



## Feature Selection by Canonical Correlation Search in High-Dimensional Multiresponse Models With Complex Group Structures

Shan Luo & Zehua Chen

To cite this article: Shan Luo & Zehua Chen (2020) Feature Selection by Canonical Correlation Search in High-Dimensional Multiresponse Models With Complex Group Structures, Journal of the American Statistical Association, 115:531, 1227-1235, DOI: [10.1080/01621459.2019.1609972](https://doi.org/10.1080/01621459.2019.1609972)

To link to this article: <https://doi.org/10.1080/01621459.2019.1609972>



View supplementary material [↗](#)



Published online: 11 Jun 2019.



Submit your article to this journal [↗](#)



Article views: 1000



View related articles [↗](#)



View Crossmark data [↗](#)



# Feature Selection by Canonical Correlation Search in High-Dimensional Multiresponse Models With Complex Group Structures

Shan Luo<sup>a</sup> and Zehua Chen<sup>b</sup>

<sup>a</sup>Department of Statistics, Shanghai Jiao Tong University, Shanghai, China; <sup>b</sup>Department of Statistics & Applied Probability, National University of Singapore, Singapore

## ABSTRACT

High-dimensional multiresponse models with complex group structures in both the response variables and the covariates arise from current researches in important fields such as genetics and medicine. However, no enough research has been done on such models. One of a few researches, if not the only one, is the article by Li, Nan, and Zhu where the sparse group Lasso approach is extended to such models. In this article, we propose a novel approach named the sequential canonical correlation search (SCCS) procedure. In the SCCS procedure, the nonzero group by group blocks of regression coefficients are searched stepwise using a canonical correlation measure. Each step of the procedure consists of a block selection and a sparsity identification. The model selection criterion, EBIC, is used as the stopping rule of the procedure. We establish the selection consistency of the SCCS procedure and conduct simulation studies for the comparison of existing methods. The SCCS procedure has two advantages over the sparse grouped Lasso method: (i) it is more accurate in the identification of nonzero coefficient blocks and their nonzero entries, and (ii) its implementation is not limited by the dimensionality of the models and requires much less computation. A real example in genetic studies is also considered. Supplementary materials for this article are available online.

## ARTICLE HISTORY

Received January 2018  
Revised September 2018

## KEYWORDS

Canonical correlation;  
Feature selection; Group  
structure; High-dimensional;  
Multiresponse model;  
Sequential procedure

## 1. Introduction

Complex group structures in variables appear in many important scientific fields, specifically, in genetic and medical studies. For example, in studies such as gene-gene association study (see Park and Hastie 2007; Zhang et al. 2010), protein-DNA association study (Zamdborg and Ma 2009), brain fMRI-DNA association study (Stein et al. 2010), both the responses and covariates have group structures. A particular example is provided by the study of Brem and Kruglyak (2005) for yeast eQTL mapping. The purpose of the study is to discover the genes which affect certain pathways. In the study, the expression levels of 6216 genes and the genotype of 2956 markers are obtained for 112 individuals. The expressed genes are grouped into pathways, and the markers are grouped by genes.

The group structures of the data carry important information. However, such information has been ignored in common practices where the group structures are not taken into account. It caught the attention of the researchers in recent years that by ignoring the group structures it might fail to detect even obvious important relationships. For example, in a study on Alzheimer's disease, Stein et al. (2010) conducted a voxelwise genome-wide association study in 740 elderly subjects where 173 have Alzheimer's disease, 361 have mild cognitive impairment, and 206 are healthy. Though it is well-known that genetics is the main risk factor of Alzheimer's disease, the study found no statistically significant markers by using a variant-


by-variant approach which ignores structures such as brain functional regions and the gene memberships of the SNPs. It is also reported in other studies, for example, Biswas and Lin (2012) and Zhou et al. (2010), that the variant-by-variant approach suffers low power in detecting influential variants. In these studies, grouping genes into pathways or SNPs into genes are suggested as a remedy for the low power.

The group structures in either the responses or the covariates reflect the correlations among the variables within the groups. A statistic ignoring those correlation information is not sufficient. From the principle of sufficiency, by taking into account the group structures, the efficiency for detecting relevant features can be enhanced. However, statistical methods taking into account the group structures in both the responses and covariates are still scarce. In this article, we develop an approach for feature selection in high-dimensional multivariate models with complex group structures, which provides a timely method for the problems mentioned above.

There are abundant methods available for feature selection in high-dimensional models without group structures. For univariate response models, the available methods can be roughly classified into two categories. One category consists of methods using regularized regression approach with various penalty functions. It includes LASSO (Tibshirani 1996), adaptive LASSO (Zou 2006), SCAD (Fan and Li 2001), MCP (Zhang 2010), and so on. The other category consists of sequential approaches such as least angle regression (LAR)

**CONTACT** Zehua Chen  [stachenz@nus.edu.sg](mailto:stachenz@nus.edu.sg)  Department of Statistics & Applied Probability, National University of Singapore, 6 Science Drive 2, Singapore 117546.

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/r/JASA](http://www.tandfonline.com/r/JASA).

 Supplementary materials for this article are available online. Please go to [www.tandfonline.com/r/JASA](http://www.tandfonline.com/r/JASA).

© 2019 American Statistical Association

(Efron et al. 2004),  $L_2$ -boosting (Buehlmann 2006), forward stepwise regression (FSR) (see Wang 2009; Ing and Lai 2011), orthogonal matching pursuit (OMP) (Cai and Wang 2011), sequential Lasso (Luo and Chen 2014), etc.

The regularized regression approach has been extended to the case of multiresponse models without group structures. The regularized regression approach for multiresponse models aims to minimize  $\text{Tr}[(Y - XB)^\top(Y - XB)]$  by imposing some constraints on  $B$  such as  $C(B) \leq c$  for some constant  $c$ ,  $C(B)$  being a function of  $B$ , where  $Y$  is the matrix of the observations on the response vector and  $X$  is the matrix of observed covariates. Various forms of  $C(B)$  have been considered by a number of authors, for example,  $C(B) = \sum_{j=1}^{\min(p,q)} \sigma_j(B)$  where  $\sigma_j(B)$  is the  $j$ th singular value of  $B$  (Yuan et al. 2007),  $C(B) = \text{RANK}(B)$  (Chen and Huang 2012),  $C(B) = \sum_{j=1}^p \|\beta_j\|_\infty$ , where  $\beta_j$  is the  $j$ th row of  $B$  (Turlach, Venables, and Wright 2005),  $C(B) = \sum_{j=1}^p \|\beta_j\|_2$  (Obozinski, Wainwright, and Jordan 2011),  $C(B) = \lambda \sum_{j=1}^p \|\beta_j\|_1 + (1 - \lambda) \sum_{j=1}^p \|\beta_j\|_2$  where  $0 < \lambda < 1$  (Peng et al. 2010), etc.

Univariate response models with group structures in the covariates have been mainly dealt with by the so-called group Lasso approach. The group LASSO approach was first proposed in Yuan and Lin (2006) which impose a  $L_2$  penalty on the coefficient vectors of groups. Group LASSO using  $L_r$  ( $0 < r < 1$ ) bridge penalties are considered in Huang et al. (2009) and Zhou and Zhu (2010). A sparse-group LASSO method, which further identifies important variables in each group, was proposed in Simon et al. (2013). The group Lasso approach was applied to tackle high-dimensional multivariate response models with group structures in both the responses and the covariates by Li, Nan, and Zhu (2015), which is, to our knowledge, the only work dealing with high-dimensional models with complex group structures.

In this article, we develop a sequential method for feature selection in multivariate response models with complex group structures based on the principle of correlation. The principle of correlation underlies all feature selection approaches, especially, the sequential approaches for univariate response models, which will be elaborated in the next section. The method which we develop is called the sequential canonical correlation search, SCCS for short. The SCCS selects covariate groups for the response groups sequentially. At each step, a covariate group relative to a response group is first selected according to its correlation with the current residuals measured by a quantity based on canonical correlation coefficients, then its nonzero coefficients are selected by EBIC (Chen and Chen 2008). The EBIC also serves as the stopping rule for the procedure. The rationale and detail of SCCS are provided in Section 2. The SCCS is selection consistent, which we establish in Section 3.

The SCCS is, we believe, the first sequential procedure for feature selection in high-dimensional models having complex group structures. There are recently a few feature screening methods using a different correlation measure called distance correlation (DC) for data with group structures in both the responses and features (see Li, Zhong, and Zhu 2012; Kong et al. 2017). However, those methods are not for feature selection. Though feature screening and feature selection are closely related, they are different to a large extent. In principle, DC

can replace the canonical correlation measure in SCCS. But the canonical correlation measure has certain advantages over DC, which will be elaborated in the next section. The SCCS not only has an edge over the group Lasso approach of Li, Nan, and Zhu (2015) but also has a great advantage in computation, as our simulation studies show. The SCCS is not restricted by the dimension of the data while the group Lasso approach might not be implementable when the number of covariate groups is too large. This is of more importance in practical high-dimensional data analysis.

The remainder of the article is arranged as follows. In Section 2, we discuss the principle of correlation and describe SCCS in detail. In Section 3, we deal with the theoretical properties of SCCS. In Section 4, we report our simulation studies for the comparison of SCCS with the group Lasso approach. In Section 5, we make the comparison based on a real dataset. The technical details are given in a supplementary document.

## 2. The Principle of Correlation and the SCCS Procedure

The SCCS method is developed based on the principle of correlation. This principle in fact underlies all feature selection procedures. However, it has never been explicitly spelled out. In this section, we first elucidate this principle since it explains the rationale of the SCCS method.

Let us begin by inspecting the mechanism of the sequential methods mentioned in the last section. Those methods are common in the following nature: at each step, they start with a current estimated mean response  $\hat{\mu}_A$ , and select the next variable such that it has the largest correlation with the current residual  $\tilde{y} = y - \hat{\mu}_A$ , then update the current estimate with the newly selected variable and proceed to the next step. The differences among those methods are their ways of updating  $\hat{\mu}_A$  and the correlation measures used. Without loss of essence, suppose there is only one variable which attains the largest correlation at each step. Let  $x_A$  be the variable which attains the largest correlation.  $L_2$ -boosting updates the current estimate by  $\hat{\mu}_A + \hat{\beta}_A x_A$ , where  $\hat{\beta}_A = (x_A^\top x_A)^{-1} x_A^\top \tilde{y}$ . LAR updates the current estimate by  $\hat{\mu}_A + \hat{\gamma} x_A$ , where  $\hat{\gamma}$  is a certain shrunk estimate of the coefficient of  $x_A$ . FSR and OMP update the current estimate by an ordinary least squares fit of the response to all selected variables.  $L_2$ -boosting, LAR, and OMP use Pearson's correlation coefficient as the correlation measure, but FSR uses a partial Pearson's correlation coefficient which can be considered as a weighted Pearson's correlation coefficient between the residual and the selected variable. It is clear that correlation is the intrinsic mechanism for feature selection in these sequential methods.

The same mechanism also governs the regularized regression approaches. If the columns of the design matrix of the regression model are orthogonal, the variables selected by Lasso are those whose correlations with the response exceed a threshold determined by the penalty parameter. The general case can be illustrated by sequential Lasso of Luo and Chen (2014). The idea of sequential Lasso is to select variables sequentially by minimizing partially penalized sum of squares. In the partially penalized sum of squares, the variables already selected are not penalized and the penalty parameter is tuned to be the largest but still

allows new variables to be selected. It is shown in Proposition 3 of Luo and Chen (2014) that the necessary condition for a variable to be selected at a step is that its correlation with the least squares residual of the previous model must be the highest among all unselected variables.

We now focus on sequential approaches. Denote by  $Y$  and  $X_j, j = 1, \dots, p$ , respectively, the response and feature objects (an object is either a scalar, a vector or a matrix). Let  $\hat{\mu}_Y$  denote an estimate of  $EY$  and  $\tilde{Y}$  the residual of  $Y$  after taking into account selected feature objects. Let  $r(\tilde{Y}, X_j)$  be a correlation measure and  $\kappa(\tilde{Y}, X_j)$  be a stopping rule. In the following, we give a general procedure dubbed correlation search procedure which provides a common roof over all existing sequential methods and opens the door for developing new methods.

*Correlation search procedure:*

- Set  $\hat{\mu}_Y = 0$ .
- Repeat
  - Compute  $r(\tilde{Y}, X_j)$  for all  $j$  and identify the  $\hat{j}$  such that  $r(\tilde{Y}, X_{\hat{j}})$  is the largest.
  - If  $X_{\hat{j}}$  passes the examination of  $\kappa(\tilde{Y}, X_{\hat{j}})$ , update  $\hat{\mu}_Y$  and  $\tilde{Y}$  by  $X_{\hat{j}}$  and continue, otherwise, stop.

For a particular problem, a choice of the correlation measure and the method for updating the estimate  $\hat{\mu}_Y$  must be made. In our opinion, the choice should be guided by the essence of feature selection. The essence of feature selection is to find features which contribute to the unexplained variation of the response. We confine ourselves to linear models. The unexplained variation is reflected in the residual of the response after fully taking into account the contribution of the selected features. A feature which makes a contribution to the unexplained variation is linearly correlated with the residual. This gives rise to what we called the principle of correlation: *Features which make contribution to the unexplained variation of the response are those which have linear correlations with the residual. Features having the highest correlation must be selected first.* According to the principle of correlation, for linear models, the estimate  $\hat{\mu}_Y$  should be updated by ordinary least squares fit since it fully accounts for the variation attributable to the selected variables, and the correlation measure should best reflect the contribution of a feature to the variation of the residual. Among the sequential approaches, only OMP (sequential Lasso is equivalent to OMP when there is only one feature which attains the largest correlation at each step) meets these requirements.  $L_2$ -boosting and LAR, similar to regularized regression approaches, have shrunk estimates of the feature coefficients and hence do not fully account for the variation attributable to the selected variables. The partial correlation in FSR inflates the correlation of a variable with the residual if the variable is correlated with already selected variables. The inflation is proportional to how strong the correlation between the variable and those already selected is. This can be seen from the fact that minimizing the residual sum of squares by adding a new variable  $x_j$  is equivalent to maximizing  $[\tilde{y}^\top x_j]^2 / x_j^\top [I - H(X^*)] x_j$  where  $X^*$  denotes the design matrix consisting of the already selected variables. Method with shrunk estimates such as  $L_2$ -boosting and LAR

might have certain merits for prediction but not necessarily for the identification of relevant features. Methods with least-square estimates such as OMP are greedy but the greedy nature is not bad for the identification of relevant features (see, e.g., Tropp 2004). For further discussions on the merits and dis-merits of these approaches, the reader is referred to Section 6 of Luo and Chen (2014).

In the following, we describe the SCCS procedure in detail. Let  $\mathcal{Y}$  be a set of  $q$  response variables. Let  $\mathcal{Y}_j, j = 1, \dots, J$ , denote the response groups. Note that  $\mathcal{Y} = \bigcup_{j=1}^J \mathcal{Y}_j$ , but the  $\mathcal{Y}_j$ 's are not necessarily disjoint. Let  $\mathcal{X}$  be a set of  $p$  covariates. Let  $\mathcal{X}_k, k = 1, \dots, K$ , denote the covariate groups. Similarly,  $\mathcal{X} = \bigcup_{k=1}^K \mathcal{X}_k$  and the  $\mathcal{X}_k$ 's are not necessarily disjoint. By an abuse of notation, we also let  $\mathcal{Y}_j$  and  $\mathcal{X}_k$  denote the vectors consisting of the elements in themselves, the meaning of which can be identified from the context. Suppose that we have a sample of size  $n$ . Let  $Y$  be the  $n \times q$  matrix of observations on the response variables  $\mathcal{Y}$ , and  $X$  be the  $n \times p$  matrix of observations on the covariates  $\mathcal{X}$ . Suppose that all the columns of  $Y$  and columns of  $X$  are standardized such that their means are zero and squared norms are  $n$ . By the group structures,  $Y$  and  $X$  are partitioned, respectively, as  $Y = (Y_1, \dots, Y_J)$  and  $X = (X_1, \dots, X_K)$ , where, for  $j = 1, \dots, J$ ,  $Y_j$  is the matrix consisting of the columns of  $Y$  corresponding to the components in  $\mathcal{Y}_j$  and, for  $k = 1, \dots, K$ ,  $X_k$  is the matrix consisting of the columns of  $X$  corresponding to the components of  $\mathcal{X}_k$ . We consider the case that  $q$  and  $p$  are larger than  $n$  but assume that the number of columns for each  $Y_j$  and  $X_k$  is smaller than  $n$ . We deal with the following model

$$\begin{pmatrix} Y_1 & \cdots & Y_J \end{pmatrix} = \begin{pmatrix} X_1 & \cdots & X_K \end{pmatrix} \begin{pmatrix} B_{11} & \cdots & B_{1J} \\ \cdots & \cdots & \cdots \\ B_{K1} & \cdots & B_{KJ} \end{pmatrix} + \begin{pmatrix} \mathcal{E}_1 & \cdots & \mathcal{E}_J \end{pmatrix}, \quad (1)$$

where  $\mathcal{E}_j$ 's are random error matrices independent of  $X_k$ 's. We also assume the sparsity of the model, that is, the number of nonzero coefficient blocks  $B_{kj}$  is small. The goal is to select relevant  $X$ -groups for each  $Y$ -group, or, equivalently, to identify nonzero  $B_{kj}$ 's.

The SCCS is a particular version of the general correlation search procedure. In SCCS, we use the sum of the canonical correlation coefficients as the correlation measure between two random vectors. Canonical correlation coefficients are extensions of Pearson's correlation coefficient. They measure essentially linear associations between the two random vectors. In fact, the sum of the canonical correlation coefficients is the squared Frobenius norm of the covariance matrix of the two random vectors standardized, see the remark on condition B1 in Section 3. The DC provides another correlation measure between two random vectors. The DC can measure any kind of correlation and might have an advantage in feature screening for identifying features which have nonlinear correlation with the response. But this advantage cannot be realized in model based feature selection since DC neither can tell whether the correlation is linear or nonlinear nor can tell the nature of the nonlinearity when the correlation is nonlinear. It is hard to determine a particular nonlinear function form of a variable to be included in the model. On the other hand, since the canonical correlation measure only measures linear correlation, it is more



suitable for linear model based feature selection. It is shown that the DC is an increasing function of the absolute Pearson's correlation coefficient  $\rho$  for two normal random variables (see Li, Zhong, and Zhu 2012). The gradient of DC is less than 1 for small  $\rho$  (roughly  $\rho < 0.5$ ) and bigger than 1 for large  $\rho$  (roughly  $\rho > 0.5$ ), which implies that DC has less information than Pearson's correlation on the linear relationship for small  $\rho$ . In practical problems, any single feature hardly has a large correlation (say, bigger than 0.5) with the unexplained variation of the response. This suggests that DC is less efficient than canonical correlation in linear model based feature selection from a theoretical viewpoint.

Let  $\Sigma_k$ ,  $\Omega_j$ , and  $\Xi_{kj}$  denote, respectively, the variance matrices of  $\mathcal{X}_k$ ,  $\tilde{\mathcal{Y}}_j$  and the covariance matrix between  $\mathcal{X}_k$  and  $\tilde{\mathcal{Y}}_j$ . Here  $\tilde{\mathcal{Y}}_j$  denotes generically the residual of  $\mathcal{Y}_j$  after taking into account the effects of certain covariates. For the ease of notation, denote  $\Xi_{kj}^\top$  by  $\Xi_{jk}$ . The canonical correlation matrix between  $\mathcal{X}_k$  and  $\tilde{\mathcal{Y}}_j$  is defined as

$$C_{kj} = \Sigma_k^{-1} \Xi_{kj} \Omega_j^{-1} \Xi_{jk}.$$

The canonical correlation coefficients defined through the maximum of squared Pearson's correlation between linear combinations of  $\mathcal{X}_k$  and  $\tilde{\mathcal{Y}}_j$  are indeed the eigenvalues of the canonical correlation matrix. The canonical correlation matrix can be estimated by

$$\hat{C}_{kj} = (X_k^\top X_k)^{-1} X_k^\top \tilde{Y}_j (\tilde{Y}_j^\top \tilde{Y}_j)^{-1} \tilde{Y}_j^\top X_k,$$

if the columns of  $X_k$  and  $\tilde{Y}_j$  are centralized. The correlation measure in SCCS is taken as  $r(X_k, \tilde{Y}_j) = \text{tr} \hat{C}_{kj}$ , that is, the sum of the estimated canonical correlation coefficients.

In SCCS, we use EBIC (Chen and Chen 2008) as the stopping rule. The EBIC has a particular form in the context of the multiresponse model with group structures as follows. Let  $r_{kjl}$  denote the number of nonzero entries in  $B_{kjl}$ . Let  $\zeta$  be a model of (1) having  $m$  nonzero coefficient blocks  $\{B_{kjl}, l = 1, \dots, m\}$ . Then

$$\begin{aligned} \text{EBIC}(\zeta) = & n \sum_{j=1}^J \ln \frac{1}{n} \|Y_j - X \hat{B}^{(j)}(\zeta)\|_F^2 + \left( \sum_{l=1}^m r_{kjl} \right) \ln n \\ & + 2\gamma \left[ \ln \binom{KJ}{m} + \sum_{l=1}^m \ln \binom{|B_{kjl}|}{r_{kjl}} \right], \end{aligned}$$

where  $|B_{kjl}|$  denotes the number of entries of  $B_{kjl}$ ,  $\hat{B}^{(j)}(\zeta)$  is the least-square estimates of the coefficient matrix corresponding to  $Y_j$  and  $\|\cdot\|_F$  is the Frobenius norm, and  $\gamma = 1 - \frac{\ln n}{2 \ln p}$ .

Let  $\zeta^*$  and  $\tilde{Y}_j$  denote generically the current model and the residual of  $Y_j$  at each step, respectively. Since a model can be specified by its nonzero coefficients, we can consider  $\zeta^*$  as the set of the nonzero coefficients in the current model. The residuals are obtained from the least-square estimates of the model. At each step of SCCS, when a coefficient block is selected by the canonical correlation measure, its nonzero entries are selected by a forward procedure with EBIC as the stopping rule. The details of SCCS is given in the following algorithm.

### SCCS algorithm

**Initial step:** Set  $\zeta^* = \emptyset$ .

**General step:**

- *Selection of coefficient block:* Compute the residuals  $\tilde{Y}_k$  from the current model  $\zeta^*$ . For all pairs  $(k, j)$  which have not been selected yet, compute  $r(X_k, \tilde{Y}_j)$  and identify  $k^*$  and  $j^*$  such that  $r(X_{k^*}, \tilde{Y}_{j^*})$  is the maximum among those pairs. Select the block  $B_{k^*j^*}$ .
- *Selection of nonzero entries of the coefficient block:* For the block  $B_{k^*j^*}$ , select its nonzero entries one at a time in the following forward procedure. Initially, let  $\mathcal{B} = \{\beta_{lm}\}$  be the set of all entries of  $B_{k^*j^*}$ . Then repeat the following.
  - Let  $\zeta^* \cup \beta_{lm}$  denote the model obtained by augmenting  $\zeta^*$  with  $\beta_{lm}$ . Compute  $\text{EBIC}(\zeta^* \cup \beta_{lm})$  for all  $\beta_{lm} \in \mathcal{B}$ .
  - If  $\text{EBIC}(\zeta^* \cup \beta_{\tilde{l}m})$  is the smallest and  $\text{EBIC}(\zeta^* \cup \beta_{\tilde{l}m}) < \text{EBIC}(\zeta^*)$ , update  $\zeta^*$  as  $\zeta^* \cup \beta_{\tilde{l}m}$ , and update  $\mathcal{B}$  as  $\mathcal{B} \setminus \beta_{\tilde{l}m}$ .
  - If  $\text{EBIC}(\zeta^* \cup \beta_{\tilde{l}m}) \geq \text{EBIC}(\zeta^*)$  or  $\mathcal{B} \setminus \beta_{\tilde{l}m}$  is empty, stop the repetition.
- If none of the entries of  $B_{k^*j^*}$  is selected, stop the sequential procedure.

Like OMP in the univariate response situation, at each step of the SCCS procedure, the variation of the response attributable to already selected features is fully taken into account in the current residual, the next feature to be selected is considered purely based on its correlation with the current residual. Therefore, the merits of OMP compared with FSR,  $L_2$ -boosting and LAR carry over to the SCCS procedure.

### 3. Theoretical Properties of the SCCS Procedure

We investigate the theoretical properties of the SCCS procedure in this section. To facilitate our discussion, we introduce more notations. In a tentative model of the SCCS procedure, the sets of covariates associated with the response variables are in general different. Let  $s_i$  be the index set of the covariates associated with  $y_i$  in the given tentative model. The tentative model is represented by  $\zeta = \{s_1, s_2, \dots, s_q\}$ . For simplicity, if a covariate, say  $x_l$ , belongs to  $\mathcal{X}_k$ , we write  $l \in \mathcal{X}_k$  by an abuse of notation. Similarly for the indices of the response variables. For any index set  $s$  of the covariates, denote by  $X(s)$  the subvector consisting of the components of  $X$  with indices in  $s$ . The variance matrix of  $X(s)$  is denoted by  $\Sigma_{ss}$ . Denote by  $X(s)$  the matrix consisting of the columns of  $X$  with indices in  $s$ . Let  $H(s) = X(s)[X^\top(s)X(s)]^{-1}X^\top(s)$  and  $R(s) = I - H(s)$ .

Under model  $\zeta$ , the residual of the  $i$ th response component  $Y_i$  is defined as  $\tilde{Y}_i = Y_i - \alpha_i - \eta_i^\top X(s_i)$  where  $\alpha_i$  and  $\eta_i$  minimize  $E[Y_i - \alpha_i - \eta_i^\top X(s_i)]^2$ , that is, the residual is the difference between  $Y_i$  and its best linear predictor in terms of  $X(s_i)$ . It turns out that  $\eta_i = \Sigma_{\Pi}^{-1} \xi_{li}$  where  $\Sigma_{\Pi}$  is the variance matrix of  $X(s_i)$  and  $\xi_{li}$  is the covariance vector between  $X(s_i)$  and  $Y_i$ .

Let  $\Omega_j(\zeta)$  be the variance matrix of the residual of  $\mathcal{Y}_j$  under model  $\zeta$ . The  $(i, l)$ th entry of  $\Omega_j(\zeta)$  is given by

$$\tilde{\omega}_{il} = \text{cov}(Y_i - \alpha_i - \eta_i^\top X(s_i), Y_l - \alpha_l - \eta_l^\top X(s_l)), i, l \in \mathcal{Y}_j.$$

Let  $\Xi_{kj}(\zeta)$  be the covariance matrix between  $\mathcal{X}_k$  and the residual of  $\mathcal{Y}_j$  under model  $\zeta$ . Its  $(i, l)$ th entry is given by

$$\xi_{il} = \text{cov}(X_i, Y_l - \alpha_l - \eta_l^\top X(s_l)), i \in \mathcal{X}_k, l \in \mathcal{Y}_j.$$

Further notations are given below.

$m_j$ —number of variables in  $\mathcal{Y}_j$ .  $n_k$ —number of variables in  $\mathcal{X}_k$ .

$\Psi_0 = \{(k, j) : B_{kj} \neq 0\}$ —index set of nonzero coefficient blocks.

$\beta_l$ —the  $l$ th column of  $\mathbf{B} = (B_{kj})$ .

$s_{0l} = \{i : \beta_{il} \neq 0\}$ ;  $s_l^- = s_l^c \cap s_{0l}$ .

$p_0 = \max_{1 \leq l \leq q} \#s_{0l}$ .

$C_0 = \{l : \beta_l \neq 0\}$ —index set of nonzero columns of  $\mathbf{B}$ .

$\mu_l = X\beta_l$ —the mean vector of  $y_l$ , the  $l$ th column of  $Y$ .

$\mathcal{S}_\zeta = \{\zeta : \zeta = \{s_1, s_2, \dots, s_q\}, s_l \subsetneq s_{0l}, l = 1, \dots, q\}$ .

The canonical correlation matrix between  $\mathcal{X}_k$  and the current residual  $\tilde{\mathcal{Y}}_j$  under model  $\zeta$  is given by

$$C_{kj}(\zeta) = \Sigma_k^{-1} \Xi_{kj}(\zeta) \Omega_j^{-1}(\zeta) \Xi_{jk}(\zeta).$$

It is estimated by

$$\hat{C}_{kj}(\zeta) = \hat{\Sigma}_k^{-1} \hat{\Xi}_{kj}(\zeta) \hat{\Omega}_j^{-1}(\zeta) \hat{\Xi}_{jk}(\zeta),$$

where  $\hat{\Sigma}_k = \frac{1}{n} X_k^\top X_k$ , the entries of  $\hat{\Omega}_j(\zeta)$  and  $\hat{\Xi}_{kj}(\zeta)$  are, respectively, given by

$$\hat{\omega}_{il} = \frac{1}{n} \mathbf{y}_i^\top R(s_l) R(s_l) \mathbf{y}_l, i, l \in \mathcal{Y}_j,$$

$$\hat{\xi}_{il} = \frac{1}{n} \mathbf{x}_i^\top R(s_l) \mathbf{y}_l, i \in \mathcal{X}_k, l \in \mathcal{Y}_j.$$

First, we deal with the properties of the correlation measure  $\text{tr}(\hat{C}_{kj}(\zeta))$ . We have the following result.

**Lemma 1.** Assume the following conditions:

- A1.  $m_j$  and  $n_k$  are bounded.  $p_0 = O(n^{1/6})$ ,  $2 \ln(pq) < n^{1/3-\delta}$  for a small positive  $\delta$ .
- A2. The eigenvalues of  $\Sigma_k, \Sigma_{s_{0l} s_{0l}}, \Omega_j(\zeta)$  are bounded from below and above for all  $k, l, j$ , and  $\zeta \in \mathcal{S}_\zeta$ ;
- A3.  $\max_{i,j} \{\sigma(X_i X_j), \sigma(Y_i X_j), \sigma(Y_i Y_j)\} \leq C$  where  $C$  is a generic constant and  $\sigma(\cdot)$  denotes the standard deviation of its argument.
- A4. For  $t$  in a neighborhood of 0,  $\max_{i,j} \{E \exp\{t(X_i X_j - EX_i X_j)\}, E \exp\{t(Y_i Y_j - EY_i Y_j)\}, E \exp\{t(Y_i X_j - EY_i X_j)\}\} \leq C$ .

Suppose that  $(k^*, j^*)$  is the pair such that  $\text{tr}(C_{k^* j^*}(\zeta)) = \max_{k=1, \dots, K; j=1, \dots, J} \{\text{tr}(C_{kj}(\zeta))\}$ . Then, as  $n \rightarrow \infty$ , uniformly for  $\zeta$ , we have

$$P \left( \text{tr}(\hat{C}_{k^* j^*}(\zeta)) = \max_{k=1, \dots, K; j=1, \dots, J} \{\text{tr}(\hat{C}_{kj}(\zeta))\} \right) \rightarrow 1.$$

In what follows, we provide our main theoretical result, the selection consistency of the SCCS procedure. The conditions needed for the selection consistency are given as follows.

- B1. For all  $\zeta = (s_1, \dots, s_q) \in \mathcal{S}_\zeta$ ,

$$\max_{(k,j) \in \Psi_0^c} |\text{tr} C_{kj}(\zeta)| < \max_{(k,j) \in \Psi_0} |\text{tr} C_{kj}(\zeta)|.$$

- B2. For  $l \in C_0$  and any  $s_l \subset s_{0l}$ , for  $j$  satisfying  $s_l^- \cap \mathcal{X}_j \neq \emptyset$ ,

$$\max_{i \in s_{0l}^c \cap \mathcal{X}_j} n \ln \left( \frac{\|R(s_l) \mu_l\|_2^2}{\|R(s_l \cup i) \mu_l\|_2^2} \right) / \ln n \rightarrow 0.$$

- B3. As  $n \rightarrow +\infty$ ,

$$\sqrt{n} \min_{(j,k) \in \Psi_0} \min_{u \in \mathcal{Y}_k, v \in s_{0u} \cap \mathcal{X}_j} |\beta_{vu}| / \sqrt{p_0 \ln p} \rightarrow +\infty.$$

We make some remarks on these conditions in the following. First, we can express  $\text{tr} C_{kj}(\zeta)$  in a different form as follows.

$$\begin{aligned} \text{tr} C_{kj}(\zeta) &= \text{tr} \Sigma_k^{-1} \Xi_{kj}(\zeta) \Omega_j^{-1}(\zeta) \Xi_{jk}(\zeta) \\ &= \text{tr} \Sigma_k^{-1/2} \Xi_{kj}(\zeta) \Omega_j^{-1/2}(\zeta) \Omega_j^{-1/2}(\zeta) \Xi_{jk}(\zeta) \Sigma_k^{-1/2} \\ &= \|\Omega_j^{-1/2}(\zeta) \Xi_{jk}(\zeta) \Sigma_k^{-1/2}\|_F^2 \\ &= \|\text{cov}(\Omega_j^{-1/2}(\zeta) \tilde{\mathcal{Y}}_j, \Sigma_k^{-1/2} \mathcal{X}_k)\|_F^2 \\ &= \|\text{cov}(\tilde{\mathcal{Y}}_{j0}, \mathcal{X}_{k0})\|_F^2, \end{aligned}$$

where  $\tilde{\mathcal{Y}}_{j0}$  and  $\mathcal{X}_{k0}$  are the standardized  $\tilde{\mathcal{Y}}_j$  and  $\mathcal{X}_k$ , respectively, and  $\|\cdot\|_F$  is the Frobenius norm. Condition B1 essentially requires that a covariate group which has at least one variable associated with at least one response component in  $\mathcal{Y}_j$  is more correlated with  $\tilde{\mathcal{Y}}_j$  than a covariate group which is not associated with  $\mathcal{Y}_j$  at all. This explains the rationality of condition B1.

Second, considering the projection of  $\mu_l$  onto the space spanned by the columns of  $X(s_l)$  and that by the columns of  $X(s_l \cup i)$ , we have

$$\|R(s_l) \mu_l\|_2^2 - \|R(s_l \cup i) \mu_l\|_2^2 = [\mathbf{x}_i^\top R(s_l) \mu_l]^2 / \mathbf{x}_i^\top R(s_l) \mathbf{x}_i. \quad (2)$$

The right-hand side of (2) measures the variation of  $\mu_l$  which is attributable to  $\mathbf{x}_i$  in the amount explainable by  $X(s_l \cup i)$ . Condition B2 can be rewritten as

$$\max_{i \in s_{0l}^c \cap \mathcal{X}_j} \ln \left( 1 + \frac{[\mathbf{x}_i^\top R(s_l) \mu_l]^2 / \mathbf{x}_i^\top R(s_l) \mathbf{x}_i}{\|R(s_l \cup i) \mu_l\|_2^2} \right) = o \left( \frac{\ln n}{n} \right).$$

B2 states that the relative portion of the variation which can be explained by an additional irrelevant feature converges to zero at a certain rate.

Lastly, condition B3 allows the magnitude of the nonzero entries of  $\mathbf{B}$  to tail off at a certain rate, which is a common condition in high-dimensional feature selection.

**Theorem 1.** Assume conditions A1–A4 and B1–B3. Then the SCCS procedure is selection consistent. That is, denoting by  $s_l^*$  the index set for the nonzero regression coefficients of the response variable  $y_l$  selected by the SCCS procedure, we have, as  $n \rightarrow \infty$ ,

- (i) for each  $l \in C_0$ ,

$$P(s_l^* = s_{0l}) \rightarrow 1.$$

- (ii) for each  $l \in C_0^c$ ,

$$P(s_l^* = \emptyset) \rightarrow 1.$$

The theoretical results above allow that the total number of features  $p$ , the total number of response variables  $q$  and the maximum number of the relevant features for each response variable  $p_0$  diverge to infinity as the sample size  $n$  goes to infinity. Explicitly, condition A1 specifies the rates they can diverge to infinity. Under condition A1, B3 is satisfied if the magnitude of the minimum nonzero coefficients is larger than  $Cn^{-1/5}$  for some constant  $C$ .

The proofs of Lemma 1 and Theorem 1 are given in the supplementary document.

## 4. Simulation Study

We compare the performance of the SCCS procedure with that of the multivariate sparse group lasso (MSGSL) proposed in Li, Nan, and Zhu (2015) by a simulation study in this section. We first describe our simulation settings. We consider three nonzero patterns of the blocks of the coefficient matrix  $\mathbf{B}$ : (i) a diagonal blocks pattern, (ii) a diagonal overlap blocks pattern, and (iii) a random blocks pattern. The first two patterns are adapted from Li, Nan, and Zhu (2015) which are indeed group structures (b) and (d) therein. The simulation settings corresponding to these patterns are given in the following.

### 4.1. Diagonal Blocks Pattern

The following two settings of  $(n, q, p)$  are considered for this pattern: (i)  $n = 150, q = 200, p = 200$ ; and (ii)  $n = 200, q = 200, p = 1000$ . The response groups and covariate groups all have equal group size 20. For both the responses and covariates, the  $j$ th group consists of components with indices from  $20(j - 1) + 1$  to  $20j$ . For the coefficient blocks, it is taken that  $B_{jj} \neq 0$  and  $B_{jk} = 0, j \neq k$ . For the values of the nonzero blocks, we consider two cases:

Dense case: All entries in each  $B_{jj}$  are nonzero. The values are generated independently from the uniform distribution on  $[-5, -1] \cup [1, 5]$ .

Sparse case: 75% of the entries in each  $B_{jj}$  generated in the dense case are randomly selected and set to 0.

The covariate matrix  $X$  and the response matrix  $Y$  are generated as follows. Each row of  $X$  is independently generated group-wise with the groups iid from  $N(\mathbf{0}, \Sigma_x)$ , where  $\Sigma_x = (0.5^{|i-j|})$ . For the generation of  $Y$ , we consider two types of responses given below.

Independent response: All the entries of the error matrix  $\mathcal{E}$  are generated as iid from  $N(0, \sigma^2)$ .

Dependent response: Each row of  $\mathcal{E}$  is generated group-wise with the groups iid from  $N(\mathbf{0}, \sigma^2 \mathbf{R})$ , where  $\mathbf{R}$  is a correlation matrix with off-diagonal entries 0.5.

The  $\sigma^2$  is chosen to achieve a certain signal-to-noise ratio. It is determined as follows. After  $X$  and  $\mathbf{B}$  are generated, compute  $V_1 = \sum_{l=1}^q \text{var}(\mathbf{z}_l)$  where  $\mathbf{z}_l$  is the  $l$ th column of  $X\mathbf{B}$  and  $\text{var}(\mathbf{z}_l)$  is the sample variance. Let  $\tilde{\mathcal{E}}$  be a version of the error matrix generated with  $\sigma^2 = 1$  and compute  $V_2 = \sum_{l=1}^q \text{var}(\boldsymbol{\epsilon}_l)$  where

$\boldsymbol{\epsilon}_l$  is the  $l$ th column of  $\tilde{\mathcal{E}}$ . Then take  $\sigma^2$  as  $\sigma^2 = V_1/5V_2$ . The response matrix  $Y$  is computed as  $Y = X\mathbf{B} + \mathcal{E}$  where  $\mathcal{E} = \sigma^2 \tilde{\mathcal{E}}$ .

### 4.2. Diagonal Overlap Blocks Pattern

The same settings of  $(n, q, p)$  as for the diagonal blocks pattern are considered for this pattern. In this pattern, the groups of covariates and responses are not disjoint. Some of the groups are overlapped. The first 200 covariates and responses are grouped in the same way into eight groups as follows:

1–20; 21–40; 41–70; 61–100;  
101–120; 121–150; 141–180; 181–200.

The remaining covariate groups are nonoverlapping and have equal size 40. Again, for the coefficient blocks,  $B_{jj} \neq 0$  and  $B_{jk} = 0, j \neq k$ . Like in the diagonal blocks pattern, a dense case and a sparse case are considered for the nonzero blocks. The values of the entries of the nonzero blocks are generated in the same way as that for the diagonal blocks pattern.

The covariate matrix  $X$  and response matrix  $Y$  are generated in the same way as those in the diagonal blocks pattern except the overlapping groups are specifically handled. If two groups are overlapped, one group is generated first. For the other group, the variables already generated in the first group are not regenerated, only those nonoverlapping variables are generated. For example, for the generation of group 3 and 4 which consist of variables indexed from 41 to 70 and from 61 to 100, respectively, we first generate group 3, then for group 4, we only generate variables from 71 to 100 as if it is a nonoverlapped group. The same as in the diagonal blocks pattern, both independent and dependent responses are considered.

### 4.3. Random Blocks Pattern

The following two settings of  $(n, q, p)$  are considered under this pattern: (i)  $n = 200, q = 200, p = 250$ ; and (ii)  $n = 200, q = 200, p = 1000$ . The response variables have the following 20 nonoverlapping groups. The first 8 groups are of size 5: 1–5; ...; 35–40. The next 6 groups are of size 10: 41–50; ...; 91–100. The next 4 groups are of size 15: 101–115; ...; 146–160. The last 2 groups are of size 20: 161–180; 181–200. The covariates have nonoverlapping groups of equal size 25. Among all the coefficient blocks, we random select 10 blocks as nonzero blocks.

The other aspects of the simulation with the random blocks pattern are the same as those with diagonal blocks pattern, that is, we consider dense and sparse cases for the nonzero coefficient blocks, and consider independent and dependent responses. The generation of the values of  $\mathbf{B}$ ,  $X$  and  $Y$  are the same as that in the case of diagonal blocks pattern.

The two methods, SCCS and MSGSL, are compared in their performance and computation time. Let  $\theta_{il} = I\{\beta_{il} \neq 0\}$  where  $\beta_{il}$  are entries of  $\mathbf{B}$  and  $\hat{\theta}_{il}$  denote the estimate of  $\theta_{il}$ . The performance is measured by positive discovery rate (PDR) and false discovery rate (FDR) defined as follows

$$\text{PDR} = \frac{\#\{\hat{\theta}_{il} = \theta_{il} = 1\}}{\#\{\hat{\theta}_{il} = 1\}}, \quad \text{FDR} = \frac{\#\{\hat{\theta}_{il} = 1, \theta_{il} = 0\}}{\#\{\hat{\theta}_{il} = 1\}}.$$

**Table 1.** The average PDR, FDR, DR (PDR + (1 - FDR)), and computation time over 100 replicates in simulation studies with diagonal blocks pattern in the coefficient matrix (numbers in parentheses are standard deviations).

$(n, q, p)$	B-type	Y-type	Method	PDR	FDR	DR	Time
(150, 200, 200)	Sparse	Dep	SCCS	0.983(0.005)	0.056(0.010)	1.927(0.012)	3.024(0.282)
			MSGSL	0.902(0.009)	0.226(0.011)	1.675(0.015)	25.057(1.169)
		Ind	SCCS	0.983(0.005)	0.057(0.007)	1.926(0.009)	3.135(0.248)
			MSGSL	0.902(0.008)	0.226(0.010)	1.676(0.013)	26.392(1.454)
	Dense	Dep	SCCS	0.841(0.010)	0.000(0.000)	1.841(0.010)	3.852(0.265)
			MSGSL	0.779(0.006)	0.051(0.031)	1.728(0.031)	33.533(1.114)
		Ind	SCCS	0.841(0.011)	0.000(0.000)	1.841(0.011)	3.846(0.241)
			MSGSL	0.778(0.006)	0.035(0.027)	1.743(0.027)	33.146(0.855)
(200, 200, 1000)	Sparse	Dep	SCCS	0.992(0.003)	0.044(0.008)	1.949(0.009)	15.096(0.885)
			MSGSL	0.908(0.009)	0.199(0.011)	1.709(0.014)	219.781(24.938)
		Ind	SCCS	0.993(0.003)	0.043(0.005)	1.950(0.006)	15.621(0.910)
			MSGSL	0.908(0.009)	0.199(0.012)	1.709(0.015)	215.184(18.182)
	Dense	Dep	SCCS	0.933(0.012)	0.000(0.000)	1.933(0.012)	20.066(2.011)
			MSGSL	0.789(0.006)	0.001(0.005)	1.788(0.008)	275.388(8.539)
		Ind	SCCS	0.931(0.010)	0.000(0.000)	1.931(0.010)	17.351(0.309)
			MSGSL	0.789(0.005)	0.000(0.000)	1.789(0.005)	271.896(6.248)

**Table 2.** The average PDR, FDR, DR (PDR + (1 - FDR)), and computation time over 100 replicates in simulation studies with diagonal overlap blocks pattern in the coefficient matrix (numbers in parentheses are standard deviations).

$(n, q, p)$	B-type	Y-type	Method	PDR	FDR	DR	Time
(150, 200, 200)	Sparse	Dep	SCCS	0.796(0.066)	0.088(0.011)	1.708(0.070)	3.216(0.409)
			MSGSL	0.225(0.008)	0.261(0.022)	0.964(0.023)	26.596(1.299)
		Ind	SCCS	0.630(0.031)	0.097(0.011)	1.533(0.034)	2.788(0.359)
			MSGSL	0.225(0.008)	0.260(0.019)	0.965(0.022)	27.861(1.559)
	Dense	Dep	SCCS	0.310(0.015)	0.000(0.000)	1.310(0.015)	2.803(0.219)
			MSGSL	0.196(0.002)	0.074(0.073)	1.121(0.073)	35.787(1.29)
		Ind	SCCS	0.311(0.012)	0.000(0.000)	1.311(0.012)	2.749(0.160)
			MSGSL	0.196(0.003)	0.077(0.077)	1.119(0.078)	35.155(1.262)
(200, 200, 1000)	Sparse	Dep	SCCS	0.884(0.010)	0.071(0.010)	1.813(0.016)	14.727(1.021)
			MSGSL	0.505(0.149)	0.335(0.044)	1.170(0.145)	212.289(11.510)
		Ind	SCCS	0.868(0.033)	0.070(0.008)	1.798(0.036)	14.108(0.985)
			MSGSL	0.497(0.155)	0.328(0.048)	1.169(0.142)	218.367(21.986)
	Dense	Dep	SCCS	0.605(0.049)	0.000(0.000)	1.605(0.049)	18.533(1.648)
			MSGSL	0.766(0.005)	0.097(0.021)	1.668(0.022)	319.617(37.686)
		Ind	SCCS	0.617(0.012)	0.000(0.000)	1.617(0.012)	19.783(0.717)
			MSGSL	0.765(0.005)	0.097(0.020)	1.669(0.020)	298.620(21.649)

**Table 3.** The average PDR, FDR, DR (PDR + (1 - FDR)), and computation time over 100 replicates in simulation studies with random blocks pattern in the coefficient matrix (numbers in parentheses are standard deviations).

$(n, q, p)$	B-type	Y-type	Method	PDR	FDR	DR	Time
(200, 200, 250)	Sparse	Dep	SCCS	0.781(0.013)	0.105(0.011)	1.677(0.019)	2.882(0.212)
			MSGSL	0.900(0.011)	0.199(0.012)	1.701(0.016)	41.764(2.485)
		Ind	SCCS	0.782(0.013)	0.105(0.009)	1.677(0.019)	2.838(0.237)
			MSGSL	0.901(0.011)	0.200(0.012)	1.701(0.016)	40.258(1.432)
	Dense	Dep	SCCS	0.728(0.019)	0.000(0.000)	1.728(0.019)	3.381(0.229)
			MSGSL	0.775(0.009)	0.016(0.017)	1.759(0.018)	53.344(2.580)
		Ind	SCCS	0.727(0.009)	0.000(0.000)	1.727(0.009)	3.378(0.223)
			MSGSL	0.775(0.008)	0.013(0.015)	1.762(0.016)	52.334(1.725)
(200, 200, 1000)	Sparse	Dep	SCCS	0.781(0.012)	0.101(0.011)	1.680(0.017)	13.836(0.914)
			MSGSL	0.903(0.011)	0.198(0.014)	1.705(0.018)	231.680(41.330)
		Ind	SCCS	0.781(0.012)	0.100(0.009)	1.681(0.017)	13.755(0.673)
			MSGSL	0.903(0.011)	0.199(0.016)	1.704(0.019)	226.735(38.398)
	Dense	Dep	SCCS	0.744(0.008)	0.000(0.000)	1.744(0.008)	14.205(1.603)
			MSGSL	0.776(0.008)	0.016(0.016)	1.759(0.018)	322.016(51.967)
		Ind	SCCS	0.745(0.008)	0.000(0.000)	1.745(0.008)	11.489(1.613)
			MSGSL	0.776(0.008)	0.015(0.016)	1.760(0.017)	336.644(66.827)





Note that a performance with high PDR is not necessarily a good performance since we can simply let all  $\hat{\theta}_{il} = 1$  to achieve PDR = 1, but then we will have a FDR almost 1. Similarly, a low FDR is not necessarily good either. If one method has higher PDR as well as lower FDR than another method, it is clear-cut that the first method is better. Otherwise we should strike a balance between PDR and FDR in the comparison. In general, we consider the joint measure DR = PDR + (1 - FDR).

For each simulation setting, we replicate the simulation 100 times. The simulation results for the three patterns are reported in Tables 1–3. The findings are briefly discussed here. As shown in Tables 1 and 2, in Setting I and II, the SCCS outperforms MSGSL almost universally not only in the joint measure DR but also in both PDR and FDR: higher PDR and lower FDR, except the case of  $p = 1000$  with dense  $B$ -type. On average, the DR of SCCS is 1.179 times that of MSGSL. In Setting III, the performance of SCCS is slightly worse than but still comparable with that of MSGSL in terms of the joint measure DR. However, SCCS always has a lower FDR, which is a merit in terms of the quality of feature selection: the selected features are more trustful to be true relevant features. In terms of computation, the SCCS requires much less computation time than MSGSL.

## 5. A Real Example

We make a comparison of SCCS and MSGSL based on a real dataset in this section. In the comparison, we also include the univariate sparse group Lasso (USGL) method of Simon et al. (2013) which ignores the group structure in the response variables and is used as a baseline for the comparison. The dataset consists of 118 GeneChip (Affymetrix) microarrays for the expressions of 39 genes in the isoprenoid pathways in *Ara-bidopsis thaliano*, 21 of which are in the *mevalonate* pathway and 18 in the *nonmevalonate* pathway, and 795 additional genes from 56 downstream metabolic pathways. The dataset was reported and used in Wille et al. (2004) to construct the genetic regulatory network between the isoprenoid pathways where the downstream genes are attached to the network as conditional variables.

The construction of the genetic regulatory network can be alternatively formulated by a conditional Gaussian graphical model with group structures:  $Y = X\beta + \varepsilon$ , where  $Y$  is the matrix of the expression levels of the genes in the isoprenoid pathways which has a block structure  $Y = (Y_1, Y_2)$  and  $X$  is the matrix of the expression levels of the genes in the downstream metabolic pathways which has the block structure  $X = (X_1, \dots, X_{56})$ . The construction of the genetic regulatory network is equivalent to the inference on the nonzero entries of the precision matrix  $\Omega = \Sigma^{-1}$  where  $\Sigma$  is the common covariance matrix of the rows of  $\varepsilon$ . The inference on  $\Omega$  is beyond the scope of our current article. In the analysis of this section, we concentrate on the selection of the downstream pathways as predictors for the isoprenoid pathways.

We apply all the three methods on the dataset. Since the truth of relevant downstream pathways is unknown, the PDR, FDR, and DR cannot be used as the performance measures. Instead, we use the mean-squared prediction error (MSPE) as the performance measure for the comparison of the two

**Table 4.** The average MSPE, NNE, and computation time (numbers in parentheses are standard deviations).

	MSPE	NNE	Time (in seconds)
SCCS	0.637(0.415)	216.3(13.983)	1.215(1.736)
MSGSL	0.667(0.346)	519.65(38.833)	46.55(19.346)
USGL	51.366(10.428)	3067.21(104.19)	582.442(87.437)

methods. To this end, we make 100 random splits of the 118 microarrays into a training sample of 110 microarrays and a testing sample of 8 microarrays. For each split, the genes in the downstream pathways are selected and the  $\beta$  is estimated by ordinary least squared approach based on the selected genes in the downstream pathways using the training sample, and the prediction error is computed using the testing sample. The prediction errors over the 100 splits are averaged and reported in Table 4. Also reported in Table 4 are the average number of nonzero entries (NNE) in the estimated coefficient matrix and computation times. The average MSPE of our method, 0.637, which is less than that of MSGSL, 0.667. More significant is that to achieve a smaller MSPE the SCCS requires an even much smaller number of predictors than that of MSGSL, which is indicated by the average nonzero numbers of entries (NNE) in the estimated coefficient matrix (216 of SCCS compared with 519 of MSGSL). Implicitly, this fact implies that the genes in the downstream pathways selected by SCCS are more accurate than those by MSGSL. The computational advantage of SCCS is demonstrated again in the real data analysis. The SCCS requires only 1.215 sec on average but MSGSL requires 46.55 sec. It is obvious from Table 4 that, compared with the baseline approach USGL, both SCCS and MSGSL demonstrate a great advantage in taking into account the group structure in the response variables. Both SCCS and MSGSL reduce the MSPE of USGL by almost 99% and are much more accurate in terms of the selection of nonzero entries of the coefficient matrix.

## Supplementary Materials

The supplementary document contains the technical proofs of the theoretical results in Section 3 of the main article, i.e., Lemma 1 and Theorem 1. The document is self-contained. It starts with the introduction of all notations necessary for the discussion. Lemma 1 and its proof are given first, then followed by the proof of Theorem 1.

## References

- 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. [1227]
- Brem, R. B., and Kruglyak, L. (2005), “The Landscape of Genetic Complexity Across 5,700 Gene Expression Traits in Yeast,” *Proceedings of the National Academy of Sciences of the United States of America*, 102, 1572–1577. [1227]
- Buehlmann, P. (2006), “Boosting for High-Dimensional Linear Models,” *The Annals of Statistics*, 34, 559–583. [1228]
- Cai, T. T., and Wang, L. (2011), “Orthogonal Matching Pursuit for Sparse Signal Recovery With Noise,” *IEEE Transactions on Information Theory*, 57, 4680–4688. [1228]
- Chen, J., and Chen, Z. (2008), “Extended Bayesian Information Criteria for Model Selection With Large Model Spaces,” *Biometrika*, 95, 759–771. [1228,1230]

- Chen, L., and Huang, J. Z. (2012), "Sparse Reduced-Rank Regression for Simultaneous Dimension Reduction and Variable Selection," *Journal of the American Statistical Association*, 107, 1533–1545. [1228]
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [1228]
- 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. [1227]
- Huang, J., Ma, S., Xie, H., and Zhang, C.-H. (2009), "A Group Bridge Approach for Variable Selection," *Biometrika*, 96, 339–355. [1228]
- Ing, C.-K., and Lai, T. L. (2011), "A Stepwise Regression Method and Consistent Model Selection for High-Dimensional Sparse Linear Models," *Statistica Sinica*, 21, 1473–1513. [1228]
- Kong, Y., Li, D., Fan, Y., and Lv, J. (2017), "Interaction Pursuit in High-Dimensional Multi-Response Regression via Distance Correlation," *The Annals of Statistics*, 45, 897–922. [1228]
- Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [1228,1230]
- Li, Y., Nan, B., and Zhu, J. (2015), "Multivariate Sparse Group Lasso for the Multivariate Multiple Linear Regression With an Arbitrary Group Structure," *Biometrics*, 71, 354–363. [1228,1232]
- Luo, S., and Chen, Z. (2014), "Sequential Lasso cum EBIC for Feature Selection With Ultra-High Dimensional Feature Space," *Journal of the American Statistical Association*, 109, 1229–1240. [1228,1229]
- Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2011), "Support Union Recovery in High-Dimensional Multivariate Regression," *The Annals of Statistics*, 39, 1–47. [1228]
- Park, M. Y., and Hastie, T. (2007), "Penalized Logistic Regression for Detecting Gene Interactions," *Biostatistics*, 9, 30–50. [1227]
- Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., and Wang, P. (2010), "Regularized Multivariate Regression for Identifying Master Predictors With Application to Integrative Genomics Study of Breast Cancer," *The Annals of Applied Statistics*, 4, 53. [1228]
- Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013), "A Sparse-Group Lasso," *Journal of Computational and Graphical Statistics*, 22, 231–245. [1228,1234]
- Stein, J. L., Hua, X., Lee, S., Ho, A. J., Leow, A. D., Toga, A. W., Saykin, A. J., Shen, L., Foroud, T., Pankratz, N., and Huentelman, M. J. (2010), "Voxelwise Genome-Wide Association Study (VGWAS)," *NeuroImage*, 53, 1160–1174. [1227]
- Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267–288. [1227]
- Tropp, J. A. (2004), "Greed Is Good: Algorithmic Results for Sparse Approximation," *IEEE Transactions on Information Theory*, 50, 2231–2242. [1229]
- Turlach, B. A., Venables, W. N., and Wright, S. J. (2005), "Simultaneous Variable Selection," *Technometrics*, 47, 349–363. [1228]
- Wang, H. (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association*, 104, 1512–1524. [1228]
- Wille, A., Zimmermann, P., Vranová, E., Fürholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., and Zitzler, E. (2004), "Sparse Graphical Gaussian Modeling of the Isoprenoid Gene Network in *Arabidopsis thaliana*," *Genome Biology*, 5, R92. [1234]
- Yuan, M., Ekici, A., Lu, Z., and Monteiro, R. (2007), "Dimension Reduction and Coefficient Estimation in Multivariate Linear Regression," *Journal of the Royal Statistical Society, Series B*, 69, 329–346. [1228]
- 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. [1228]
- Zamdborg, L., and Ma, P. (2009), "Discovery of Protein–DNA Interactions by Penalized Multivariate Regression," *Nucleic Acids Research*, 37, 5246–5254. [1227]
- Zhang, C. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [1227]
- Zhang, S.-Q., Ching, W.-K., Tsing, N.-K., Leung, H.-Y., and Guo, D. (2010), "A New Multiple Regression Approach for the Construction of Genetic Regulatory Networks," *Artificial Intelligence in Medicine*, 48, 153–160. [1227]
- Zhou, H., Sehl, M. E., Sinsheimer, J. S., and Lange, K. (2010), "Association Screening of Common and Rare Genetic Variants by Penalized Regression," *Bioinformatics*, 26, 2375. [1227]
- Zhou, N., and Zhu, J. (2010), "Group Variable Selection via a Hierarchical Lasso and Its Oracle Property," arXiv no. 1006.2871. [1228]
- Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [1227]