(B)$ and $S(B)$ on B, writing only $\mathbf{R}$ and S. Additionally, we assume that the columns of $\mathbf{X}$ have been centered and scaled to have mean 0 and Euclidean norm $\sqrt{n}$ and that the columns of $\mathbf{Y}$ have been centered and are on approximately similar scales.

Recall that our data likelihood is given by

$$p(\mathbf{Y}|B,\Omega) \propto |\Omega|^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left((\mathbf{Y}-\mathbf{X}B)\,\Omega\,(\mathbf{Y}-\mathbf{X}B)'\right)\right\}.$$

We introduce latent 0–1 indicators, $\boldsymbol{\gamma} = (\gamma_{j,k} : 1 \le j \le p, 1 \le k \le q)$ so that, independently for $1 \le j \le p, 1 \le k \le q$, we have

$$\pi(\beta_{j,k}|\gamma_{j,k}) \propto \left(\lambda_1 \mathrm{e}^{-\lambda_1|\beta_{j,k}|}\right)^{\gamma_{j,k}} \left(\lambda_0 \mathrm{e}^{-\lambda_0|\beta_{j,k}|}\right)^{1-\gamma_{j,k}}.$$

Similarly, we introduce latent 0–1 indicators, $\boldsymbol{\delta} = (\delta_{k,k'} : 1 \le k < k' \le q)$ so that, independently for $1 \le k < k' \le q$, we have

$$\pi(\omega_{k,k'}|\delta_{k,k'}) \propto \left(\xi_1 \mathrm{e}^{-\xi_1|\omega_{k,k'}|}\right)^{\delta_{k,k'}} \left(\xi_0 \mathrm{e}^{-\xi_0|\omega_{k,k'}|}\right)^{1-\delta_{k,k'}}.$$

Recall that in the spike-and-slab framework, the spike distribution is viewed as having a priori generated all of the negligible parameter values, permitting us to interpret $\gamma_{j,k} = 0$ as an indication that variable j has an essentially null effect on outcome k. Similarly, we may interpret $\delta_{k,k'} = 0$ to mean that the partial covariance between $\mathbf{r}_k$ and $\mathbf{r}_{k'}$ is small enough to ignore. To model our uncertainty about $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, we use the familiar beta-binomial prior (Scott and Berger 2010): $\gamma_{j,k} \sim \text{Bernoulli}(\theta), \theta \sim \text{Beta}(a_\theta, b_\theta)$ and $\delta_{k,k'} \sim \text{Bernoulli}(\eta), \eta \sim \text{Beta}(a_\eta, b_\eta)$, where $a_\theta, b_\theta, a_\eta$, and b_η are fixed positive constants, and $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are a priori independent. We may view θ and η as measuring the proportion of nonzero entries in B and nonzero off-diagonal elements of Ω, respectively.

Following the example of Wang (2015) and Banerjee and Ghosal (2015), we place independent exponential $\text{Exp}(\xi_1)$ priors on the diagonal elements of Ω. This introduces mild regularization to prevent the diagonal elements $\omega_{k,k}$ from becoming massive. At this point, it is worth noting that the requirement Ω be positive definite introduces dependence between the $\omega_{k,k'}$'s not currently reflected in the above prior. In fact, generally speaking, placing independent spike-and-slab priors on the off-diagonal elements and independent exponential priors along the diagonal leads to considerable prior probability being placed outside the cone of symmetric positive semidefinite matrices. In light of this, we complete our prior specification by formally truncating to the space of positive definite matrices so that the conditional prior density of $\Omega|\boldsymbol{\delta}$ can be written

$$\begin{aligned}\pi(\Omega|\eta) \propto & \left(\prod_{k=1}^{q} \xi_1 \mathrm{e}^{-\xi_1 \omega_{k,k}}\right) \\ & \times \left(\prod_{k<k'} \left\{\delta_{k,k'} \frac{\xi_1}{2} \mathrm{e}^{-\xi_1|\omega_{k,k'}|}\right.\right. \\ & \left.\left. + (1-\delta_{k,k'}) \frac{\xi_0}{2} \mathrm{e}^{-\xi_0|\omega_{k,k'}|}\right\}\right) \times \mathbb{I}(\Omega \succ 0).\end{aligned}$$

We note in passing that Wang (2015), Banerjee and Ghosal (2015), and Gan, Narisetty, and Liang (2018) employ similar truncation. For compactness, we will suppress the restriction $\mathbb{I}(\Omega \succ 0)$ in what follows.

Before proceeding, we take a moment to introduce two functions that will play a critical role in our optimization strategy. Given $\lambda_1, \lambda_0, \xi_1$, and ξ_0, define the functions $p^\star, q^\star : \mathbb{R} \times [0,1] \to [0,1]$ by

$$p^\star(x,\theta) = \frac{\theta\lambda_1 \mathrm{e}^{-\lambda_1|x|}}{\theta\lambda_1 \mathrm{e}^{-\lambda_1|x|} + (1-\theta)\lambda_0 \mathrm{e}^{-\lambda_0|x|}},$$

$$q^\star(x,\eta) = \frac{\eta\xi_1 \mathrm{e}^{-\xi_1|x|}}{\eta\xi_1 \mathrm{e}^{-\xi_1|x|} + (1-\eta)\xi_0 \mathrm{e}^{-\xi_0|x|}}.$$

Letting Ξ denote the collection $\{B, \theta, \Omega, \eta\}$, it is straightforward to verify that $p^\star(\beta_{j,k}, \theta) = \mathbb{E}\left[\gamma_{j,k}|\mathbf{Y}, \Xi\right]$ and $q^\star(\omega_{k,k'}, \eta) = \mathbb{E}\left[\delta_{k,k'}|\mathbf{Y}, \Xi\right]$, the conditional posterior probabilities that $\beta_{j,k}$ and $\omega_{k,k'}$ were drawn from their respective slab distributions.

Integrating out the latent indicators, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, the log-posterior density of Ξ is, up to an additive constant, given by

$$\begin{aligned}\log \pi(\Xi|\mathbf{Y}) = & \frac{n}{2}\log|\Omega| - \frac{1}{2}\mathrm{tr}\left((\mathbf{Y} - \mathbf{X}B)'(\mathbf{Y} - \mathbf{X}B)\Omega\right) \\ & + \sum_{j,k} \log\left(\theta\lambda_1 \mathrm{e}^{-\lambda_1|\beta_{j,k}|} + (1-\theta)\lambda_0 \mathrm{e}^{-\lambda_0|\beta_{j,k}|}\right) \\ & + \sum_{k,k'} \log\left(\eta\xi_1 \mathrm{e}^{-\xi_1|\omega_{k,k'}|} + (1-\eta)\xi_0 \mathrm{e}^{-\xi_0|\omega_{k,k'}|}\right) \\ & - \xi_1 \sum_{k=1}^{q} \omega_{k,k} + (a_\theta - 1)\log\theta + (b_\theta - 1) \\ & \times \log(1-\theta) + (a_\eta - 1)\log\eta + (b_\eta - 1) \\ & \times \log(1-\eta). \end{aligned} \qquad (2)$$

Rather than directly sample from this intractable posterior distribution with MCMC, we maximize the posterior density, seeking $\Xi^* = \arg\max\left\{\log\pi(\Xi|\mathbf{Y})\right\}$. Performing this joint optimization is quite challenging, especially in light of the nonconvexity of the log-posterior density. To overcome this, we use an expectation/conditional maximization (ECM) algorithm (Meng and Rubin 1993) that treats only the partial covariance indicators $\boldsymbol{\delta}$ as "missing data." For the E step of this algorithm, we first compute $q^\star_{k,k'} := q^\star(\omega^{(t)}_{k,k'}, \eta^{(t)}) = \mathbb{E}\left[\delta_{k,k'}|\mathbf{Y}, \Xi^{(t)}\right]$ given a current estimate $\Xi^{(t)}$ and then consider maximizing the surrogate objective function

$$\begin{aligned}& \mathbb{E}\left[\log\pi(\Xi, \boldsymbol{\delta}|\mathbf{Y})|\Xi^{(t)}\right] \\ & = \frac{n}{2}\log|\Omega| - \frac{1}{2}\mathrm{tr}\left((\mathbf{Y} - \mathbf{X}B)'(\mathbf{Y} - \mathbf{X}B)\Omega\right) \\ & \quad + \sum_{j,k} \log\left(\theta\lambda_1 \mathrm{e}^{-\lambda_1|\beta_{j,k}|} + (1-\theta)\lambda_0 \mathrm{e}^{-\lambda_0|\beta_{j,k}|}\right) \\ & \quad - \sum_{k,k'} \xi^\star_{k,k'}|\omega_{k,k'}| - \xi_1 \sum_{k=1}^{q} \omega_{k,k} \\ & \quad + (a_\theta - 1)\log\theta + (b_\theta - 1)\log(1-\theta) \\ & \quad + (a_\eta - 1)\log\eta + (b_\eta - 1)\log(1-\eta),\end{aligned}$$

where $\xi^\star_{k,k'} = \xi_1 q^\star_{k,k'} + \xi_0(1 - q^\star_{k,k'})$. We then perform two CM steps, first updating the pair (B, θ) while holding $(\Omega, \eta) = (\Omega^{(t)}, \eta^{(t)})$ fixed at its previous value and then updating (Ω, η) while fixing $(B, \theta) = (B^{(t+1)}, \theta^{(t+1)})$ at its new value. As we will see shortly, augmenting our log-posterior with the indicators $\boldsymbol{\delta}$ facilitates simple updates of Ω by solving a GLASSO problem. We do not also augment our log-posterior with the indicators $\boldsymbol{\gamma}$ as the update of B can be carried out with a coordinate ascent strategy despite the nonconvex penalty seen in the second line of Equation (2).

Holding $(\Omega, \eta) = (\Omega^{(t)}, \eta^{(t)})$ fixed, we update (B, θ) by solving

$$\begin{aligned}(B^{(t+1)}, \theta^{(t+1)}) = \underset{B,\theta}{\arg\max} & \left\{-\frac{1}{2}\mathrm{tr}\left((\mathbf{Y} - \mathbf{X}B)\Omega(\mathbf{Y} - \mathbf{X}B)'\right)\right. \\ & \left. + \log\pi(B|\theta) + \log\pi(\theta)\right\},\end{aligned} \qquad (3)$$

where

$$\pi(B|\theta) = \prod_{j,k} \left(\theta\lambda_1 \mathrm{e}^{-\lambda_1|\beta_{j,k}|} + (1-\theta)\lambda_0 \mathrm{e}^{-\lambda_0|\beta_{j,k}|}\right)$$

and $\pi(\theta) \propto \theta^{a_\theta - 1}(1-\theta)^{b_\theta - 1}$. We do this in a coordinate-wise fashion, sequentially updating θ with a simple Newton algorithm and updating B by solving the following problem

$$\tilde{B} = \arg\max_{B} \left\{ -\frac{1}{2}\mathrm{tr}\left((\mathbf{Y} - \mathbf{X}B)\,\Omega\,(\mathbf{Y} - \mathbf{X}B)'\right) + \sum_{j,k} \mathrm{pen}(\beta_{j,k}|\theta) \right\}, \tag{4}$$

where

$$\begin{aligned}\mathrm{pen}(\beta_{j,k}|\theta) &= \log\left(\frac{\pi\left(\beta_{j,k}|\theta\right)}{\pi(0|\theta)}\right) \\ &= -\lambda_1 \left|\beta_{j,k}\right| + \log\left(\frac{p^\star(\beta_{j,k},\theta)}{p^\star(0,\theta)}\right).\end{aligned}$$

Using the fact that the columns of $\mathbf{X}$ have norm $\sqrt{n}$ and Lemma 2.1 of Ročková and George (2018), the Karush–Kuhn–Tucker condition tell us that

$$\tilde{\beta}_{j,k} = n^{-1}\left[\left|z_{j,k}\right| - \lambda^\star(\tilde{\beta}_{j,k},\theta)\right]_+ \mathrm{sign}(z_{j,k}),$$

where $\lambda^\star_{j,k} := \lambda^\star(\tilde{\beta}_{j,k},\theta) = \lambda_1 p^\star(\tilde{\beta}_{j,k},\theta) + \lambda_0(1 - p^\star(\tilde{\beta}_{j,k},\theta))$ and

$$z_{j,k} = n\tilde{\beta}_{j,k} + \sum_{k'} \frac{\omega_{k,k'}}{\omega_{k,k}} \mathbf{x}_j' \mathbf{r}_{k'}(\tilde{B}).$$

The form of $\tilde{\beta}_{j,k}$ above immediately suggests a coordinate ascent strategy with soft-thresholding to compute $\tilde{B}$ that is very similar to the one used to compute LASSO solutions (Friedman et al. 2007). As noted by Ročková and George (2018), however, this necessary characterization of $\tilde{B}$ is sufficient only when posterior is unimodal. In general, when $p > n$ and when λ_0 and λ_1 are far apart, the posterior tends to be highly multimodal. In light of this, cyclically applying the soft-thresholding operator may terminate at a sub-optimal local mode. Arguments in Zhang and Zhang (2012) and Ročková and George (2018) lead immediately to the following refined characterization of $\tilde{B}$, which blends hard- and soft-thresholding.

Proposition 1. The entries in the global mode $\tilde{B} = \left(\tilde{\beta}_{j,k}\right)$ in Equation (4) satisfy

$$\tilde{\beta}_{j,k} = \begin{cases} n^{-1}\left[\left|z_{j,k}\right| - \lambda^\star(\tilde{\beta}_{j,k},\theta)\right]_+ \times \mathrm{sign}\left(z_{j,k}\right) & \text{when } \left|z_{j,k}\right| > \Delta_{j,k}, \\ 0 & \text{when } \left|z_{j,k}\right| \le \Delta_{j,k} \end{cases}$$

where

$$\Delta_{j,k} = \inf_{t>0}\left\{\frac{nt}{2} - \frac{\mathrm{pen}(\tilde{\beta}_{j,k},\theta)}{\omega_{k,k}t}\right\}.$$

As it turns out, the element-wise thresholds $\Delta_{j,k}$ are generally quite hard to compute but can be bounded, as seen in the following analog to Theorem 2.1 of Ročková and George (2018).

Proposition 2. Suppose that $(\lambda_1 - \lambda_0) > 2\sqrt{n\omega_{k,k}}$ and $(\lambda^\star(0,\theta) - \lambda_1)^2 > -2n\omega_{k,k}p^\star(0,\theta)$. Then $\Delta^L_{j,k} \le \Delta_{j,k} \le \Delta^U_{j,k}$ where

$$\begin{aligned}\Delta^L_{j,k} &= \sqrt{-2n\omega^{-1}_{k,k}\log p^\star(0,\theta) - \omega^{-2}_{k,k}d} + \omega^{-1}_{k,k}\lambda_1 \\ \Delta^U_{j,k} &= \sqrt{-2n\omega^{-1}_{k,k}\log p^\star(0,\theta)} + \omega^{-1}_{k,k}\lambda_1,\end{aligned}$$

where $d = -\left(\lambda^\star(\delta_{c_+},\theta) - \lambda_1\right)^2 - 2n\omega_{k,k}\log p^\star(\delta_{c_+},\theta)$ and δ_{c_+} is the larger root of $\mathrm{pen}''(x|\theta) = \omega_{k,k}$.

Together Propositions 1 and 2 suggest a *refined coordinate ascent* strategy for updating our estimate of B. Namely, starting from some initial value B^{old}, we can update $\beta_{j,k}$ with the thresholding rule

$$\beta^{\mathrm{new}}_{j,k} = \frac{1}{n}\left(\left|z_{j,k}\right| - \lambda^\star(\beta^{\mathrm{old}}_{j,k},\theta)\right)_+ \mathrm{sign}(z_{j,k})\mathbb{I}\left(\left|z_{j,k}\right| > \Delta^U_{j,k}\right). \tag{5}$$

Before proceeding, we pause for a moment to reflect on the threshold $\lambda^\star_{j,k}$ appearing in the KKT condition and Proposition 1, which evolves alongside our estimates of B and θ. In particular, when our current estimate of $\beta_{j,k}$ is large in magnitude, the conditional posterior probability that it was drawn from the slab, $p^\star_{j,k}$, tends to be close to one so that $\lambda^\star_{j,k}$ is close to λ_1. On the other hand, if it is small in magnitude, $\lambda^\star_{j,k}$ tends to be close to the much larger λ_0. In this way, as our EM algorithm proceeds, it performs *selective shrinkage*, aggressively penalizing small values of $\beta_{j,k}$ without overly penalizing larger values. It is worth pointing out as well that $\lambda^\star_{j,k}$ adapts not only to the current estimate of B but also to the overall level of sparsity in B, as reflected in the current estimate of θ. The adaptation is entirely a product our explicit a priori modeling of the latent indicators $\boldsymbol{\gamma}$ and stands in stark contrast to regularization techniques that deploy fixed penalties.

Fixing $(\Omega,\eta) = (\Omega^{(t)},\eta^{(t)})$, we iterate between the refined coordinate ascent for B and the Newton algorithm for θ until some convergence criterion is reached at some new estimate $(B^{(t+1)},\theta^{(t+1)})$. Then, holding $(B,\theta) = (B^{(t+1)},\theta^{(t+1)})$, we turn our attention to (Ω,η) and solving the posterior maximization problem

$$\begin{aligned}&\left(\Omega^{(t+1)},\eta^{(t+1)}\right) \\ &\quad= \arg\max\left\{\frac{n}{2}\log|\Omega| - \frac{1}{2}\mathrm{tr}\left(S\Omega\right) - \sum_{k<k'} \xi^\star_{k,k'}\left|\omega_{k,k'}\right| \right. \\ &\qquad - \xi_1\sum_{k=1}^{q}\omega_{k,k} + \log\eta \times \left(a_\eta - 1 + \sum_{k<k'} q^\star_{k,k'}\right) \\ &\qquad \left. + \log\left(1-\eta\right) \times \left(b_\eta - 1 + \sum_{k<k}(1 - q^\star_{k,k'})\right)\right\}.\end{aligned}$$

It is immediately clear that there is a closed form update of η

$$\eta^{(t+1)} = \frac{a_\eta - 1 + \sum_{k<k'} q^\star_{k,k'}}{a_\eta + b_\eta - 2 + q(q-1)/2}.$$

For Ω, we recognize the M Step update of Ω as a GLASSO problem.

$$\Omega^{(t+1)} = \underset{\Omega \succ 0}{\arg\max} \left\{ \frac{n}{2}\log|\Omega| - \frac{n}{2}\mathrm{tr}\,(S\Omega) - \sum_{k<k'} \xi^{\star}_{k,k'} \left|\omega_{k,k'}\right| - \xi_1 \sum_{k=1}^{q} \omega_{k,k} \right\}. \qquad (6)$$

To find $\Omega^{(t+1)}$, rather than using the block-coordinate ascent algorithms of Friedman, Hastie, and Tibshirani (2008) and Witten, Friedman, and Simon (2011), we use the state-of-art QUIC algorithm of Hsieh et al. (2014), which is based on a quadratic approximation of the objective function and achieves a superlinear convergence rate. Each of these algorithms returns a positive semidefinite $\Omega^{(t+1)}$. Just like with the $\lambda^{\star}_{j,k}$'s, the penalties $\xi^{\star}_{k,k'}$ in Equation (6) adapt to the values of the current estimates of $\omega_{k,k'}$ and the overall level of sparsity in Ω, captured by η.

In our implementation, we iterate between the E and CM steps until the percentage change in every $\beta_{j,k}$ and $\omega_{k,k'}$ estimate is less than a user-specified threshold (e.g., 10^{-3} or 10^{-6}) or if the percentage increase in objective value is less than that same threshold. Because of the nonconvexity of our log-posterior, there are no theoretical guarantees that our algorithm terminates at a global mode. Indeed, with our stopping criterion, the most we can say is that it will terminate in the vicinity of a stationary point.

Finally, we note that our proposed framework for simultaneous variable and covariance selection can easily be modified to estimate B when Ω is known and to estimate Ω when B is known. Concurrently with but independently of us, Li and McCormick (2017) and Gan, Narisetty, and Liang (2018) have developed similar ECM algorithms for Gaussian graphical model estimation using spike-and-slab mixtures of Gaussian and Laplacian distributions, respectively.

3. Dynamic Posterior Exploration

Given any specification of hyper-parameters $\left(a_\theta, b_\theta, a_\eta, b_\eta\right)$ and $(\lambda_1, \lambda_0, \xi_1, \xi_0)$, it is straightforward to deploy the ECM algorithm described in the previous section to identify a putative posterior mode. We may moreover run our algorithm over a range of hyper-parameter settings to estimate the mode of a range of different posteriors. Unlike MCMC, which expends considerable computational effort sampling from a single posterior, this *dynamic posterior exploration* provides a quick snapshot of several different posteriors.

In the univariate regression setting, Ročková and George (2018) proposed a path-following scheme in which they fixed λ_1 and identified modes of a range of posteriors indexed by a ladder of L increasing λ_0 values, $\mathcal{I}_\lambda$ with sequential reinitialization to produce a sequence of posterior modes. In this path-following scheme, as λ_0 increases, the spike distribution increasingly absorbs negligible parameter estimates and results in sparse modes. Remarkably, Ročková and George (2018) found that individual parameter estimates stabilized early in the path, indicating that the parameters had cleanly segregated into groups of zero and nonzero values.

Extending this dynamic posterior exploration strategy to our multivariate setting is straightforward: we now specify two ladders $\mathcal{I}_\lambda$ and $\mathcal{I}_\xi$ of L increasing λ_0 and ξ_0 values. We then identify a sequence $\{\hat{\Xi}^{s,t} : 1 \leq s, t \leq L\}$ where $\hat{\Xi}^{s,t}$ is an estimate of the posterior mode corresponding to the choice $(\lambda_0, \xi_0) = (\lambda_0^{(s)}, \xi_0^{(t)})$, which we denote $\Xi^{s,t*}$. When it comes time to estimate $\Xi^{s,t*}$, we launch our ECM algorithm from whichever of $\hat{\Xi}^{s-1,t}$, $\hat{\Xi}^{s,t-1}$ and $\hat{\Xi}^{s-1,t-1}$ has the largest log-posterior density, computed according to Equation (2) with $\lambda_0 = \lambda_0^{(s)}$ and $\xi_0 = \xi_0^{(t)}$. We implement this dynamic posterior exploration by starting with $B = \mathbf{0}, \Omega = I$ and looping over the λ_0^s values and ξ_0^t values. Proceeding in this way, we propagate a single estimate of Ξ through a series of prior filters indexed by the pair $(\lambda_0^{(s)}, \xi_0^{(t)})$.

When λ_0 is close to λ_1, our refined coordinate ascent can sometimes promote the inclusion of many negligible but non-null $\beta_{j,k}$'s. Such a specification could over-explain the variation in $\mathbf{Y}$ using several covariates, leaving very little to the residual conditional dependency structure and a severely ill-conditioned residual covariance matrix S. In our implementation, we do not propagate any $\hat{\Xi}^{s,t}$ where the accompanying S has condition number exceeding $10n$. While this choice is decidedly arbitrary, we have found it to work rather well in our simulation studies. When it comes time to estimate $\Xi^{s,t*}$, if each of $\hat{\Xi}^{s-1,t}$, $\hat{\Xi}^{s,t-1}$ and $\hat{\Xi}^{s-1,t-1}$ is numerically unstable, we relaunch our EM algorithm from $B = \mathbf{0}$ and $\Omega = I$.

To illustrate this procedure, which we call mSSL-DPE for "multivariate spike-and-slab LASSO with dynamic posterior exploration," we simulate data from the following model with $n = 400, p = 500$, and $q = 25$. We generate the matrix $\mathbf{X}$ according to a $\mathrm{N}_p\left(\mathbf{0}_p, \Sigma_X\right)$ distribution where $\Sigma_X = \left(0.7^{|j-j'|}\right)_{j,j'=1}^{p}$. We construct matrix B_0 with $pq/5$ randomly placed nonzero entries independently drawn uniformly from the interval $[-2, 2]$. This allows us to gauge mSSL-DPE's ability to recover signals of varying strength. We then set $\Omega_0^{-1} = \left(0.9^{|k-k'|}\right)_{k,k'=1}^{q}$ so that Ω_0 is tridiagonal, with the $q-1$ nonzero entries in the upper-triangle immediately above the diagonal. Finally, we generate data $\mathbf{Y} = \mathbf{X}B_0 + E$ where the rows of E are independently $\mathrm{N}\left(\mathbf{0}_q, \Omega_0^{-1}\right)$. For this simulation, we set $\lambda_0 = 1, \xi_0 = 0.01n$ and let I_λ and I_ξ contain $L = 10$ equally space values ranging from 10 to n and $0.1n$ to n, respectively.

To establish posterior consistency in the univariate linear regression, Ročková and George (2018) required the prior on θ to place most of its probability in a small interval near zero and recommended taking $a_\theta = 1$ and $b_\theta = p$. This concentrates their prior on models that are relatively sparse. With pq coefficients in B, we take $a_\theta = 1$ and $b_\theta = pq$ for this demonstration. We further take $a_\eta = 1$ and $b_\eta = q$, so that the prior on the underlying residual Gaussian graph G concentrates on very sparse graphs with average degree just less than one.

We find that for small values of λ_0 and ξ_0, the modes are predictably numerically unstable. However, once λ_0 is moderate, we find that estimated supports of B and Ω tend not to change. In this particular setting, for most combinations of λ_0 and ξ_0, mSSL-DPE identified the same 2367 nonzero $\beta_{j,k}$'s (2363 true positives, 4 false positives, 137 false negatives) and the same 27 nonzero $\omega_{k,k'}$'s in the upper triangle of Ω (23 true positives, 4 false positives, 1 false negative). In addition to the supports stabilizing, the actual parameter estimates seemed to converge as well. The apparent stabilization allows us to report a single estimate, $\hat{\Xi}^{L,L}$

While there is no general guarantee for stabilization for arbitrary ladders $\mathcal{I}_\lambda$ and $\mathcal{I}_\xi$, in all of the examples we have tried, we found that stabilization occurred long before $\lambda_0^{(s)}$ and $\xi_0^{(t)}$ reached their final values. Moreover, once the modes stabilized, our algorithm runs quite quickly so the excess computations are not at all burdensome. In Appendix A in the online supplement, we develop some intuition about why such stabilization may occur. Roughly, when λ_0 is large and $\beta_{j,k}$ is large enough (resp. small enough), the corresponding $p^\star_{j,k}$ will be close to one (resp. zero), so that increasing λ_0 will result in subjecting $z_{j,k}$ to a smaller (resp. larger) threshold in our refined coordinate ascent. This means that large nonzero $\beta_{j,k}$ estimates remain nonzero when we increase λ_0 and small ones are shrunk more aggressively to 0.

3.1. Faster Dynamic Conditional Posterior Mode Exploration

mSSL-DPE can expend considerable computational effort identifying modal estimates $\hat{\Xi}^{s,t}$ corresponding to smaller values of λ_0 and ξ_0. Although the support recovery performance of $\hat{\Xi}^{L,L}$ from mSSL-DPE is very promising, one might also consider streamlining the procedure using the following procedure we term mSSL-DCPE for "dynamic *conditional* posterior exploration." First, we fix $\Omega = I$ and sequentially solve Equation (3) for each $\lambda_0 \in I_\lambda$, with warm-starts. This produces a sequence $\{(\hat{B}^s, \hat{\theta}^s)\}$ of *conditional* posterior modes of $(B, \theta)\,|\mathbf{Y}, \Omega = I$. Then, holding $(B, \theta) = (\hat{B}_0^L, \hat{\theta}_0^L)$ fixed, we run a modified version of our dynamic posterior exploration to produce a sequence $\{(\hat{\Omega}^t, \hat{\eta}^t)\}$ of conditional modes of $(\Omega, \eta)|\mathbf{Y}, B = \hat{B}^L$. We finally run our ECM algorithm starting from $(\hat{B}^L, \hat{\theta}^L, \hat{\Omega}^L, \hat{\eta}^L)$ with $\lambda_0 = \lambda_0^L$ and $\xi_0 = \xi_0^L$ to arrive at an estimate of $\Xi^{L,L*}$, which we denote $\tilde{\Xi}^{L,L}$. In general, $\tilde{\Xi}^{L,L}$ will be different than the solution obtained by mSSL-DPE, $\hat{\Xi}^{L,L}$, since the two algorithms typically launch the ECM algorithm from different points when it comes time to estimate $\Xi^{L,L*}$.

In sharp contrast to mSSL-DPE, which visits several joint posterior modes before reaching an estimate of posterior mode $\Xi^{L,L*}$, mSSL-DCPE visits several conditional posterior modes to reach another estimate of the same mode. On the same dataset from the previous subsection, mSSL-DCPE performs somewhat worse than mSSL-DPE. In particular, mSSL-DCPE correctly identified only 2201 of the 2500 nonzero $\beta_{j,k}$'s while making 18 false positives and found all 24 nonzero $\omega_{k,k'}$'s with 32 false positives. This was all accomplished in about 11 sec, a considerable improvement over the nearly 20 min runtime of mSSL-DPE. Despite the obvious improvement in runtime, mSSL-DCPE terminated at a suboptimal point whose log-posterior density was much smaller than the solution found by mSSL-DPE. All of the false negative identifications in the support of B made by both procedures corresponded to $\beta_{j,k}$ values which were relatively small in magnitude. Interestingly, mSSL-DPE was able to detect smaller signals than mSSL-DCPE. We will return to this point later in Section 3.2.

3.2. Simulations

We now assess the performance of mSSL-DPE and mSSL-DCPE on simulated high-dimensional data, with $n = 400, p = 500$, and $q = 25$. Just as above, we generate the matrix $\mathbf{X}$ according to a $\mathrm{N}_p\left(\mathbf{0}_p, \Sigma_X\right)$ distribution where $\Sigma_X = \left(0.7^{|j-j'|}\right)_{j,j'=1}^{p}$. We construct matrix B_0 with $pq/5$ randomly placed nonzero entries independently drawn uniformly from the interval $[-2, 2]$. We then set $\Omega_0^{-1} = \left(\rho^{|k-k'|}\right)_{k,k'=1}^{q}$ for $\rho \in \{0, 0.5, 0.7, 0.9\}$. When $\rho \neq 0$, the resulting Ω_0 is tridiagonal. Finally, we generate data $\mathbf{Y} = \mathbf{X}B_0 + E$ where the rows of E are independently $\mathrm{N}\left(\mathbf{0}_q, \Omega_0^{-1}\right)$. For this simulation, we set $\lambda_1 = 1, \xi_1 = 0.01n$ and set $\mathcal{I}_\lambda$ and $\mathcal{I}_\xi$ to contain $L = 10$ equally spaced values ranging from 1 to n and from $0.1n$ to n, respectively. Like in the previous subsection, we took $a_\theta = 1, b_\theta = pq, a_\eta = 1$, and $b_\eta = q$. We examine the sensitivity of our results to these hyper-parameters more carefully in Appendix B in the online supplement.

We simulated 100 datasets according to each model, each time keeping B_0 and Ω_0 fixed but drawing a new matrix of errors E. To assess the support recovery and estimation performance, we tracked the following quantities: SEN (sensitivity), SPE (specificity), PREC (precision), ACC (accuracy), MCC (Matthew's correlation coefficient), MSE (mean square error in estimating B_0), FROB (squared Frobenius error in estimating Ω_0), and TIME (execution time in seconds). If we let TP, TN, FP, and FN denote the total number of true positive, true negative, false positive, and false negative identifications made in the support recovery, these quantities are defined as: $\mathrm{SEN} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, $\mathrm{SPE} = \mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$, $\mathrm{PREC} = \mathrm{TP}/(\mathrm{TP}+\mathrm{TN})$, $\mathrm{ACC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})$, and

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})\,(\mathrm{TP} + \mathrm{FN})\,(\mathrm{TN} + \mathrm{FP})\,(\mathrm{TN} + \mathrm{FN})}}.$$

Tables 1 and 2 report the average performance of mSSL-DPE, mSSL-DCPE, Rothman, Levina, and Zhu (2010)'s MRCE, Cai et al. (2013)'s CAPME, each with 5-fold cross-validation, and the following two competitors:

Sep.L+G: We first estimate B by solving separate LASSO problems with 10-fold cross-validation for each outcome. We then estimate Ω from the resulting residual matrix using the GLASSO procedure of Friedman, Hastie, and Tibshirani (2008), also run with 10-fold cross-validation.

Sep.SSL + SSG: We first estimate B column-by-column, deploying Ročková and George (2018)'s path-following SSL along the ladder $\mathcal{I}_\lambda$ separately for each outcome. We then run a modified version of our dynamic posterior exploration that holds B fixed and only updates Ω and η with the ECM algorithm along the ladder $\mathcal{I}_\xi$. This is similar to Sep.L+G but with adaptive spike-and-slab lasso penalties rather than fixed ℓ_1 penalties.

We additionally carried out similar simulations in a low-dimensional regime, with $n = 100, p = 50$, and $q = 25$. The performances results of these methods are presented in Tables 1 and 2 of Appendix B. All simulations were carried out in R 3.3.3 (R Core Team 2017). We used the implementation of MRCE available in the MRCE R package (Rothman 2017) and implemented Sep.L+G using the glmnet (Friedman, Hastie, and

Table 1. Average variable selection performance in the high-dimensional setting. MSE has been rescaled by a factor of 1000.

Method	SEN/SPE	PREC/ACC	MCC	MSE	TIME
Simulation 1: $n = 400, p = 500, q = 25, \rho = 0.9$					
mSSL-DPE	**0.95**/**1.00**	**1.00**/**0.99**	**0.96**	**0.41**	2229.21
mSSL-DCPE	0.88/**1.00**	0.99/0.97	0.92	1.40	23.66
MRCE	0.40/0.63	0.67/0.59	0.07	171.73	7116.94
CAPME	**0.95**/0.54	0.34/0.62	0.40	8.49	7625.05
SEP.L+G	0.92/0.76	0.49/0.79	0.56	10.33	19.21
SEP.SSL+SSG	0.88/**1.00**	0.98/0.97	0.91	2.25	**3.14**
Simulation 2: $n = 400, p = 500, q = 25, \rho = 0.7$					
mSSL-DPE	0.91/**1.00**	**0.99**/**0.98**	**0.94**	**1.19**	2260.99
mSSL-DCPE	0.88/**1.00**	**0.99**/0.97	0.92	1.65	23.93
MRCE	0.74/0.30	0.33/0.39	0.07	87.43	9092.95
CAPME	0.68/0.84	0.53/0.81	0.48	109.03	7243.94
SEP.L+G	**0.92**/0.76	0.48/0.79	0.56	10.26	19.11
SEP.SSL+SSG	0.88/**1.00**	0.98/0.97	0.91	2.22	**3.11**
Simulation 3: $n = 400, p = 500, q = 25, \rho = 0.5$					
mSSL-DPE	0.91/0.61	0.39/0.67	0.43	33.52	3839.74
mSSL-DCPE	0.88/**1.00**	**0.99**/**0.97**	**0.92**	**1.92**	24.12
MRCE	0.65/0.36	0.39/0.42	0.04	107.27	9540.04
CAPME	0.66/0.86	0.54/0.82	0.48	116.39	7594.80
SEP.L+G	**0.92**/0.76	0.49/0.79	0.56	10.23	19.28
SEP.SSL+SSG	0.88/**1.00**	0.98/**0.97**	0.91	2.22	**3.14**
Simulation 4: $n = 400, p = 500, q = 25, \rho = 0$					
mSSL-DPE	0.91/0.58	0.35/0.64	0.39	36.26	2800.63
mSSL-DCPE	0.88/**1.00**	**0.98**/**0.97**	**0.91**	2.25	23.82
MRCE	0.59/0.41	0.42/0.45	0.03	123.23	9187.28
CAPME	0.66/0.86	0.54/0.82	0.48	116.36	7255.42
SEP.L+G	**0.92**/0.76	0.49/0.79	0.56	10.27	19.26
SEP.SSL+SSG	0.88/**1.00**	**0.98**/**0.97**	**0.91**	**2.24**	**3.22**

NOTE: NaN indicates that the specified quantity was undefined, either because no nonzero estimates were returned or because there were truly no nonzero parameters (Simulation 4). Best results are bolded.

Tibshirani 2010) and glasso (Friedman, Hastie, and Tibshirani 2018) packages. The R package implementing CAPME is no longer available on CRAN but the source code is available at the CRAN archive. An R package implementing mSSL-DPE, mSSL-DPCE, Sep.SSL+SSG, and R scripts to replicate these simulations are available in the online supplementary materials. The simulations were carried out on a high-performance computing cluster, with each node running an Intel Xeon E5-2667 3.30 GHz processor. Each simulated dataset was analyzed on a single core with 5 GB of RAM.

In both the high- and low-dimensional settings (see Tables 1 and 2 in Appendix B), we see that the nonadaptive regularization methods utilizing cross-validation (MRCE, CAPME, and SEP.L+G) are characterized by high sensitivity, moderate specificity, and low precision in recovering the support of both B and Ω. That these three methods have precision less than 0.5 is worrying from a practical standpoint: the majority of nonzero estimates returned by these methods are false positives. This is not entirely surprising, as cross-validation has a well-known tendency to over-select. In stark contrast are mSSL-DPE, mSSL-DCPE, and SEP.SSL+SSG, which all used adaptive spike-and-slab penalties. These methods are all characterized by somewhat lower sensitivity than their cross-validated counterparts but with vastly improved specificity and precision, performing exactly as anticipated by Ročková and George (2018)'s simulations from the univariate setting. In a certain sense, the competitors cast a very wide net to capture most of the nonzero parameters, while our methods are much more discerning. So while the latter methods may not capture as much of the true signal as the former, they do not admit nearly as many false positives.

Recall that mSSL-DPCE proceeds by identifying two conditional modes and then refining them to a single joint mode. It is only in this last refining step that mSSL-DCPE considers the correlation between residuals while estimating B. As it turns out, this final refinement did little to change the estimated support of B, explaining the nearly identical performance of SEP.SSL+SSG and mSSL-DCPE. Further, the only practical difference between these two procedures is the adaptivity of the penalties on $\beta_{j,k}$: in SEP.SSL+SSG, the penalties separately adapt to the sparsity within each column of B while in mSSL-DCPE, they adapt to the overall sparsity of B.

By simulating the nonzero $\beta_{j,k}$'s uniformly from $[-2, 2]$, we were able to compare our methods' abilities to detect signals of varying strength. Figure 1 super-imposes the distribution of nonzero $\beta_{j,k}$'s correctly identified as nonzero with the distribution of nonzero $\beta_{j,k}$'s incorrectly identified as zero by mSSL-DPE and mSSL-DCPE from a single replication of Simulation 1.

In this situation, mSSL-DPE displays greater acuity for detecting smaller $\beta_{j,k}$'s than mSSL-DCPE, which is virtually ignorant of the covariance structure of the outcomes. This is very reminiscent of Zellner (1962)'s observation that multivariate estimation of B in seemingly unrelated regressions is asymptotically more efficient than proceeding response-by-response and ignoring the correlation between responses. We

Table 2. Average covariance selection performance in the high-dimensional setting.

Method	SEN/SPE	PREC/ACC	MCC	FROB	TIME
		Simulation 1: $n = 400, p = 500, q = 25, \rho = 0.9$			
mSSL-DPE	0.97/0.98	**0.84/0.98**	**0.89**	**97.92**	2229.21
mSSL-DCPE	**1.00**/0.89	0.45/0.90	0.63	1226.78	23.66
MRCE	0.94/0.22	0.11/0.28	0.21	6.17×10^6	7116.94
CAPME	0.00/**1.00**	NaN/0.92	NaN	2989.33	7625.05
SEP.L+G	**1.00**/0.60	0.18/0.63	0.33	2682.86	19.21
SEP.SSL+SSG	0.99/0.87	0.41/0.88	0.59	1953.69	**3.14**
		Simulation 2: $n = 400, p = 500, q = 25, \rho = 0.7$			
mSSL-DPE	0.99/**1.00**	**0.95/1.00**	**0.97**	22.10	2260.99
mSSL-DCPE	**1.00**/0.96	0.72/0.97	0.83	**14.36**	23.93
MRCE	0.92/0.45	0.16/0.49	0.28	16.16×10^6	9092.95
CAPME	0.00/**1.00**	NaN/0.92	NaN	285.86	7243.94
SEP.L+G	0.99/0.87	0.40/0.88	0.58	161.84	19.11
SEP.SSL+SSG	**1.00**/0.96	0.71/0.97	0.83	57.68	**3.11**
		Simulation 3: $n = 400, p = 500, q = 25, \rho = 0.5$			
mSSL-DPE	0.07/**1.00**	0.95/0.93	0.50	3.59×10^4	3839.74
mSSL-DCPE	**1.00/1.00**	0.97/**1.00**	0.98	**2.18**	24.12
MRCE	0.87/0.49	0.17/0.52	0.32	1.15×10^9	9540.04
CAPME	0.00/1.00	NaN/0.92	NaN	87.10	7594.80
SEP.L+G	0.86/0.96	0.66/0.95	0.72	29.30	19.28
SEP.SSL+SSG	**1.00/1.00**	**0.98/1.00**	**0.99**	4.38	**3.14**
		Simulation 4: $n = 400, p = 500, q = 25, \rho = 0$			
mSSL-DPE	NaN/**1.00**	NaN/**1.00**	NaN	4.03×10^4	2800.63
mSSL-DCPE	NaN/**1.00**	NaN/**1.00**	NaN	1.14	23.82
MRCE	NaN/0.46	0.00/0.46	NaN	5.07×10^9	9187.28
CAPME	NaN/**1.00**	NaN/**1.00**	NaN	24.00	7255.42
SEP.L+G	NaN/0.98	0.00/0.98	NaN	**0.49**	19.26
SEP.SSL+SSG	NaN/**1.00**	0.00/**1.00**	NaN	1.13	**3.22**

NOTE: NaN indicates that the specified quantity was undefined, either because no nonzero estimates were returned or because there were truly no nonzero parameters (Simulation 4). Best results are bolded.

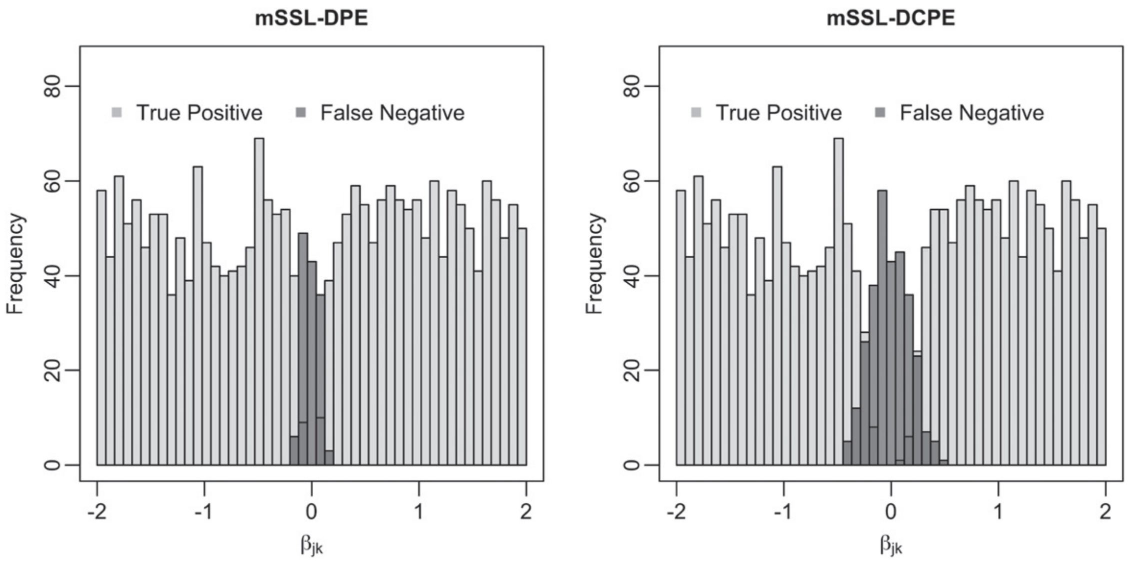

Figure 1. Histograms of the nonzero $\beta_{j,k}$ values correctly selected (i.e., true positives, in light gray) and nonzero $\beta_{j,k}$ values incorrectly estimated as zero (i.e., false negatives, in dark gray) by mSSL-DPE and mSSL-DCPE run on a dataset from Simulation 1 ($n = 400, p = 500, q = 25, \rho = 0.9$). Predictably, both mSSL-DPE and mSSL-DCPE correctly select larger $\beta_{j,k}$ values and are both unable to select small nonzero $\beta_{j,k}$'s. Compared to mSSL-DCPE, however, mSSL-DPE is able to detect smaller signals, as indicated by the tighter concentration of the false negative distribution in the left panel.

develop some intuition about why mSSL-DPE may be able to recover smaller signals than mSSL-DCPE in Appendix C in the online supplement. Roughly speaking, when $\omega_{k,k'}$ is nonzero the value of $z_{j,k}$ in Equation (5) can be larger in magnitude, making it easier for some small $\beta_{j,k}$ values to overcome the threshold.

Finally we must address Simulations 3 and 4, in which mSSL-DPE appears to perform exceptionally poorly. On closer inspection, in all of the replications, mSSL-DPE stabilized immediately at a rather dense estimate of B that left very little residual variance and produced an Ω with massive diagonal entries. As it turns out, the log-posterior evaluated at this estimate with $(\lambda_0, \xi_0) = (\lambda_0^{(L)}, \xi_0^{(L)})$ was considerably smaller than the log-posterior evaluated at mSSL-DCPE's estimate. In other words, mSSL-DCPE was able to escape the "dense B—unstable, diagonal Ω" region of the parameter space and navigate to regions of higher posterior density. In Simulation 3, the truly nonzero $\omega_{k,k'}$'s were rather small and in Simulation 8, Ω was the identity. Taken together, these two simulations suggests that when $p > n$, estimating B and Ω jointly with small values of λ_0 can lead to suboptimal estimates. In practice, we recommend running both mSSL-DPE and mSSL-DCPE and reporting results of whichever estimate has higher log-posterior.

4. Multivariate Analysis of the Football Safety Data

In a recent observational study, Deshpande et al. (2017) studied the effect of playing high school football on later-life cognitive and mental health using data from the Wisconsin Longitudinal Study (WLS), which has followed 10,317 people since they graduated from a Wisconsin high school in 1957. In addition to an indicator of high school football participation, the WLS contains a rich set of baseline variables measured in adolescence that may be associated with later-life health, socio-economic outcomes measured in the mid-1907s, when subjects were in their mid-to-late 30s, and results from a battery of cognitive, psychological, and behavioral tests conducted in 1993, 2004, and 2011, when subjects were approximately 54, 65, and 72 years old. Deshpande et al. (2017) took a univariate approach, analyzing each outcome separately, and found no evidence of a harmful effect of playing high school football on any outcome considered, after carefully adjusting for several important confounders. We now revisit the dataset of Deshpande et al. (2017) from a full multivariate perspective with mSSL-DPE and mSSL-DCPE. Our more powerful multivariate methodology not only confirms the main findings of their analysis but also provides new insight into the residual inter-dependence of the cognitive, psychological, and socio-economic outcomes that was otherwise unavailable in their univariate analysis.

To isolate the effect of playing football, Deshpande et al. (2017) began by creating matched sets containing one football player and one or more control subjects, or one control subject and one or more football players, using full matching with a propensity score caliper. These matched sets optimally balance the distribution of each baseline variable between football players and controls. They then regressed several standardized cognitive, psychological, behavioral, and socio-economic outcomes onto the indicator of football participation, the baseline covariates, and indicator variables for matched set inclusions. This allowed them to estimate the effect of playing football with the associated partial slope. The combination of full matching and model-based covariate adjustment has been shown to remove biases due to residual covariate imbalance (Cochran and Rubin 1973; Silber et al. 2001) in an efficient and robust fashion (see, e.g., Rosenbaum 2002; Hansen 2004; Rubin 1973, 1979).

The cognitive outcomes considered included scores on Letter Fluency (LF), Immediate Word Recall (IWR), Delayed Word Recall (DWR), Digit Ordering (DO), WAIS Similarity (SIM), and Number Series (NS) tests. All of these tests were administered in both 2003 and 2011, except for SIM which was also administered in 1993 and NS which was only administered in 2011. The psychological and behavioral outcomes included scores on the Center for Epidemiological Studies-Depression scale (CES-D), Anger Index (ANG), Hostility Index (HOS), and Anxiety Index (ANX). CES-D and HOS scores were available from 1993, 2003, and 2011, while ANG and ANX scores were available only in 2003 and 2011. The socio-economic and education outcomes included occupational prestige scores (SEI) for jobs held in 1964, 1970, 1974, and 1975, number of weeks worked (WW) in 1974, earnings (EARN) in 1974, and number of years of education completed by 1974.

We now focus on the $n = 448$ subjects with all available outcomes. Of these 448 subjects, 157 played high school football. Following the broad outline of Deshpande et al. (2017), we first matched football players to controls along several baseline covariates using full matching and a propensity caliper. These covariates included potential confounders like adolescent IQ, high school rank, family background, and whether the subject's parents and teachers encouraged him to go to college. Table 5 in Appendix D in the online supplement lists all of these covariates, along with the pre- and post-matching means and standardized differences for the football player and controls. In all we had 157 matched sets, each comprised of a single football player and up to 6 controls, that adequately balanced the distribution of each baseline covariate. We then standardized each of the $q = 29$ outcomes and regressed them onto the $p = 204$ predictors, which included the baseline covariates and matched set indicators. Like the simulation study in Section 3.2, we ran mSSL-DPE and mSSL-DCPE with $\mathcal{I}_\lambda$ and $\mathcal{I}_\xi$ containing 10 evenly spaced points ranging from 10 to n and $0.1n$ to n, respectively, and set $a_\theta = a_\eta = 1, b_\theta = pq = 5916$, and $b_\eta = q = 29$.

mSSL-DCPE recovered 9 nonzero $\beta_{j,k}$'s and 41 nonzero $\omega_{k,k'}$'s. mSSL-DPE recovered 14 nonzero $\beta_{j,k}$'s, 8 of which were identified by mSSL-DCPE. Additionally, mSSL-DPE identified 37 of the 41 nonzero entries in $\omega_{k,k'}$'s found by mSSL-DCPE along with several more. On closer inspection, we found that mSSL-DPE's estimated mode had a slightly larger log-posterior value than mSSL-DCPE's. In terms of estimating the effect of playing football on these outcomes, our results comport with Deshpande et al.'s (2017) findings from separate univariate analyses: neither mSSL-DPE nor mSSL-DCPE identified a nonzero $\beta_{j,k}$ corresponding to football participation. Much of the signal uncovered by mSSL-DPE is quite intuitive: adolescent IQ was a relevant predictor of scores on the digits ordering task in 2003 and the WAIS similarity task in 1993, 2003, and 2011, anticipated years of post-secondary education was a strong predictor of actual years of education completed by 1974 and the occupational prestige of subjects' job in 1964, and the occupational prestige of the jobs to which subjects aspired in high school was a relevant predictor of the occupational prestige of the jobs they actually held in 1964, 1970, 1974, and 1975. In addition, mMEVS-DPE also selected several of the indicator variables of matched set membership. These corresponded to matched sets containing subjects with similar covariates and propensity scores who had higher than average CES-D scores in 1993 (i.e., they displayed more depressive symptoms), higher than average earnings in 1974, or higher than average scores on the Anger Index in 2004.

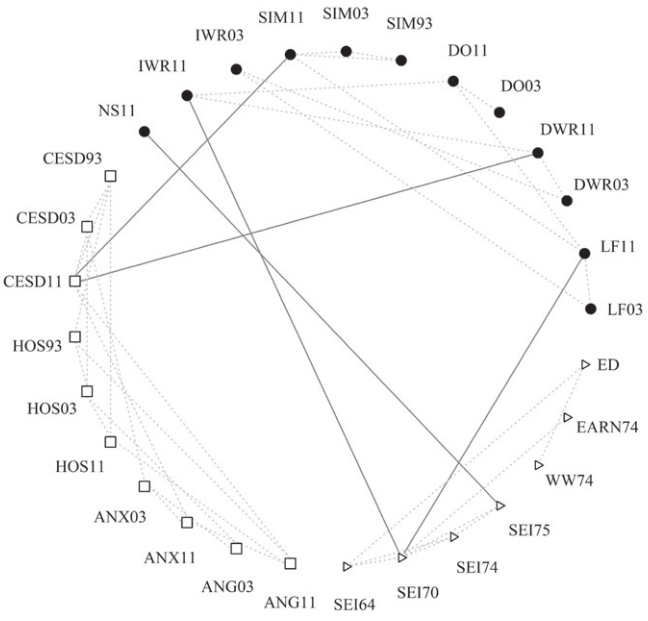

Figure 2. Estimated residual graphical model for the 29 outcomes. The number following outcome abbreviation indicates the year in which it was measured. Node shapes are used to distinguish outcome type: cognitive (circle), socio-economic/educational (triangle), and psychological (square). There are many more "within community" edges (dashed gray lines) than "between community" edges (solid black lines).

Not only does our multivariate approach confirm the main findings of Deshpande et al.'s (2017) univariate analysis, it also provides an estimate of the residual residual Gaussian graphical model G of the 29 outcomes considered, shown in Figure 2. The edges in G encode conditional dependencies between the cognitive, psychological/behavioral, and socio-economic outcomes that remain after we adjust for the measured confounders. G exhibits very strong community structure, with many more edges between outcomes of the same type (dotted gray lines) than of different type (solid black lines). This is rather interesting in light of the fact that the implicit prior on G, which models each edge as equally likely to appear, did not tend to favor any such structure.

Many of the conditional dependence relations represented in G seem intuitive: after adjusting for the covariates, we see that results from the same cognitive test administered in multiple years tended to be conditionally dependent on each other (see, e.g., the triangle formed by SIM93, SIM03, and SIM11). Additionally, we see that the CES-D scale depression scores and anger, hostility, and anxiety scores from the same year tended to be conditionally dependent as well. Perhaps more interesting are the "between community" links between outcomes of different types. After adjusting for covariates, occupational prestige of the job held in 1975 (SEI75) appears conditionally dependent on the score on the number series task in 2011 (NS11), while the scores on both the CES-D scale and letter fluency test (CESD11 and LF11) are conditionally dependent on the similarity test result in 2011 (SIM11).

5. Discussion

In this article, we have built on Ročková and George (2014)'s and Ročková and George's (2018) deterministic spike-and-slab formulation of Bayesian variable selection for univariate linear regression to develop a full joint procedure for simultaneous variable and covariance selection problem in multivariate linear regression models. We proposed and deployed an ECM algorithm within a path-following scheme to identify the modes of several posterior distributions, corresponding to different choices of spike distributions. This dynamic exploration of several posteriors is in marked contrast to MCMC, which attempts to characterize a single posterior. In our simulation experiments and analysis of the football safety data, the modal estimates identified by our dynamic posterior exploration stabilized, allowing us to report a single estimate out of the many we computed without the need for cross-validation. While there is no general guarantee that these trajectories will stabilize, stabilization provides a useful self-check: if it occurs, one can safely report the final model identified and if not, one may consider additional larger λ_0 and ξ_0 values and continue exploring.

To negotiate the dynamically changing multimodal environment, we have focused on modal estimation, at the cost of temporarily sacrificing full uncertainty quantification and posterior inference. Assessing the variability in the estimates of mSSL-DPE remains an important problem. One could potentially run a general MCMC simulation starting from the final mSSL-DPE estimate.

As anticipated by results in Ročková and George (2014) and Ročková and George (2018), our procedure tends to outperform procedures that use cross-validation to select regularization penalties. A key driver of the improvement is the hierarchical modeling of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, which facilitated our ECM algorithm's selective shrinkage. While we have here focused only on the simplest exchangeable models of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, it will be interesting to incorporate more thoughtfully structured sparsity within our framework. For instance, if the covariates displayed a known grouping structure, we may introduce a separate θ parameter for each group with little additional computational overhead.

Supplementary Materials

R package mSSL: An R package that implements mSSL-DPE and mSSL-DCPE. The package may also be installed from *https://github.com/skdeshpande91/multivariate_SSL/*.

examples.tar.gz Compressed files containing R scripts for replicating examples and simulations in the main text.

Appendices Appendices A–D as cited in the main text.

Acknowledgments

We would also like to thank the editor, associate editor, and referees for their very helpful suggestions.

Funding

The authors gratefully acknowledge NSF grant DMS-1406563, a Simons Fellowship from the Isaac Newton Institute at the University of Cambridge, and the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business.

References

Abegaz, F., and Wit, E. (2013), "Sparse Time Series Chain Graph Models for Reconstructing Genetic Networks," *Biostatistics*, 14, 586–599. [2]

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. (2008), "Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data," *Journal of Machine Learning Research*, 9, 485–516. [1]

Banerjee, S., and Ghosal, S. (2015), "Bayesian Structure Learning in Graphical Models," *Journal of Multivariate Analysis*, 136, 147–162. [2,3]

Bhadra, A., and Mallick, B. K. (2013), "Joint High-Dimensional Bayesian Variabile and Covariance Selection With an Application to eQTL Analysis," *Biometrics*, 69, 447–457. [2]

Breiman, L., and Friedman, J. H. (1997), "Predicting Multivariate Responses in Multiple Linear Regression," *Journal of the Royal Statistical Society*, Series B, 59, 3–54. [1]

Brown, P. J., Vannucci, M., and Fearn, T. (1998), "Multivariate Bayesian Variable Selection and Prediction," *Journal of the Royal Statistical Society*, Series B, 60, 627–641. [2]

Cai, T. T., Li, H., Liu, W., and Xie, J. (2013), "Covariate-Adjusted Precision Matrix Estimation With an Application in Genetical Genomics," *Biometrika*, 100, 139–156. [2,6]

Carvalho, C. M., Massam, H., and West, M. (2007), "Simulation of Hyper-Inverse Wishart Distributions in Graphical Models," *Biometrika*, 94, 647–659. [2]

Cochran, W. G., and Rubin, D. B. (1973), "Controlling Bias in Observational Studies: A Review," *Sankhya*, 35, 417–446. [9]

Dawid, A., and Lauritzen, S. (1993), "Hyper Markov Laws in the Statistical Analysis of Decomposable Graphical Models," *Annals of Statistics*, 21, 1272–1317. [2]

Dempster, A. P. (1972), "Covariance Selection," *Biometrics*, 28, 157–175. [1]

Deshpande, S. K., Hasegawa, R. B., Rabinowitz, A. R., Whyte, J., Roan, C. L., Tabatabatei, A., Baiocchi, M., Karlawish, J. H., Master, C. L., and Small, D. S. (2017), "Association of Playing High School Football With Cognition and Mental Health Later in Life," *JAMA Neurology*, 74, 909–918. [2,9,10]

Fan, J., and Li, R. (2001), "Variable Selection via Noncave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1]

Friedman, J., Hastie, T., and Tibshirani, R. (2008), "Sparse Inverse Covariance Estimation With the Graphical Lasso," *Biostatistics*, 9, 432–441. [1,5,6]

——— (2010), "Reguarilzation Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22. [7]

——— (2018), "glasso: Graphical Lasso: Estimation of Gaussian Graphical Models," R Package Version 1.10. [7]

Friedman, J. H., Hastie, T., Höfling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *Annals of Applied Statistics*, 1, 302–332. [4]

Gan, L., Narisetty, N. N., and Liang, F. (2018), "Bayesian Regularization of Graphical Models With Unequal Shrinkage," *Journal of the American Statistical Association*. doi: 10.1080/01621459.2018.1482755 [3,5]

George, E. I., and R. E. McCulloch (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889. [2]

Hansen, B. B. (2004), "Full Matching in an Observational Study of Coaching for the SAT," *Journal of the American Statistical Association*, 99, 609–618. [9]

Hsieh, C.-J., Sustik, M. A., Dhillon, I. S., and Ravikumar, P. (2014), "QUIC: Quadratic Approximation for Sparse Inverse Covariance Estimation," *Journal of Machine Learning Research*, 15, 2911–2947. [5]

Jones, B., Carvalho, C. M., Dobra, A., Hans, C., Carter, C., and West, M. (2005), "Experiments in Stochastic Computation for High-Dimensional Graphical Models," *Statistical Science*, 20, 388–400. [2]

Lee, W., and Liu, Y. (2012), "Simultaneous Multiple Response Regression and Inverse Covariate Matrix Estimation via Penalized Gaussian Maximum Likelihood," *Journal of Multivariate Analysis*, 111, 241–255. [2]

Li, Z. R., and T. McCormick (2017), "An Expectation Conditional Maximizaiton Algorithm for Gaussian Graphical Models," arXiv no. 1709.06970. [5]

Meng, X.-L., and Rubin, D. B. (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278. [3]

Mitchell, T., and Beauchamp, J. (1988), "Bayesian Variable Selection in Linear Regression," *Journal of the American Statistical Association*, 83, 1023–1032. [2]

Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2011), "Support Union Recovery in High-Dimensional Multivarate Regression," *Annals of Statistics*, 39, 1–47. [1]

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010), "Regularized Multivarate Regression for Identifying Master Predictors With Application to Integrative Genomics Study of Breast Cancer," *Annals of Applied Statistics*, 4, 53–77. [1]

R Core Team (2017), *R: A Language and Environment for Statsitical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [6]

Richardson, S., Bottolo, L., and Rosenthal, J. S. (2010), "Bayesian Models for Sparse Regression Analysis of High Dimensional Data," *Bayesian Statistics*, 9, 539–569. [2]

Ročková, V., and George, E. I. (2014), "EMVS: The EM Approach to Bayesian Variable Selection," *Journal of the American Statistical Association*, 109, 828–846. [2,10]

——— (2018), "The Spike-and-Slab LASSO," *Journal of the American Statistical Association*, 113, 431–444. [2,4,5,6,7,10]

Rosenbaum, P. R. (2002), *Observational Studies*, New York: Springer. [9]

Rothman, A. J. (2017), "MRCE: Multivariate Regression With Covariance Estimation," R Package Version. [6]

Rothman, A. J., Levina, E., and Zhu, J. (2010), "Sparse Multivariate Regression With Covariance Estimation," *Journal of Computational and Graphical Statistics*, 19, 947–962. [1,2,6]

Rubin, D. B. (1973), "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies," *Biometrics*, 29, 185–203. [9]

——— (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 74, 318–328. [9]

Scott, J. G., and Berger, J. O. (2010), "Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem," *Annals of Statistics*, 38, 2587–2619. [3]

Silber, J. H., Rosenbaum, P. R., Trudeau, M. E., Evan-Shoshan, O., Chen, W., Zhang, X., and Mosher, R. E. (2001), "Multivariate Matching and Bias Reduction in the Surgical Outcomes Study," *Medical Care*, 39, 1048–1064. [9]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1]

Turlach, B. A., Venables, W. N., and Wright, S. J. (2005), "Simultaneous Variable Selection," *Technometrics*, 47, 349–363. [1]

Uhler, C., Lenkoski, A., and Richards, D. (2018), "Exact Formulas for the Normalizing Constants of Wishart Distributions for Graphical Models," *Annals of Statistics*, 46, 90–118. [2]

Wang, H. (2015), "Scaling It Up: Stochastic Search Structure Learning in Graphical Models," *Bayesian Analysis*, 10, 351–377. [2,3]

Witten, D. M., Friedman, J. H., and Simon, N. (2011), "New Insights and Faster Computations for the Graphical Lasso," *Journal of Computational and Graphical Statistics*, 20, 892–900. [5]

Yin, J., and Li, H. (2011), "A Sparse Conditional Gaussian Graphical Model for Analysis of Genetical Genomics Data," *Annals of Applied Statistics*, 5, 2630–2650. [2]

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94, 19–35. [1]

Zellner, A. (1962), "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias," *Journal of the American Statistical Association*, 57, 348–368. [7]

Zhang, C.-H. (2010), "Nearly Unbaised Variable Selection Under Minimax Concave Penalty," *Annals of Statistics*, 38, 894–942. [1]

Zhang, C.-H., and Zhang, T. (2012), "A General Theory of Concave Regularization for High-Dimensional Sparse Estimation Problems," *Statistical Science*, 27, 576–593. [4]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1427. [1]