Biometrics 71, 354–363
June 2015

DOI: 10.1111/biom.12292

Multivariate Sparse Group Lasso for the Multivariate Multiple Linear Regression with an Arbitrary Group Structure

Yanming Li,[1] Bin Nan,[1,*] and Ji Zhu[2]

[1]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[2]Department of Statistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[*]*email:* bnan@umich.edu

Summary. We propose a multivariate sparse group lasso variable selection and estimation method for data with high-dimensional predictors as well as high-dimensional response variables. The method is carried out through a penalized multivariate multiple linear regression model with an arbitrary group structure for the regression coefficient matrix. It suits many biology studies well in detecting associations between multiple traits and multiple predictors, with each trait and each predictor embedded in some biological functional groups such as genes, pathways or brain regions. The method is able to effectively remove unimportant groups as well as unimportant individual coefficients within important groups, particularly for large p small n problems, and is flexible in handling various complex group structures such as overlapping or nested or multilevel hierarchical structures. The method is evaluated through extensive simulations with comparisons to the conventional lasso and group lasso methods, and is applied to an eQTL association study.

Key words: Coordinate descent algorithm; eQTL; Genetic association; High-dimensional data; Oracle inequalities; Sparsity.

1. Introduction

Genomic association studies with a single phenotype have been widely studied. Such association studies often encounter high-dimensional predictors with sparsity, that is, only a small number of predictors are associated with the response. To select truly associated predictors, it is necessary to use regularization penalties to shrink the coefficients of irrelevant predictors to exactly zero. Popular penalties for regression models with a univariate response include the lasso (Tibshirani, 1996), the adaptive lasso (Zou, 2006), the elastic net (Zou and Hastie, 2005), and the smoothly clipped absolute deviation (Fan and Li, 2001), among many others.

An important characteristic of high-dimensional genomic predictors is the intrinsic group structures. For example, the DNA markers, also known as single nucleotide polymorphisms (SNPs), can often be grouped into genes, and genes can be grouped into biological pathways. Such grouping strategies have been applied successfully to genomic studies in rare variant detection (Zhou et al., 2010; Biswas and Lin, 2012). For group variable selection, Yuan and Lin (2006) proposed the group lasso method for the univariate response case. It penalizes the L_2 norm of each predictor group and selects important groups in an "all-in-all-out" fashion. That is, all the predictors in a group are included or excluded simultaneously. However, in real applications, this is rarely the case. Oftentimes, not all the variables in an important group are important. For example, a gene associated with a certain complex trait does not mean that all the variants within the gene are causal, and a pathway that regulates certain gene expressions does not necessarily indicate that all its components have regulatory effects. Recent efforts have been made to select both important groups and important within-group signals simultaneously. Huang et al. (2009) and Zhou and Zhu (2010) adopted a L_γ, $0 < \gamma < 1$, penalty to select important groups while removing unimportant variables within them; Zhou et al. (2010) used a penalized logistic regression with a mixed L_1/L_2 penalty to select both common and rare variants in a genome-wide association study; and Simon et al. (2013) proposed the sparse group lasso for selecting both important groups and within group predictors. However, all the above methods concern a univariate response.

Many other genomic data analyses focus on investigating the associations between high dimensional response variables and high-dimensional covariates, such as gene-gene associations (Park and Hastie, 2008; Zhang et al., 2010), protein–DNA associations (Zamdborg and Ma, 2009) and brain fMRI-DNA (or gene) associations (Stein et al., 2010). Oftentimes pairwise associations are calculated in such studies. For example, many multivariate genome-wide association studies nowadays still look for one association at a time between a single marker and a single trait, and then correct for multiple hypothesis testing (Dudoit, Shaffer, and Boldrick, 2003; Stein et al., 2010). However, when both responses and predictors are of high dimensions, most of the familywise type I error controlling procedures are usually too conservative and yield poor performance (Stein et al., 2010), and oftentimes adjusted analysis considering multiple variables simultaneously is more appropriate.

High-dimensional responses also have natural group structures very often, for example, pathway group structures for gene expression responses and brain functional regions for fMRI intensity responses. For multivariate responses,

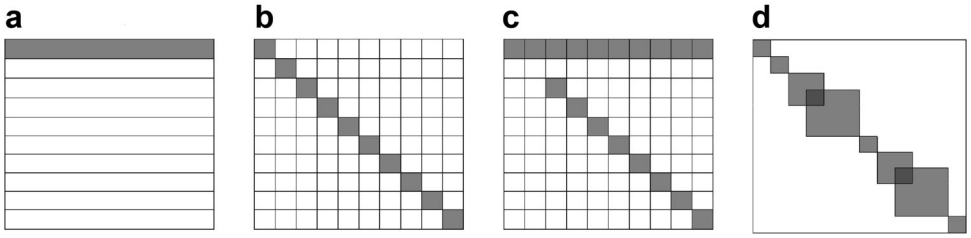

Figure 1. $\boldsymbol{B}^*$ group structures. Important groups are shaded. (a) X group structure, (b) XY group structure, (c) $X+XY$ group structure (nesting group structure) and (d) overlapping group structure.

Peng et al. (2010) adopted the mixed L_1/L_2 penalty in an orthonormal setting for identifying hub covariates in a gene regulation network; Obozinski, Wainwright, and Jordan (2011) and Bunea, She, and Wegkamp (2011) studied joint support union and joint rank selections; Lounici et al. (2011) proved oracle inequalities for multitask learning. Despite all the efforts, little focus, to our knowledge, has been put on the cases where the responses also have a group structure, whereas such cases are commonly encountered in biological studies. A possible strategy for multivariate-response analysis is to perform covariate selection for one response variable at a time. In such analysis the predictor group structure can be considered but the response group structure is overlooked.

In this article, we propose a regularization method for making a good use of the intrinsic biological group structures on both covariates and responses to facilitate a better variable selection on multivariate-response and multiple-predictor data by effectively removing unimportant blocks of regression coefficients. Both the predictor and response group structures, or in general, the block structure of the regression coefficient matrix, are assumed known. Information of many biologically confirmed group structures can be achieved from publicly available repositories, for example, RefSeq gene files from NCBI Reference Sequence Database (http://www.ncbi.nlm.nih.gov/refseq/), KEGG pathway maps from Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/), and Brodmann brain anatomic region atlas from https://surfer.nmr.mgh.harvard.edu/fswiki/BrodmannAreaMaps. The proposed method can handle cases where the number of variables in either responses or predictors is much greater than the sample size, and complex group structures such as overlapping groups where a variable belongs to multiple groups. The estimators enjoy finite sample oracle bounds for the prediction error, the estimation error, and the estimated sparsity of the regression coefficient matrix. Extensive simulations show that the proposed method outperforms competitive regularization methods. We applied the proposed method to a yeast gene expression quantitative loci (eQTL) study, where the numbers of gene expression responses and genetic marker predictors are both much larger than the sample size. The gene expressions are grouped into biological pathways and the genetic markers are grouped into genes. We demonstrate by considering both group structures that the proposed method generates a much more interpretable and predictive eQTL network between the gene expressions and genetic markers, comparing with several other commonly used regularized approaches.

2. Multivariate Linear Model with Arbitrary Grouping

We consider the multivariate linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{W}, \tag{1}$$

where $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_q) \in \mathbb{R}^{n\times q}$ is the response matrix of n samples and q variables, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n\times p}$ is the covariate matrix of n samples and p variables, $\boldsymbol{B} = (\beta_{jk})_{p\times q} \in \mathbb{R}^{p\times q}$ is the coefficient matrix and $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_q) \in \mathbb{R}^{n\times q}$ is the matrix of error terms with each $\boldsymbol{w}_k \sim N(0, \sigma_k^2 \boldsymbol{I}_{n\times n})$, $k = 1, \ldots, q$. Assume $\boldsymbol{Y}$ and $\boldsymbol{X}$ are centered so that there is no intercept in $\boldsymbol{B}$. We adopt the notational convention that the column vectors of $\boldsymbol{X}$ are indexed by j, the column vectors of $\boldsymbol{Y}$ and $\boldsymbol{W}$ are indexed by k, and the samples are indexed by i.

Assume $\boldsymbol{B}$ contains G groups, and each group, denoted as $\boldsymbol{B}_g$ where $g \in \{1, \ldots, G\}$, is a subset of two or more elements in $\boldsymbol{B}$. We denote the group structure by $\mathcal{G} = \{\boldsymbol{B}_1, \ldots, \boldsymbol{B}_G\}$. We use $\boldsymbol{B}$ or $\boldsymbol{B}_g$ to denote either the set of all their elements or the numerical values of all their elements, depending on the context, which should not cause any confusion. Figure 1 illustrates a few examples of group structures, where each highlighted block indicates an important group in $\mathcal{G}$ and each figure may represent several different group structures. Note that the group structures considered in this article are predefined by biological functions, such as gene or pathways. Also note that the union of all groups in $\mathcal{G}$ does not need to contain all the elements of $\boldsymbol{B}$, in other words, some β_{jk} may not belong to any group. We say $\boldsymbol{B}_{g_1}$ is *nested* in $\boldsymbol{B}_{g_2}$ if $\boldsymbol{B}_{g_1} \subset \boldsymbol{B}_{g_2}$; $\boldsymbol{B}_{g_1}$ and $\boldsymbol{B}_{g_2}$ are *overlapping* if $\boldsymbol{B}_{g_1} \cap \boldsymbol{B}_{g_2}$ is not empty. Obviously, nested groups are a special case of overlapping. A group structure with overlapping groups is common in biological studies. For example, when grouping genetic variants according to genes or pathways, different genes or pathways can overlap.

Though the proposed method works for an arbitrary group structure $\mathcal{G}$ on $\boldsymbol{B}$, in real applications, a biologically meaningful group structure on $\boldsymbol{B}$ is usually introduced from the group structures of both predictors and responses. Specifically, suppose $\boldsymbol{X}$ has m_1 column groups and $\boldsymbol{Y}$ has m_2 column groups, then they yield $m_1 \times m_2$ intersection block groups on $\boldsymbol{B}$. We denote this intersection block group structure by $\mathcal{G}_{XY}$, the row block group structure only determined by the predictor groups by $\mathcal{G}_X$, and the nested group structure containing all groups in $\mathcal{G}_{XY}$ and $\mathcal{G}_X$ by $\mathcal{G}_{XY} \cup \mathcal{G}_X$. In the eQTL association study, a nonzero group in $\mathcal{G}_{XY}$ indicates that the corresponding

gene group has SNPs associated with expressions in the corresponding pathway group. A nonzero group in $\mathcal{G}_X$ indicates that the corresponding gene group has an effect on some or all of the expressions.

For an arbitrary group structure $\mathcal{G}$ with G groups, let $\sum_{g=1}^{G} \|\boldsymbol{B}_g\|_2$ be the total sum of L_2 norms of every group in $\mathcal{G}$, where $\|\boldsymbol{B}_g\|_2^2 = \sum_{\beta_{jk}\in\boldsymbol{B}_g} \beta_{jk}^2$. The group L_2 norm reduces to the Frobenius norm $\|\boldsymbol{A}\|_2 = \left\{\mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})\right\}^{1/2}$ for a matrix group $\boldsymbol{A}$ and to the vector L_2 norm $\|\boldsymbol{a}\|_2 = \left\{\boldsymbol{a}^{\mathrm{T}}\boldsymbol{a}\right\}^{1/2}$ for a vector group $\boldsymbol{a}$. Proofs of theoretical results in the following sections are provided in the web-based Supplementary Materials.

3. The Regularization Method and Its Properties

3.1. *The Multivariate Sparse Group Lasso*

For an arbitrary group structure $\mathcal{G}$ on $\boldsymbol{B}$, to simplify the notation, we denote $\{g : \boldsymbol{B}_g \in \mathcal{G}\}$ by $\{g \in \mathcal{G}\}$ as long as it does not cause any confusion. For $j = 1, \ldots, p$ and $k = 1, \ldots, q$, let $\lambda_{jk} \geq 0$ be the adaptive lasso tuning parameter for β_{jk}, with $\lambda_{jk} = 0$ if β_{jk} is not penalized. Let $\lambda_g \geq 0$ be the adaptive tuning parameter for group $\boldsymbol{B}_g \in \mathcal{G}$, with $\lambda_g = 0$ if group $\boldsymbol{B}_g$ is not penalized. We consider the following penalized optimization problem for a general regularized multivariate multiple linear regression:

$$\arg\min_{\boldsymbol{B}} \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_2^2 + \sum_{1\leq j\leq p, 1\leq k\leq q} \lambda_{jk}|\beta_{jk}| + \sum_{g\in\mathcal{G}} \lambda_g \|\boldsymbol{B}_g\|_2, \tag{2}$$

where the L_2 penalty term aims to shrink unimportant groups to zero and the L_1 penalty term aims to shrink unimportant entries within an important group to zero. We call it the multivariate sparse group lasso (MSGLasso). We exclude the trivial case that $\lambda_g = 0$ for all $g \in \mathcal{G}$ and $\lambda_{jk} = 0$ for all j, k. To better understand the solution to (2), we develop the following theorem for β_{jk} when all other elements in $\boldsymbol{B}$ are fixed.

THEOREM 1. *For an arbitrary group structure $\mathcal{G}$ on $\boldsymbol{B}$, let $\hat{\boldsymbol{B}}$ be the solution to (2) and $\hat{\beta}_{jk}$ be its jkth element. If for some group $\boldsymbol{B}_{g0} \in \mathcal{G}$ with a tuning parameter λ_{g0},*

$$\sqrt{\sum_{\{jk:\ \beta_{jk}\in\boldsymbol{B}_{g0}\}} (|S_{jk}|/n - \lambda_{jk})_+^2} \leq \lambda_{g0}, \tag{3}$$

then $\hat{\beta}_{jk} = 0$ for every $\beta_{jk} \in B_{g0}$. Otherwise, $\hat{\beta}_{jk}$ satisfies

$$\hat{\beta}_{jk} = \frac{\mathrm{sgn}(S_{jk})\,(|S_{jk}| - n\lambda_{jk})_+}{\|x_j\|_2^2 + n\sum_{\{g\in\mathcal{G}:\ \beta_{jk}\in\boldsymbol{B}_g\}} \lambda_g/\|\hat{\boldsymbol{B}}_g\|_2}, \tag{4}$$

where $S_{jk} = \boldsymbol{x}_j^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{B}}_{(-j)})_{\cdot k}$ with $\hat{\boldsymbol{B}}_{(-j)}$ being the jth row of $\hat{\boldsymbol{B}}$ replaced by zeros, the subscript $\cdot k$ refers to the kth column of a matrix, and $a_+ = a$ if $a > 0$ and 0 otherwise.

Note that Theorem 1 is a general solution form and applies to arbitrary group structures. If there is no group structure assigned on $\boldsymbol{B}$, then $\mathcal{G}$ becomes an empty set and (4) reduces to the lasso solution; If $\lambda_{jk} = 0$ for all j, k, then (4) and (3) provide the group lasso solution. It is of interest to consider certain special group structures that are intuitive and commonly used in many applications. Specifically, we consider model (2) with the following four group structures: (I) $\mathcal{G} = \emptyset$, no group structure assigned on $\boldsymbol{B}$; (II) $\mathcal{G}_X$; (III) $\mathcal{G}_{XY}$; (IV) $\mathcal{G}_{XY} \cup \mathcal{G}_X$. The corresponding optimization problems become

$$\arg\min_{\boldsymbol{B}} \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_2^2 + \lambda|\boldsymbol{B}|_1, \tag{5}$$

$$\arg\min_{\boldsymbol{B}} \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_2^2 + \lambda|\boldsymbol{B}|_1 + \lambda_1 \sum_{g_1\in\mathcal{G}_X} \omega_{g_1}^{1/2}\|\boldsymbol{B}_{g_1}\|_2, \tag{6}$$

$$\arg\min_{\boldsymbol{B}} \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_2^2 + \lambda|\boldsymbol{B}|_1 + \lambda_2 \sum_{g_2\in\mathcal{G}_{XY}} \omega_{g_2}^{1/2}\|\boldsymbol{B}_{g_2}\|_2, \tag{7}$$

$$\arg\min_{\boldsymbol{B}} \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}\|_2^2 + \lambda|\boldsymbol{B}|_1 + \lambda_1 \sum_{g_1\in\mathcal{G}_X} \omega_{g_1}^{1/2}\|\boldsymbol{B}_{g_1}\|_2 + \lambda_2 \sum_{g_2\in\mathcal{G}_{XY}} \omega_{g_2}^{1/2}\|\boldsymbol{B}_{g_2}\|_2, \tag{8}$$

where $|\boldsymbol{B}|_1 = \sum_{jk}|\beta_{jk}|$ is the L_1 norm of $\boldsymbol{B}$, and ω_{g_1} and ω_{g_2} are some weights, in particular, the group sizes. The tuning parameters $\lambda_{jk} = \lambda$ for all lasso penalties, $\lambda_g = \lambda_1\omega_{g_1}^{1/2}$ if $g \in \mathcal{G}_X$, and $\lambda_g = \lambda_2\omega_{g_2}^{1/2}$ if $g \in \mathcal{G}_{XY}$.

In the remaining of this article, we call (5) the *Lasso* model, (6) the *Lasso+X* model, (7) the *Lasso+XY* model, and (8) the *Lasso+X+XY* model.

Let $\hat{\boldsymbol{B}}_L$, $\hat{\boldsymbol{B}}_{LX}$, $\hat{\boldsymbol{B}}_{LXY}$, and $\hat{\boldsymbol{B}}_{LXXY}$ be the solutions to (5), (6), (7), and (8), respectively. Their corresponding expressions from Theorem 1 further reduce to some interesting simpler forms under the orthonormal design, in particular, $\hat{\boldsymbol{B}}_{LX}$ and $\hat{\boldsymbol{B}}_{LXY}$ are just further shrinkages of $\hat{\boldsymbol{B}}_L$, and $\hat{\boldsymbol{B}}_{LXXY}$ is a further shrinkage of either $\hat{\boldsymbol{B}}_{LX}$ or $\hat{\boldsymbol{B}}_{LXY}$. We are also interested in the group lasso cases where $\lambda = 0$ in (6), (7), and (8), with their solutions denoted by $\hat{\boldsymbol{B}}_{GX}$, $\hat{\boldsymbol{B}}_{GXY}$ and $\hat{\boldsymbol{B}}_{GXXY}$, respectively. Then the main theorems in Yuan and Lin (2006) and Peng et al. (2010) become special cases.

In the eQTL example that we will analyze later, method (5) does not take the advantage of knowing the group structure. Method (6) only concerns the predictor group structure, therefore can select important gene groups. However, it ignores which pathways those genes are associated with. Method (7) considers both predictor and response group structures, therefore can select gene-to-pathway association blocks. Method (8) pertains advantages of both (6) and (7) and is more robust to misspecified group structures.

3.2. *Oracle Inequalities*

The lasso method has been shown to achieve the oracle bounds for both prediction and estimation in the multiple linear regression model, which are the error bounds one would obtain if the true model were given, see for example, Bickel,

Ritov, and Tsybakov (2009). Similar bounds also hold for a total of pq regression coefficients in the multivariate multiple linear regression model with a multivariate mixed L_1/L_2 penalty. For notational simplicity, we consider the following special case of (2) with $\lambda_{jk} = \lambda$ for all j, k:

$$\arg\min_{\boldsymbol{B}} \frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{XB}\|_2^2 + \lambda|\boldsymbol{B}|_1 + \sum_{g\in\mathcal{G}} \lambda_g \|\boldsymbol{B}_g\|_2. \tag{9}$$

We follow the method of Bickel et al. (2009). Let $J_1(\boldsymbol{B}) = \{jk : |\beta_{jk}| \neq 0\}$ be the index set of nonzero elements in $\boldsymbol{B}$, and $J_2(\boldsymbol{B}) = \{g \in \mathcal{G}, \|\boldsymbol{B}_g\|_2 \neq 0\}$ be the index set of nonzero groups in $\mathcal{G}$. Define $M_1(\boldsymbol{B}) = \sum_{jk} I(\beta_{jk} \neq 0) = |J_1(\boldsymbol{B})|$ and $M_2(\boldsymbol{B}) = \sum_{g\in\mathcal{G}} I(\|\boldsymbol{B}_g\|_2 \neq 0) = |J_2(\boldsymbol{B})|$. For any matrix $\boldsymbol{\Delta} \in \mathbb{R}^{p\times q}$ and any given index set $J_1 \subseteq \{jk : 1 \le j \le p, 1 \le k \le q\}$, denote $\boldsymbol{\Delta}_{J_1}$ the projection of $\boldsymbol{\Delta}$ on the index set J_1, that is the matrix with the same elements of $\boldsymbol{\Delta}$ on coordinates J_1 and zeros on the complementary coordinates J_1^c. Also for any group index set $J_2 \subseteq \{1, \ldots, |\mathcal{G}|\}$, denote $\boldsymbol{\Delta}_{J_2}$ the set of projection of $\boldsymbol{\Delta}$ on each of $\{\boldsymbol{B}_g : g \in J_2\}$, that is $\boldsymbol{\Delta}_{J_2} = \{\boldsymbol{\Delta}_{\boldsymbol{B}_g} : g \in J_2\}$. Denote $M_1(\boldsymbol{B}) = r$ and $M_2(\boldsymbol{B}) = s$. We then impose a restricted eigenvalue assumption for the multivariate linear regression model with a multivariate mixed L_1/L_2 penalty, which leads to the desirable oracle inequalities.

ASSUMPTION 1. *Let $J_1 \subseteq \{jk : 1 \le j \le p, 1 \le k \le q\}$ and $J_2 \subseteq \{1, \ldots, |\mathcal{G}|\}$ be any index sets that satisfy $|J_1| \le r$ and $|J_2| \le s$. Let $\tilde{\rho} = \{\rho_g : g \in \mathcal{G}\}$ be a set of positive numbers. Then for any nontrivial matrix $\boldsymbol{\Delta} \in \mathbb{R}^{p\times q}$ that satisfies*

$$|\boldsymbol{\Delta}_{J_1^c}|_1 + 2\sum_{g\in J_2^c} \rho_g \|\boldsymbol{\Delta}_{B_g}\|_2 \le 3|\boldsymbol{\Delta}_{J_1}|_1 + 2\sum_{g\in J_2} \rho_g \|\boldsymbol{\Delta}_{B_g}\|_2,$$

the following minimums exist and are positive:

$$\kappa_1(r, s, \tilde{\rho}) = \min_{J_1, J_2, \boldsymbol{\Delta}\neq 0} \frac{\|\boldsymbol{X\Delta}\|_2}{n^{1/2}\|\boldsymbol{\Delta}_{J_1}\|_2} > 0,$$

$$\kappa_2(r, s, \tilde{\rho}) = \min_{J_1, J_2, \boldsymbol{\Delta}\neq 0} \frac{\|\boldsymbol{X\Delta}\|_2}{n^{1/2}\|\boldsymbol{\Delta}_{J_2}\|_2} > 0.$$

THEOREM 2. *Consider model (9). Let $\boldsymbol{B}^*$ be the true coefficient matrix. Assume each column of the error matrix, $\boldsymbol{w}_k$, follows a multivariate normal distribution $N(0, \sigma_k \boldsymbol{I}_n)$, and all the diagonal elements of the matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/n$ are equal to 1. Suppose $M_1(\boldsymbol{B}^*) = r$ and $M_2(\boldsymbol{B}^*) = s$. Let $\psi_{\max}$ be the largest eigenvalue of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/n$, $\sigma = \max\{\sigma_1, \cdots, \sigma_q\}$, $\lambda_g = \rho_g\lambda$ for $g \in \mathcal{G}$, $\rho = \min\{1, \rho_g; g \in \mathcal{G}\}$, c be the maximum number of duplicates of a coefficient in overlapping groups in $\mathcal{G}$, and*

$$\lambda = 2\sigma A\{\log(pq)/n\}^{1/2}$$

for some constant $A > 2^{1/2}$. Furthermore, assume Assumption 1 holds with $\kappa_1 = \kappa_1(r, s, \tilde{\rho})$ and $\kappa_2 = \kappa_2(r, s, \tilde{\rho})$. Then with probability at least $1 - (pq)^{1-A^2/2}$, we have the following oracle bounds for the prediction error, the estimation error and the order of sparsity:

$$\frac{1}{n}\|\boldsymbol{X}(\hat{\boldsymbol{B}} - \boldsymbol{B}^*)\|_2^2 \le 16\lambda^2 \left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g\in J_2(\boldsymbol{B}^*)} \rho_g^2\right)^{1/2}}{\kappa_2} \right)^2,$$

$$|\hat{\boldsymbol{B}} - \boldsymbol{B}^*|_1 \le \frac{32(c+2)\sigma A}{1+\rho} \left(\frac{\log(pq)}{n}\right)^{1/2} \times \left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g\in J_2(\boldsymbol{B}^*)} \rho_g^2\right)^{1/2}}{\kappa_2} \right)^2,$$

$$M_1(\hat{\boldsymbol{B}}) \le 64\psi_{\max} \left(\frac{r^{1/2}}{\kappa_1} + \frac{\left(\sum_{g\in J_2(\boldsymbol{B}^*)} \rho_g^2\right)^{1/2}}{\kappa_2} \right)^2.$$

The mean square prediction error is bounded by a factor of order $\lambda^2 \sim \log(pq)/n$, the l_1 norm of the estimation error is bounded by a factor of order $\sqrt{\log(pq)/n}$, and the estimated order of sparsity is bounded by a constant related to Assumption 1. These results are similar to those in Bickel et al. (2009). Note that Theorem 2 will still hold for flexible λ_{jk} in (2), as long as $\lambda_{jk} > 0$ for all j, k.

4. The Mixed Coordinate Descent Algorithm

Based on Theorem 1, the zero groups can be determined according to (3) and the entries in a nonzero group can be determined by solving for the fixed point solution of (4) using a coordinate descent algorithm. The algorithm updates each coefficient coordinate β_{jk} at a step while fixing all the other coefficients at their current values. Theoretically, the coordinate descent algorithm would work if one can solve (4) for $\hat{\beta}_{jk}$ exactly. Practically, since $\hat{\beta}_{jk}$ also appears in the term $\sum_{\{g\in\mathcal{G}:\ \beta_{jk}\in\boldsymbol{B}_g,\ \|\hat{\boldsymbol{B}}_g\|_2>0\}} \lambda_g/\|\hat{\boldsymbol{B}}_g\|_2$ on the right hand side of (4), unlike lasso, a closed form solution is usually not available and numerically solving for $\hat{\beta}_{jk}$ requires iteratively updating (4), which can be time consuming. Here we propose a mixed coordinate descent algorithm, which only updates $\hat{\beta}_{jk}$ once from $\hat{\beta}_{jk}^{(m-1)}$ to $\hat{\beta}_{jk}^{(m)}$ according to (4) without iteratively solving (4). In particular, the algorithm updates $\hat{\beta}_{jk}$ by the following.

(I) If any of the groups $\boldsymbol{B}_g \in \mathcal{G}$ containing β_{jk} satisfies (3), then the entire group is estimated at zero. Otherwise $\hat{\beta}_{jk}$ will be updated according to one of the situations (II)–(IV).

(II) If all the groups containing β_{jk} satisfy $\|\hat{\boldsymbol{B}}_{g-(jk)}^{(m-1)}\|_2 = 0$ at the current step, where $\hat{\boldsymbol{B}}_{g-(jk)}^{(m-1)}$ is $\hat{\boldsymbol{B}}_g^{(m-1)}$ with its jkth element replaced by zero, then $\hat{\beta}_{jk}$ is updated by

$$\hat{\beta}_{jk}^{(m)} = \frac{\operatorname{sgn}(S_{jk}^{(m-1)}) \left(|S_{jk}^{(m-1)}| - n\sum_{\{g\in\mathcal{G}:\ \beta_{jk}\in\boldsymbol{B}_g,\ \|\hat{\boldsymbol{B}}_{g-(jk)}^{(m-1)}\|_2=0\}} \lambda_g - n\lambda_{jk} \right)_+}{\|\boldsymbol{x}_j\|_2^2}.$$

Notice that in this case (4) becomes a closed form lasso solution.

(III) If all the groups containing β_{jk} satisfy $\|\hat{\boldsymbol{B}}^{(m-1)}_{g-(jk)}\|_2 > 0$ at the current step and $\lambda_{jk} = 0$, then $\hat{\beta}^{(m-1)}_{jk}$ is updated by the group lasso formulation

$$\hat{\beta}^{(m)}_{jk} = \frac{S^{(m-1)}_{jk}}{\|x_j\|_2^2 + n\sum_{\{g\in\mathcal{G}:\ \beta_{jk}\in\boldsymbol{B}_g,\ \|\hat{\boldsymbol{B}}^{(m-1)}_{g-(jk)}\|_2>0\}} \lambda_g/\|\hat{\boldsymbol{B}}^{(m-1)}_g\|_2}.$$

Notice in this case, all the entries in $\boldsymbol{B}_g$ with $\|\hat{\boldsymbol{B}}_{g-(jk)}\|_2 > 0$ will enter as nonzero entries

(IV) If some but not all groups containing β_{jk} satisfy $\|\hat{\boldsymbol{B}}_{g-(jk)}\|_2 = 0$ at the current step, then $\hat{\beta}^{(m-1)}_{jk}$ belongs to a mixture of the lasso case (for groups with $\|\hat{\boldsymbol{B}}^{(m-1)}_{g-(jk)}\|_2 = 0$) and the group lasso case (for groups with $\|\hat{\boldsymbol{B}}^{(m-1)}_{g-(jk)}\|_2 > 0$), and it is updated as if by a mixture of the lasso and the group lasso through

$$\hat{\beta}^{(m)}_{jk} = \frac{\mathrm{sgn}(S^{(m-1)}_{jk})\left(|S^{(m-1)}_{jk}| - n\sum_{\{g\in\mathcal{G}:\ \beta_{jk}\in\boldsymbol{B}_g,\ \|\hat{\boldsymbol{B}}^{(m-1)}_{g-(jk)}\|_2=0\}} \lambda_g - n\lambda_{jk}\right)_+}{\|x_j\|_2^2 + n\sum_{\{g\in\mathcal{G}:\ \beta_{jk}\in\boldsymbol{B}_g\ \|\hat{\boldsymbol{B}}^{(m-1)}_{g-(jk)}\|_2>0\}} \lambda_g/\|\hat{\boldsymbol{B}}^{(m-1)}_g\|_2}.$$

Specifically, the algorithm is given in the following for a fixed set of values of all the tuning parameters.

Step 1. Standardize the data such that

$$\sum_{i=1}^n y_{ik} = 0, \quad \sum_{i=1}^n x_{ij} = 0, \quad \sum_{i=1}^n x_{ij}^2 = 1,$$
$$\text{for all } j \in \{1,\ldots,p\},\ k \in \{1,\ldots,q\}.$$

In our numerical examples, we also standardize $\boldsymbol{y}_k$ such that $\sum_{i=1}^n y_{ik}^2 = 1$ to minimize the impact of different scales of variations across $\boldsymbol{y}_k$ on the regression coefficients for all $k \in \{1,\ldots,q\}$.

Step 2. Set initial values for all $\hat{\beta}_{jk}$ and the iteration index $m = 1$. We use initial values $\hat{\beta}^{(0)}_{jk} = 0$ in our numerical examples.

Step 3. For a given pair (j,k), fix $\beta_{j'k'}$ at $\hat{\beta}^{(m-1)}_{j'k'}$ for all $j' \neq j$ or $k' \neq k$. Then update $\hat{\beta}^{(m-1)}_{jk}$ to $\hat{\beta}^{(m)}_{jk}$ by (I) to (IV) accordingly.

Step 4. Repeat Step 3 for all $j \in \{1,\ldots,p\}$ and $k \in \{1,\ldots,q\}$, and iterate until $\|\hat{\boldsymbol{B}}^{(m)} - \hat{\boldsymbol{B}}^{(m-1)}\|$ reaches a prespecified precision level for some norm $\|\cdot\|$. We use infinity norm in our numerical examples.

Convergence of different types of coordinate descent algorithms have been studied in the literature. Tseng (2001) provided conditions for convergence of cyclic coordinate descent algorithm with general separable objective functions. Wu and Lange (2008) proved the convergence of greedy coordinate descent algorithm with a L_2 loss and the lasso penalty. Based on Wu and Lange (2008), we show the convergence of our mixed coordinate descent algorithm which is given in the following proposition. Details are provided in the supplemental materials, where we also illustrate that the speed of convergence of our mixed coordinate descent algorithm is much faster than the coordinate descent algorithm that solves the fixed point solution to (4) with inner iterations.

PROPOSITION 1. *A sequence of coordinate estimates iteratively updated by the mixed coordinate descent algorithm converge to a global minimizer of the objective function.*

We implemented the MSGLasso and the mixed coordinate descent algorithm with C/C++ language and wrapped it into an `R` package. It is available on the web-based Supplementary Materials and will soon be upload to CRAN repository.

5. Numerical Studies

5.1. *Simulations*

In this section, we first investigate the numerical performances of *Lasso*, *Lasso+X*, *Lasso+XY*, *Lasso+X+XY* methods and their group lasso counterparts when the true coefficient matrix $\boldsymbol{B}^*$ takes a group structure of either $\mathcal{G}_X$, $\mathcal{G}_{XY}$ or $\mathcal{G}_{XY} \cup \mathcal{G}_X$. We also compare the proposed MSGLasso method with lasso and group lasso for an overlapping group structure.

All the true group structures considered in our simulations are given in Figure 1a–d. For each group structure, we consider two scenarios: (i) "all-in-all-out," where all the coefficients in an important group are important, and (ii) "not-all-in-all-out," where only a subset of coefficients in an important group are important. Specifically, we generate $\boldsymbol{B}^*$ by setting $\beta^*_{jk} = 0$ if it is from an unimportant group, and drawing its value from a uniform distribution on $[-5,-1] \cup [1,5]$ and fixing it for the simulations if it is from an important group. The sparsity of an important group in the "not all in all out" setting is randomly set between 1/4 and 1/6.

Each $\boldsymbol{B}^*$ is of dimension 200×200. For a nonoverlapping group structure, each $\boldsymbol{X}$ row group is of dimension 20×200; each $\boldsymbol{XY}$ block group is of dimension 20×20. For the overlapping group structure, the groups start on coordinates $(1, 21, 41, 61, 101, 121, 141, 181)$ and end on coordinates $(20, 40, 70, 100, 120, 150, 180, 200)$, for both $\boldsymbol{X}$ and $\boldsymbol{Y}$ variables.

Covariates $\boldsymbol{X}^{\mathrm{T}}_{i\cdot}$, $i = 1,\ldots,n$, are generated from a multivariate normal distribution $N_p(0, \boldsymbol{\Sigma}_X)$, where $\boldsymbol{\Sigma}_X = diag(\boldsymbol{\Sigma}_{g_1},\ldots,\boldsymbol{\Sigma}_{g_{10}})$ is block diagonal and each block corresponds to each group of $\boldsymbol{X}$ which has the first order autoregressive structure. Specifically, $\boldsymbol{\Sigma}_{g_i}(j,k) = \rho^{|j-k|}$ for any j,k pair from the same group, $i = 1,\ldots,10$. The error terms w_{ik} are generated from a normal distribution $N(0,\sigma^2)$, where σ^2 is chosen to yield a signal to noise ratio of 2. Finally, the responses are generated from $\boldsymbol{Y} = \boldsymbol{XB}^* + \boldsymbol{W}$.

The optimal values of tuning parameters may be selected by different criteria. Since the degrees of freedom are difficult to determine for a penalty with multiple tuning parameters, we search for the optimal tuning parameter values using a five-fold cross-validation over a wide range of candidate values. The searching process starts with the largest candidate tuning parameter values with each by itself shrinking all the coefficients to zero. The converged estimates $\hat{\boldsymbol{B}}$ obtained from the previous searching step are used as the initial values for $\boldsymbol{B}$ in the next searching step with a new set of tuning parameter values. We find it very effective in reducing the computational cost.

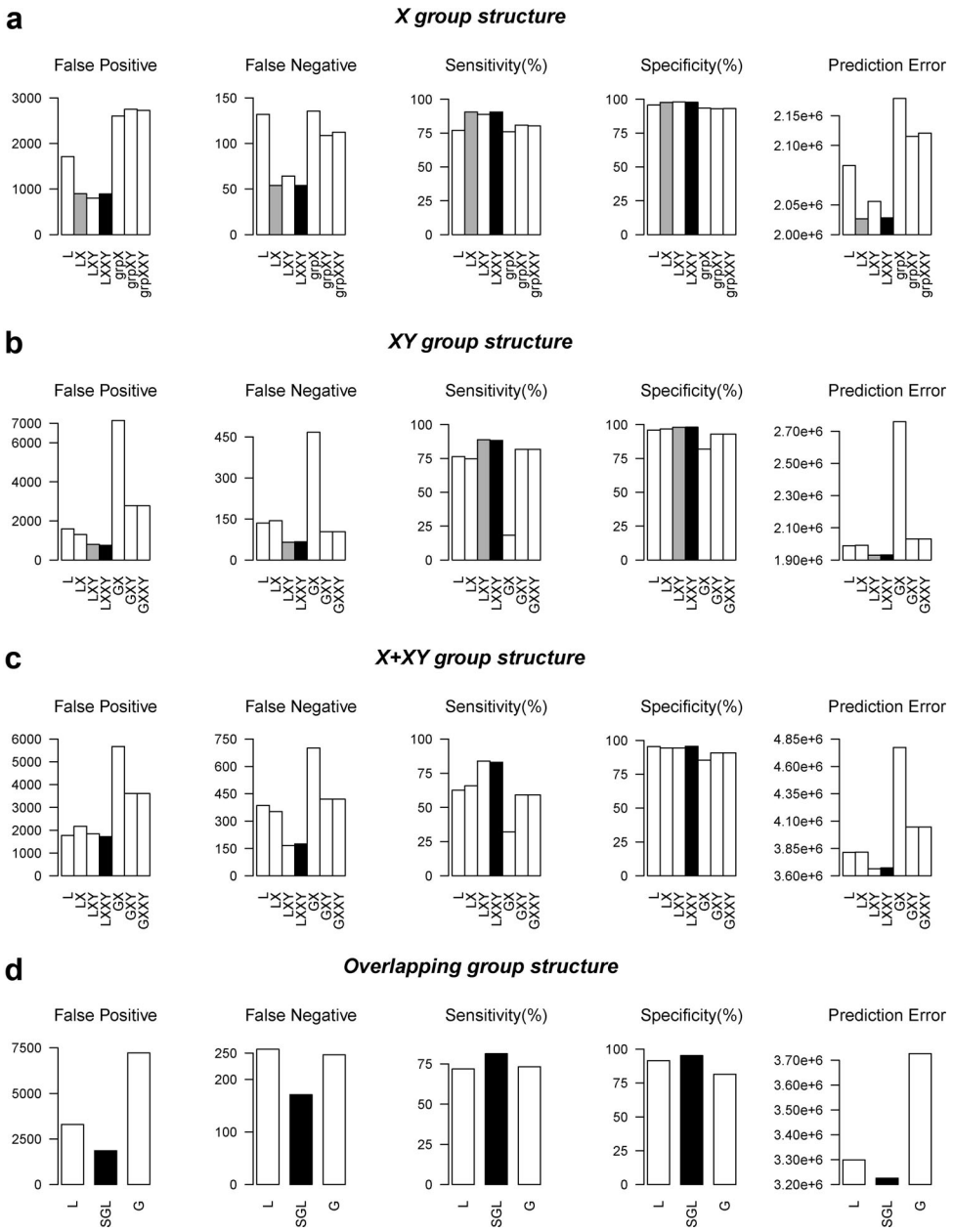

Figure 2. Simulation results, large p small n, "not all in all out" cases with $n = 100$, $p = q = 200$, and $\rho = 0.5$. SGL: the multivariate sparse group lasso; G: the multivariate group lasso.

For each simulation setup, we run a hundred replications and calculate the averages of the following quantities:

$$\text{false positives} = |\{ij \text{ pairs} : \hat{\beta}_{ij} \neq 0 \text{ and } \beta^*_{ij} = 0\}|,$$

$$\text{false negatives} = |\{ij \text{ pairs} : \hat{\beta}_{ij} = 0 \text{ and } \beta^*_{ij} \neq 0\}|,$$

$$\text{sensitivity} = \frac{|\{ij \text{ pairs} : \hat{\beta}_{ij} \neq 0 \text{ and } \beta^*_{ij} \neq 0\}|}{|\{ij \text{ pairs} : \beta^*_{ij} \neq 0\}|},$$

$$\text{specificity} = \frac{|\{ij \text{ pairs} : \hat{\beta}_{ij} = 0 \text{ and } \beta^*_{ij} = 0\}|}{|\{ij \text{ pairs} : \beta^*_{ij} = 0\}|},$$

$$\text{prediction error} = \|\boldsymbol{Y}_{\text{test}} - \boldsymbol{X}_{\text{test}}\hat{\boldsymbol{B}}\|_2^2,$$

where $|\cdot|$ is the number of elements in a set and $(\boldsymbol{Y}_{\text{test}}, \boldsymbol{X}_{\text{test}})$ is an independently generated testing set of 100 samples.

Figure 2 summarizes these quantities for simulation setups with "not all in and all out" for all the group structures in Figure 1 at $p = q = 200$, $n = 100$, and $\rho = 0.5$. The proposed method using *Lasso+X+XY* for the nonoverlapping group structures $\mathcal{G}_X$, $\mathcal{G}_{XY}$, and $\mathcal{G}_{XY} \cup \mathcal{G}_X$ as well as for the overlapping group structure are highlighted in black. The methods for the correctly specified group structures are highlighted in grey except in Figure 2c and d, where the implemented group structures are by themselves the correctly specified group structures. From Figure 2 we see that correctly incorporating group structure improves both variable selection and prediction, and

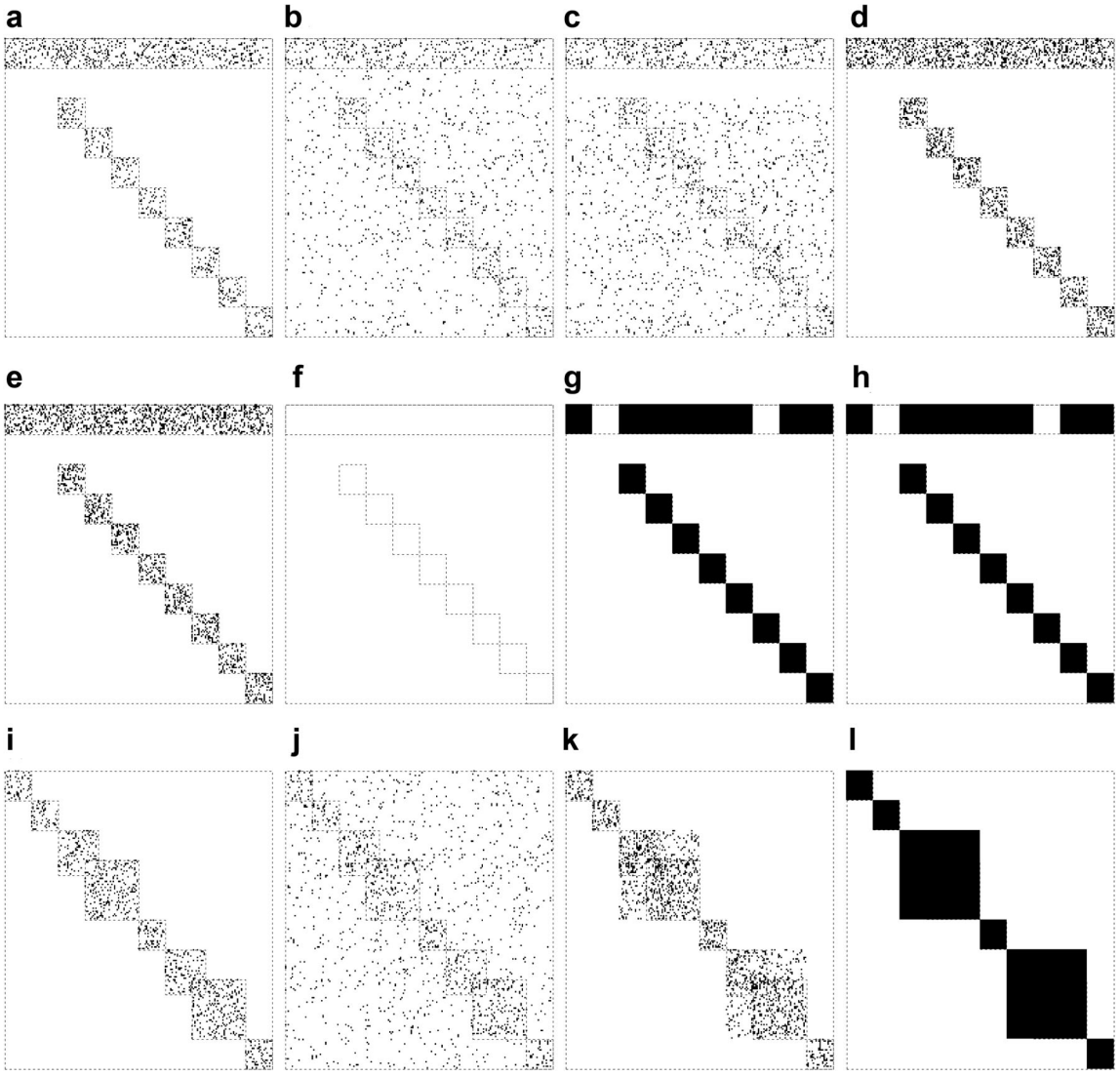

Figure 3. Heatmaps of coefficient matrices, selection effects. (a–h): "Not all in all out" $X{+}XY$ nonoverlapping group structure with $n = 100$, $p = 200$, $q = 200$, and $\rho = 0.5$. (a) $\boldsymbol{B}^*$; (b) $\hat{\boldsymbol{B}}_L$; (c) $\hat{\boldsymbol{B}}_{LX}$; (d) $\hat{\boldsymbol{B}}_{LXY}$; (e) $\hat{\boldsymbol{B}}_{LXXY}$; (f) $\hat{\boldsymbol{B}}_{GX}$; (g) $\hat{\boldsymbol{B}}_{GXY}$; (h) $\hat{\boldsymbol{B}}_{GXXY}$. (i–l): "Not all in all out" overlapping group structure with $n = 100$, $p = 200$, $q = 200$, and $\rho = 0.5$. (i) $\boldsymbol{B}^*$; (j) $\hat{\boldsymbol{B}}_L$; (k) $\hat{\boldsymbol{B}}_{SGL}$; (l) $\hat{\boldsymbol{B}}_G$.

our proposed method *Lasso+X+XY*, or the MSGLasso, performs at least the same as, if not better than, the methods for the correct group structures and yields the lowest prediction errors.

Figure 3 illustrates fitted results for a data set randomly chosen from one hundred replications, where $\boldsymbol{B}^*$ has a "not all in all out" either $\mathcal{G}_{XY} \cup \mathcal{G}_X$ or overlapping group structure with $p = 200$, $q = 200$, and $\rho = 0.5$. It clearly shows that the MSGLasso results for correctly specified group structure, both in Figure 3e and in Figure 3k, yield the most desirable estimates. Methods without lasso penalty yield too many false positives inside the important groups for the "not all in all out" case even when the groups are correctly specified, while methods with lasso penalty but incorrectly specified groups yield too many false positives outside the important groups.

5.2. *Yeast eQTL Data Analysis*

In this section, we demonstrate our method by analyzing a yeast eQTL data set generated by Brem and Kruglyak (2005), see also Yin and Li (2011), where gene expressions are grouped into, possibly overlapping, pathways and the genetic markers are grouped into genes.

The data set contains 6216 yeast genes assayed for 112 individual segregant. Genotypes of these 112 segregant at 2956 marker positions were also collected using GeneChip Yeast Genome S98 microarrays. The 6216 expressed genes are grouped by Kyoto Encyclopedia of Genes and Genomes pathways and the 2956 markers are grouped by genes, taking isoform genes as the same gene. To illustrate the method, in the reported analysis we only include genes from the following four pathways: the

Table 1
Comparison of prediction errors between different methods

Method	MSG lasso	M lasso	MG lasso	lasso
Prediction error	3094.5	3396.8	3557.4	3683.3

MSG lasso = multivariate sparse group lasso, M lasso = multivariate lasso, MG lasso = multivariate group lasso, lasso = univariate lassos.

mitogen-activated protein kinases (MAPK) pathway containing 54 genes, the *cell cycle* pathway containing 116 genes, the *cancer* pathway containing 20 genes and the *ribosome* pathway containing 137 genes. There are in total 315 distinct expressed genes in these pathways, with 5 genes overlapping between *MAPK* and *cell cycle*, 5 genes overlapping between *MAPK* and *cancer*, 3 genes overlapping between *cell cycle* and *cancer*, and 1 gene overlapping between *MAPK*, *cell cycle* and *cancer*. *Ribosome* does not contain overlapping genes with the other three pathways.

We follow a similar procedure of Yin and Li (2011) for prescreening genotype markers by performing univariate linear regressions across all the 315 gene expressions and 2956 markers, and include the 395 markers with p-value of 0.01 or smaller into the final analysis. These 395 markers are embedded in 45 distinct genes.

Since some marker within a gene is associated with some gene expression in a pathway does not necessarily imply the gene must be associated with all four pathways, we exclude the $\mathcal{G}_X$ group structure and only apply an overlapping $\mathcal{G}_{XY}$ group structure in the data analysis. We cross-validate the performance of the multivariate sparse group lasso, the multivariate lasso, the multivariate group lasso and the univariate lasso. In particular, we randomly divide the 112 samples into five approximately equal sized subsets, set one subset aside as the test set, and use the remaining four subsets as the training set. Then for each model, we run five-fold cross-validation on the training set to estimate the coefficient matrix, and use the estimated model to compute the prediction error on the test set. We repeat the above procedures until each of the five subsets has been used as the test set once. The overall cross-validated prediction errors, the sum of squares, are reported in Table 1. The univariate lasso is conducted by first selecting variables on the training set using 315 separate lasso regressions, each for a single gene expression variable, and then implementing multivariate linear regression on only the selected set of covariates to obtain $\hat{\boldsymbol{B}}$. Our proposed method has the best performance. The univariate lasso gives the highest prediction error, which is expected because the relations among responses are totaly overlooked, and this leads to high variability and over-fitting (Peng et al., 2010). The proposed method shows roughly a 10% decrease of the cross-validated prediction error over the multivariate lasso method, the second best approach among all four compared methods.

We then apply the multivariate sparse group lasso to the entire data set with 315 gene expressions and 395 markers. The final tuning parameters are $\lambda = 7 \times 10^{-2}$ and $\lambda_1 = 2 \times 10^{-4}$, determined by a fivefold cross-validation. We also investigate the selection stability following Meinshausen and Bühlmann (2010) by calculating the selection frequencies of the top selected associations using one hundred bootstrap datasets. The top associations in terms of size, with selection frequency no less than 95%, are given in Table 2. The p-values in the last column are obtained from marginal simple linear regressions. Overall there are 1422 nonzero elements in the estimated coefficient matrix, which gives an overall estimated sparsity of about 1%. There are 235 markers with nonzero coefficients related to genes in the *MAPK* pathway, 135 markers related to genes in the *cell cycle* pathway, 65 markers related to genes in the *cancer* pathway, and 65 markers related to genes in the *ribosome* pathway. Among those, 34 markers are related to genes in the overlap of *MAPK* and *cell cycle* pathways, 23 markers are related to genes in the overlap of *MAPK* and *cancer* pathways, and 5 markers are related to a gene in the overlap of *MAPK*, *cell cycle* and *cancer* pathways.

Table 3 lists the top pathway-gene groupwise associations in terms of the group L_2 norms with a 100% group-wise selection frequency. Out of 180 block groups, 89 groups contain nonzero coefficients. Several top selected genes have been reported in the literature. For example, one of the isoforms of *YCR* gene, *YCR073C/SSK22* is *MAPK* cascade involved in osmosensory signaling pathway. Gene groups *YJL* and *YGR* in the Scr homology 3 domains are interacting with gene *Pbs2* in one of the three kinase components in the *MAPK* pathway (Zarrinpar, Park, and Lim, 2003). The top association signals detected between the gene expressions in the joint of *MAPK*, *cell cycle* and *cansor* pathways and markers in *NHR* gene group also confirm the regulation effects of *NHR* genes on *cell cycle* pathway and other autophagy-related genes.

It is worth noting that none of the association p-values from marginal simple linear regressions between gene *YJL* and pathway *MAPK* survives the Bonferroni correction for multiple comparisons. For example, the 14th signal in Table 2 has a univariate marginal p-value of 0.044, therefore it is unlikely to be picked up by the pairwise analysis. However, the MSGLasso successfully selected this signal in an adjusted analysis with high individual and group selection frequencies, see Tables 2 and 3. This finding is supported by Zarrinpar et al. (2003). It demonstrates that besides the advantage of dimension reduction, the MSGLasso can also pick out important signals that would be missed by the pairwise method.

The stability selection results show that the first 40 selected top signals do not contain zero within their 2.5–97.5% bootstrap percentile band, and the bootstrap Q1–Q3 band of the top 100 selected signals do not contain zero, indicating that the top selected signals using proposed method have high selection frequencies from bootstrap samples.

6. Discussion

For a predetermined group structure, the MSGLasso effectively and efficiently selects the important groups and important individual signals within those groups. There is some interest in recent literature in learning the group structure and selecting the important variables simultaneously. For example, Yin and Li (2011) proposed a conditional Guassian graphical model to select nonzero entries in the precision matrix conditional on simultaneously selected predictors. It is of interest to select important predictors via the MSGlasso

Table 2
Top selected expression-marker associations

Index	$\hat{\beta}_{jk}$	Sel. freq.* (%)	Expr.** name	Expr. pathways	Marker Chr:BP***	Marker gene	p-value
1	−1.481	100	*YKL178C*	*MAPK*	3:201166	*YCR041W*	2.43e−51
2	1.465	100	*YFL026W*	*MAPK*	3:201166	*YCR041W*	2.81e−55
3	−1.264	100	*YPL187W*	*MAPK*	3:201166	*YCR041W*	7.10e−45
4	1.061	100	*YNL145W*	*MAPK*	3:201166	*YCR041W*	5.54e−39
5	−0.735	100	*YGL089C*	*MAPK*	3:201166	*YCR041W*	8.53e−20
6	0.650	100	*YFL026W*	*MAPK*	3:201167	*YCR041W*	2.81e−55
7	−0.649	100	*YKL178C*	*MAPK*	3:201167	*YCR041W*	2.43e−51
8	−0.554	98	*YPL187W*	*MAPK*	3:201167	*YCR041W*	7.10e−45
9	0.452	100	*YDR461W*	*MAPK*	3:201166	*YCR041W*	8.42e−14
10	−0.385	98	*YPL187W*	*MAPK*	3:177850	*gCR02*	1.65e−33
11	0.352	100	*YGR088W*	*MAPK*	15:170945	*gOL02*	1.52e−10
12	0.346	100	*YGR088W*	*MAPK*	15:174364	*gOL02*	1.51e−10
13	−0.318	97	*YKL178C*	*MAPK*	3:177850	*gCR02*	2.44e−37
14	0.257	98	*YGR088W*	*MAPK*	10:51003	*YJL204C*	0.044
15	−0.175	95	*YGL089C*	*MAPK*	2:681361	*YML056C*	0.66

*Sel. Freq. = Selection frequency. **Expr. = gene expression. ***Marker is denoted by its physical position in the format of "chromosome:basepair".

Table 3
Top selected pathway-gene associations (with 100% selection frequency)

Index	Pathway	Gene	$\|\hat{\boldsymbol{B}}_g\|_2$	Number of nonzero $\hat{\beta}_{jk}$ in group	Top expr.* in pathway	Top marker** in gene	Top $\hat{\beta}_{jk}$ in group
1	*MAPK*	*YCR*	3.06	23	*YKL178C*	3:201166	−1.481
2	*MAPK*	*gOL*	0.508	10	*YGR088W*	15:170945	0.352
3	*MAPK*	*gCR*	0.499	3	*YPL187W*	3:177850	−0.385
4	*MAPK*	*YJL*	0.424	23	*YGR088W*	10:51003	0.257
5	*MAPK*	*NHR*	0.420	49	*YCL027W*	8:111686	−0.184
6	*MAPK*	*NBR*	0.382	15	*YGL089C*	2:681361	0.207
7	*MAPK*	*YBR*	0.372	81	*YGR088W*	2:368060	0.165
8	*Ribosome*	*YER*	0.342	119	*YER102W*	5:350744	−0.063
9	*Cancer*	*YLR*	0.286	14	*YJR048W*	12:674651	0.164
10	*MAPK*	*YGR*	0.275	3	*YGL089C*	7:916471	−0.172
11	*MAPK*	*YPL*	0.274	18	*YGR088W*	12:428612	0.240
12	*MAPK*	*YLR*	0.252	62	*YCL027W*	12:957108	0.092
13	*MAPK*	*YER*	0.229	23	*YPL187W*	7:321714	0.135
14	*MAPK*	*YML*	0.214	23	*YGL098C*	13:164026	−0.175
15	*MAPK*	*YHL*	0.205	15	*YKL178C*	8:98513	−0.128
16	*MAPK*	*YNL*	0.183	23	*YGL089C*	14:418269	−0.083
17	*MAPK*	*YCL*	0.176	27	*YCL027W*	3:64311	0.140
18	*MAPK; Cell cycle*	*NHR*	0.175	44	*YJL157C*	8:111686	−0.061
19	*MAPK*	*gJL*	0.131	9	*YFL026W*	10:259991	0.098
20	*MAPK*	*YOL*	0.125	26	*YPL187W*	15:193911	0.084
21	*MAPK; Cell cycle; Cancer*	*NHR*	0.098	5	*YBL016W*	8:111686	−0.044
22	*Cell cycle*	*YCR*	0.067	5	*YLR288C*	3:201166	0.046
23	*Cell cycle*	*YCL*	0.063	16	*YDL003W*	3:64311	−0.035
24	*Cell cycle*	*YLR*	0.029	37	*YBR093C*	12:674651	0.012

*Expr. = gene expression. **Top marker in gene is denoted by its physical position in the format of "chromosome:basepair."

based on a data driven group structure, where the selection of group structure is a topic for future research.

The $L1/L2$ penalty in the MSGLasso ensures that the objective function is a convex function with respect to $\boldsymbol{B}$. The convexity is essential for the proposed mixed coordinate descent algorithm. Replacing the $L1$ penalty by the SCAD penalty (Fan and Li, 2001) would be of interest, but the respective optimization is non-convex, thus not guaranteed to converge to the global minimum. More research along this line is needed.

7. Supplementary Materials

Web Appendices for the proofs of theoretical results referenced in Sections 3 and 4, computing cost comparison and MSGLasso package referenced in Section 4, and additional numerical results are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors thank Dr. Hongzhe Li for providing the yeast eQTL data and helpful discussions. The research was supported in part by the National Institute of Health grant R01-AG036802 and the National Science Foundation grants DMS-1407142, DMS-1007590 and DMS-0748389.

References

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* **37**, 1705–1732.

Biswas, S. and Lin, S. (2012). Logistic Bayesian lasso for identifying association with rare haplotypes and application to age-related macular degeneration. *Biometrics* **68**, 587–597.

Brem, R. B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5700 gene expression traits in yeast. *Procedings of National Academy of Sciences* **102**, 1572–1577.

Bunea, F., She, Y., and Wegkamp, M. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *Annals of Statistics* **39**, 1282–1309.

Dudoit, S., Shaffer, J., and Boldrick, J. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**, 71–103.

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

Huang, J., Ma, S., Xie, H., and Zhang, C. (2009). A group bridge approach for variable selection. *Biometrika* **2**, 339–355.

Lounici, K., Pontil, M., Tsybakov, A. B., and van de Geer, S. (2011). Oracle inequalities and optimal inference under group sparsity. *Annals of Statistics* **39**, 2164–2204.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society, Series B* **72**, 417–473.

Obozinski, G., Wainwright, M., and Jordan, M. (2011). Support union recovery in high-dimensional multivariate regression. *Annals of Statistics* **39**, 1–47.

Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**, 30–50.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010). Newblock regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals of Applied Statistics* **4**, 53–77.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22**, 231–245.

Stein, J., Hua, X., Lee, S., Ho, A., Leow, A., Toga, A., Saykin, A., Shen, L., Foroud, T., Pankratz, N., Huentelman, M., Craig, D., Gerber, J., Allen, A., Corneveaux, J., Dechairo, B., Potkin, S., Weiner, M., Thompson, P., and Initiative, A. D. N. (2010). Voxelwise genome-wide association study (vgwas). *Neuroimage* **53**, 1160–1174.

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

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization: Theory and Applications* **109**, 275–294.

Wu, T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Annal of Applied Statistics* **2**, 224–244.

Yin, J. and Li, H. (2011). A sparse conditional Gaussian graphical model for analysis of genetical genomics data. *Annals of Applied Statistics* **4**, 2630–2650.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

Zamdborg, L. and Ma, P. (2009). Discovery of protein–DNA interactions by penalized multivariate regression. *Nucleic Acids Research* **37**, 5246–5254.

Zarrinpar, A., Park, S. H., and Lim, W. A. (2003). Optimization of specificity in a cellular protein interaction network by negative selection. *Nature* **426**, 676–680.

Zhang, S., Ching, W., Tsing, N., Leung, H., and Guo, D. (2010). A new multiple regression approach for the construction of genetic regulatory networks. *Artificial Intelligence in Medicine* **48**, 153–160.

Zhou, H., Sehl, M. E., Sinsheimer, J. S., and Lange, K. (2010). Association screening of common and rare genetic variants by penalized regression. *Nucleic Acids Research* **26**, 2375–2382.

Zhou, N. and Zhu, J. (2010). Group variable selection via a hierarchical lasso and its oracle property. *Statistics and Its Interface* **4**, 557–574.

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

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

Received August 2013. *Revised December* 2014.
Accepted January 2015.