Biometrics 72, 85–94
March 2016
DOI: 10.1111/biom.12374

Detection of Gene–Gene Interactions Using Multistage Sparse and Low-Rank Regression

Hung Hung,[1] Yu-Ting Lin,[2] Penweng Chen,[3] Chen-Chien Wang,[4] Su-Yun Huang,[2] and Jung-Ying Tzeng[5,6,*]

[1]Institute of Epidemiology and Preventive Medicine, National Taiwan University, Taipei 100, Taiwan
[2]Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan
[3]Department of Applied Mathematics, National Chung Hsing University, Taichung 402, Taiwan
[4]Yahoo, Sunnyvale, California 94089, U.S.A.
[5]Department of Statistics and Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina 27695, U.S.A.
[6]Department of Statistics, National Cheng Kung University, Tainan 701, Taiwan
email: jytzeng@ncsu.edu

Summary. Finding an efficient and computationally feasible approach to deal with the curse of high-dimensionality is a daunting challenge faced by modern biological science. The problem becomes even more severe when the interactions are the research focus. To improve the performance of statistical analyses, we propose a sparse and low-rank (SLR) screening based on the combination of a low-rank interaction model and the Lasso screening. SLR models the interaction effects using a low-rank matrix to achieve parsimonious parametrization. The low-rank model increases the efficiency of statistical inference and, hence, SLR screening is able to more accurately detect gene–gene interactions than conventional methods. Incorporation of SLR screening into the Screen-and-Clean approach (Wasserman and Roeder, 2009; Wu et al., 2010) is also discussed, which suffers less penalty from Boferroni correction, and is able to assign p-values for the identified variables in high-dimensional model. We apply the proposed screening procedure to the Warfarin dosage study and the CoLaus study. The results suggest that the new procedure can identify main and interaction effects that would have been omitted by conventional screening methods.

Key words: Asymptotic normality; Gene–gene interactions; Low-rank approximation; Over-parametrization; Screen-and-Clean; Sparsity.

1. Introduction

Usage of high-throughput data has revolutionized modern biological research and has yielded significant insights for cancer research, genetic diseases, and many genetic disorders. Yet, the ever-increasing data dimensions lead to the inevitable curse of high-dimensionality, where conventional methods in genetic association analysis lose power and often fail to detect meaningful signals. The problem is further exacerbated when the question of interest focuses on gene–gene interactions (G×G), where the dimensionality of the model explodes even faster. Let Y be the response of interest and $\boldsymbol{g} = (g_1, \ldots, g_p)^T$ be the genotypes at the p loci. The most widely used method to detecting G×G is marginal scan (MS) such as PLINK (Purcell et al., 2007) and BOOST (Wan et al., 2010), where the main effect terms were identified by testing for $H_{0,j} : c_j = 0$ in $E(Y|g_j) = c_0 + c_j g_j$, and the interaction terms were identified by testing for $H_{0,jk} : c_{jk} = 0$ in $E(Y|g_j, g_k) = c_0 + c_j g_j + c_k g_k + c_{jk}(g_j g_k)$. MS has the advantages of using minimal parameters in model fitting and the ease of implementation. However, it has two major drawbacks: (i) MS is too conservative due to multiple testing correction when p is large; and (ii) MS has low detecting power due to the ignorance of joint effects.

To improve on (i), newly developed G×G detection methods adopt a multistage strategy where a screening step is used to reduce the number of variables (e.g., Wu et al., 2010). To improve on (ii), joint screening (based on a multilocus model including all main and G×G effects) is used to identify candidate genes. Joint screening can better identify loci that interact with each other but exhibit little marginal effects; it also improves the overall screening performance by reducing the unexplained variance in the model (Wu et al., 2010). A commonly used joint-screening method for G×G detection is to apply Lasso on the model

$$E(Y|\boldsymbol{g}) = \gamma + \sum_{j=1}^{p} \xi_j g_j + \sum_{j<k} \eta_{jk}(g_j g_k), \tag{1}$$

where ξ_j is the main effect of the jth locus, and η_{jk} is the G×G that corresponds to the jth and kth loci. The performance of

Lasso depends on the number of parameters in model (1)

$$m_p = 1 + p + \binom{p}{2} \tag{2}$$

and the available sample size, n. Although it has been verified that Lasso performs well when m_p is large, caution should be used when m_p is ultralarge, e.g., on the order of $\exp\{O(n^\delta)\}$ for some $\delta > 0$ (Fan and Lv, 2008). In addition, the m_p encountered in modern biomedical studies are usually much larger than n even if p is of moderate size. In this situation, statistical inferences can become unstable and inefficient, which impact the screening performance.

To improve the joint screening of all main and G×G effects, we utilize the matrix nature of the interaction terms and consider a reduced model. In model (1), $(g_j\, g_k)$ is the (j,k)th element of the symmetric matrix $\boldsymbol{J} = \boldsymbol{g}\boldsymbol{g}^T$, and it is natural to treat η_{jk} as the (j,k)th entry of the symmetric matrix parameter $\boldsymbol{\eta}$. Thus, an equivalent expression of model (1) is

$$E(Y|\boldsymbol{g}) = \gamma + \boldsymbol{\xi}^T\boldsymbol{g} + \mathrm{vecp}(\boldsymbol{\eta})^T\mathrm{vecp}(\boldsymbol{J}), \tag{3}$$

where $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_p)^T$ is the coefficient vector of main effects, and $\mathrm{vecp}(\cdot)$ denotes the operator that stacks the lower half (excluding diagonals) of a symmetric matrix columnwisely into a long vector. With the equivalent model expression (3), we can utilize the structure of $\boldsymbol{\eta}$ to improve the inference procedure. Specifically, we posit the condition

$$\boldsymbol{\eta} : \text{being sparse and low rank.} \tag{4}$$

Condition (4) is typically satisfied in modern biomedical research. First, in a G×G scan, it is reasonable to assume that only a few pairs of G×G, say at most q pairs, are associated with the response, where q is much smaller than the number of genes p. This sparsity assumption is also the underlying rationale for applying Lasso for variable selection in conventional approaches. Second, the rank of such a symmetric G×G coefficient matrix $\boldsymbol{\eta}$ can be at most $2q$, i.e., $\mathrm{rank}(\boldsymbol{\eta}) \le 2q \ll p$, which motivates the low-rank assumption on $\boldsymbol{\eta}$. Displayed below is an example of $\boldsymbol{\eta}$ with $p = 10$ and $q = 3$, and its rank is only three:

$$\boldsymbol{\eta} = \begin{bmatrix} 0 & \bigstar & \spadesuit & \\ \bigstar & 0 & \blacklozenge & \mathbf{0}_{3\times 7} \\ \spadesuit & \blacklozenge & 0 & \\ & \mathbf{0}_{7\times 3} & & \mathbf{0}_{7\times 7} \end{bmatrix}. \tag{5}$$

One key characteristic of our method is the consideration of the sparse and low-rank condition (4), which allows us to express $\boldsymbol{\eta}$ using far fewer parameters than $\binom{p}{2}$. In contrast, Lasso assumes that $\boldsymbol{\eta}$ is sparse, but without using the matrix low-rank structure, $\boldsymbol{\eta}$ still contains $\binom{p}{2}$ parameters. From a statistical viewpoint, parsimonious parametrization can improve the efficiency of model inferences. The aim of this work is thus to use model (3) and condition (4) to propose an efficient joint-screening procedure, referred to as sparse and low-rank (SLR) screening. We will also demonstrate that the performances of existing multistage G×G detection methods can be enhanced by incorporating SLR screening.

Some notation is defined here for reference. Let $\{(Y_i, \boldsymbol{g}_i)\}_{i=1}^n$ be random copies of $(Y, \boldsymbol{g})$, and let $\boldsymbol{J}_i = \boldsymbol{g}_i\boldsymbol{g}_i^T$. Let $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$ be an n-vector of observed responses, and let $\boldsymbol{X}$ be the $n \times m_p$ matrix with the ith row being $\{1, \boldsymbol{g}_i^T, \mathrm{vecp}(\boldsymbol{J}_i)^T\}$. For any square matrix $\boldsymbol{M}$, $\boldsymbol{M}^+$ is its Moore–Penrose generalized inverse. The operator that stacks a matrix columnwise into a vector is $\mathrm{vec}(\cdot)$. The commutation matrix, $\boldsymbol{K}_{p,k}$, is defined such that $\boldsymbol{K}_{p,k}\,\mathrm{vec}(\boldsymbol{M}) = \mathrm{vec}(\boldsymbol{M}^T)$ for any $p \times k$ matrix $\boldsymbol{M}$ (Magnus and Neudecker, 1979). The matrix $\boldsymbol{P}$ satisfies $\boldsymbol{P}\mathrm{vec}(\boldsymbol{M}) = \mathrm{vecp}(\boldsymbol{M})$ for any $p \times p$ symmetric matrix $\boldsymbol{M}$, and can be chosen such that $\boldsymbol{P}\boldsymbol{K}_{p,p} = \boldsymbol{P}$. The symbol $\otimes$ denotes the Kronecker product. For a vector, $\|\cdot\|_1$ is its 1-norm, and $\|\cdot\|$ is its Euclidean norm (2-norm). The cardinality of a set is $|\cdot|$.

This article is organized as follows. The inference procedures for the low-rank model are discussed in Section 2. The SLR screening and its extensions are discussed in Sections 3 and 4. Sections 5 and 6 provide numerical studies to demonstrate the utility of the proposed screening procedures. The article ends with a discussion in Section 7.

2. Inference Procedure for the Low-Rank Model

2.1. *Model Specification and Estimation*

To incorporate the low-rank property (4) for G×G into model building, we propose the following rank-r model, where $r \ll p$ is a prespecified positive integer:

$$E(Y|\boldsymbol{g}) = \gamma + \boldsymbol{\xi}^T\boldsymbol{g} + \mathrm{vecp}(\boldsymbol{\eta})^T\mathrm{vecp}(\boldsymbol{J}), \quad \mathrm{rank}(\boldsymbol{\eta}) \le r. \tag{6}$$

Although the above low-rank model expression is straightforward, it is not convenient for numerical implementation. Therefore, by spectral decomposition theorem, we use an equivalent parametrization, $\boldsymbol{\eta}(\boldsymbol{\phi})$, that directly satisfies the constraint $\mathrm{rank}\{\boldsymbol{\eta}(\boldsymbol{\phi})\} \le r$:

$$\begin{aligned} \boldsymbol{\eta}(\boldsymbol{\phi}) &= \boldsymbol{A}\boldsymbol{U}\boldsymbol{A}^T, \quad \boldsymbol{A} \in \mathbb{R}^{p\times r}, \\ \boldsymbol{U} &= \mathrm{diag}(\boldsymbol{u}), \quad \boldsymbol{u} \in \mathbb{R}^r, \\ \boldsymbol{\phi} &= \{\mathrm{vec}(\boldsymbol{A})^T, \boldsymbol{u}^T\}^T. \end{aligned} \tag{7}$$

Note that the number of parameters required in (7) for interactions $\boldsymbol{\eta}(\boldsymbol{\phi})$ can be much smaller than $\binom{p}{2}$. In the case of $r = 1$, for example, $\boldsymbol{A} = (a_1, \ldots, a_p)^T$ and it is equivalent to model the G×G of (g_i, g_j) as the product $(\boldsymbol{u}a_ia_j)$, i.e., using a total of $p + 1$ parameters to model G×G. See Remark 1 for further details. When model (6) is true, standard MLE arguments show that statistical inference based on model (6) must be the most efficient. Even if model (6) is incorrectly specified, we still favor the low-rank model when the sample size is small. We note that the concept of fitting a low-rank model is similar to applying singular value decomposition (SVD) and aiming to find the "best" rank-r approximation of the true $\boldsymbol{\eta}$, "best" in the sense of maximizing the likelihood function. It thus provides a good "working" model that compromises between model approximation bias and efficiency of parameters estimation. With limited sample size, it is preferable to more

efficiently estimate the approximated low-rank model than to obtain an unstable estimate of the full model.

Let the parameters of interest in rank-r model (6) be

$$\boldsymbol{\beta}(\boldsymbol{\theta}) = \left[\gamma, \boldsymbol{\xi}^T, \text{vecp}\{\boldsymbol{\eta}(\boldsymbol{\phi})\}^T\right]^T \quad \text{with} \quad \boldsymbol{\theta} = \left(\gamma, \boldsymbol{\xi}^T, \boldsymbol{\phi}^T\right)^T, \quad (8)$$

which comprises an intercept, main effects, and interaction terms. Under model (6) and assuming independent and identically distributed errors from a normal distribution $N(0, \sigma^2)$, the log-likelihood function (apart from a constant term) is proportional to

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= -\frac{1}{2}\sum_{i=1}^{n}\left[Y_i - \gamma - \boldsymbol{\xi}^T\boldsymbol{g}_i - \text{vecp}\{\boldsymbol{\eta}(\boldsymbol{\phi})\}^T\text{vecp}(\boldsymbol{J}_i)\right]^2 \\ &= -\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}(\boldsymbol{\theta})\|^2. \end{aligned} \quad (9)$$

To further stabilize the MLE, a common approach is to add a penalty on $\boldsymbol{\theta}$. We propose to estimate $\boldsymbol{\theta}$ by maximizing the penalized log-likelihood function

$$\ell_{\lambda_l}(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \frac{\lambda_l}{2}\|\boldsymbol{\theta}\|^2, \quad (10)$$

where $\lambda_l \geq 0$ is the penalty (the subscript l is for low-rank). Let $\widehat{\boldsymbol{\theta}}_{\lambda_l} = \text{argmax}_{\boldsymbol{\theta}}\, \ell_{\lambda_l}(\boldsymbol{\theta})$ be the penalized MLE from maximizing (10). The parameter of interest, $\boldsymbol{\beta}(\boldsymbol{\theta})$, is estimated by

$$\widehat{\boldsymbol{\beta}}_{\lambda_l} = \boldsymbol{\beta}(\widehat{\boldsymbol{\theta}}_{\lambda_l}). \quad (11)$$

In practical implementation, we use fivefold cross-validation to select λ_l. Detailed implementation algorithms for obtaining $\widehat{\boldsymbol{\theta}}_{\lambda_l}$ are described in Web Appendix A.

REMARK 1. *In (7), $\boldsymbol{\phi}$ is over-parameterized. Indeed, only $pr - r(r-1)/2$ identifiable parameters are required to specify a $p \times p$ rank-r symmetric matrix. One can impose the constraints $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{I}_r$ to make the parametrization identifiable. However, such identifiability constraints, which are merely to obtain a unique solution for $\boldsymbol{A}$, unnecessarily complicate the numerical implementation. Since the parameters of interest are $\boldsymbol{\eta}(\boldsymbol{\phi})$ but not $\boldsymbol{\phi}$ itself, we keep the simple usage of over-parameterized $\boldsymbol{\phi}$ without imposing any constraint on $\boldsymbol{A}$.*

2.2. *Asymptotic Properties*

To derive the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\lambda_l}$ in (11), we assume that the parameter space Θ of $\boldsymbol{\theta}$ is bounded, open, and connected. Define $\boldsymbol{V}_0 = E(\boldsymbol{V}_n)$ with $\boldsymbol{V}_n = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$. Let $\boldsymbol{\beta}_0 = \boldsymbol{\beta}(\boldsymbol{\theta}_0)$ for some $\boldsymbol{\theta}_0 \in \Theta$ be the true parameter value of model (6), and define $\boldsymbol{\Delta}_0 = \boldsymbol{\Delta}(\boldsymbol{\theta}_0)$, where

$$\begin{aligned}\boldsymbol{\Delta}(\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\beta}(\boldsymbol{\theta}) \\ &= \begin{bmatrix} \boldsymbol{I}_{p+1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & 2\boldsymbol{P}(\boldsymbol{A}\boldsymbol{U}\otimes\boldsymbol{I}_p) & \left[\text{vecp}(\boldsymbol{A}_1\boldsymbol{A}_1^T), \ldots, \text{vecp}(\boldsymbol{A}_r\boldsymbol{A}_r^T)\right] \end{bmatrix} \end{aligned} \quad (12)$$

with $\boldsymbol{A}_j$ being the jth column of $\boldsymbol{A}$. The result is summarized below, and the proof is given in the Web Appendix B.

THEOREM 1. *Assume model (6) and conditions (C1)–(C3) in the Web Appendix B. Assume also $\lambda_l/n = o(n^{-1/2})$. Then, we have $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\lambda_l} - \boldsymbol{\beta}_0) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma}_0)$ as $n \to \infty$, where $\boldsymbol{\Sigma}_0 = \sigma^2\boldsymbol{\Delta}_0(\boldsymbol{\Delta}_0^T\boldsymbol{V}_0\boldsymbol{\Delta}_0)^+\boldsymbol{\Delta}_0^T$.*

The asymptotic covariance matrix $\boldsymbol{\Sigma}_0$ can be estimated by

$$\widehat{\boldsymbol{\Sigma}}_0 = \widehat{\sigma}^2 \cdot \widehat{\boldsymbol{\Delta}}_0\left(\widehat{\boldsymbol{\Delta}}_0^T\boldsymbol{V}_n\widehat{\boldsymbol{\Delta}}_0 + \frac{\lambda_l}{n}\boldsymbol{I}\right)^+\widehat{\boldsymbol{\Delta}}_0^T, \quad (13)$$

where $\widehat{\sigma}^2 = \|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\lambda_l}\|^2/(n - d_r)$ and $\widehat{\boldsymbol{\Delta}}_0 = \boldsymbol{\Delta}(\widehat{\boldsymbol{\theta}}_{\lambda_l})$ are estimators of σ^2 and $\boldsymbol{\Delta}_0$, and $d_r = 1 + p + \{pr - r(r-1)/2\}$ is the number of parameters required to specify model (6). Here $\frac{\lambda_l}{n}\boldsymbol{I}$ was added in (13) to stabilize the estimator, $\widehat{\boldsymbol{\Sigma}}_0$, and will not affect its consistency to $\boldsymbol{\Sigma}_0$. The over-parameterized nature of $\boldsymbol{\beta}(\boldsymbol{\theta})$ can be observed from Theorem 1 that the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\lambda_l}$ depends on $\boldsymbol{\theta}_0$ only through the space $\text{span}(\boldsymbol{\Delta}_0)$. Theorem 1 further implies that, as mentioned in Remark 1, adding constraints on $\boldsymbol{\phi}$ does not affect the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\lambda_l}$ because the resulting space, $\text{span}(\boldsymbol{\Delta}_0)$, is unchanged.

REMARK 2. *We suggest to use d_r to guide the determination of the maximum model rank r with given (n, p). That is, $n - d_r$ in (13) should be adequate for error variance estimation. One can further use cross-validation to select the rank within the proper range.*

3. Sparse and Low-Rank (SLR) Screening for Genetic Main and G×G Effects

We propose a screening procedure called SLR screening based on the low-rank model (6). The main idea of SLR screening, as summarized below, is to filter out insignificant variables by first fitting a low-rank model and then fitting Lasso on the remaining variables.

Sparse and Low-Rank Screening (SLR Screening)

(S1) **Low-Rank Screening:** Fit the low-rank model (6). Examine the significance of $\widehat{\boldsymbol{\beta}}_{\lambda_l}$ to form the index set $\mathcal{I}_{\text{LR}}$.

(S2) **Sparse (Lasso) Screening:** Fit Lasso on $\mathcal{I}_{\text{LR}}$. Variables with nonzero estimates are included in the index set $\mathcal{I}_{\text{SLR}}$.

The goal of Stage-(S1) in SLR screening is to utilize the low-rank property of $\boldsymbol{\eta}$ to efficiently identify important variables. By Theorem 1, the candidate set of variables, $\mathcal{I}_{\text{LR}}$, is formed as

$$\mathcal{I}_{\text{LR}} = \left\{1 < j \leq m_p : |\widehat{\boldsymbol{\beta}}_{\lambda_l, j}|/\sqrt{n^{-1}\,\widehat{\boldsymbol{\Sigma}}_{0,j}} > \alpha_l\right\} \quad (14)$$

for some $\alpha_l > 0$, where $\widehat{\boldsymbol{\beta}}_{\lambda_l, j}$ is the jth element of $\widehat{\boldsymbol{\beta}}_{\lambda_l}$, and $\widehat{\boldsymbol{\Sigma}}_{0,j}$ is the jth diagonal element of $\widehat{\boldsymbol{\Sigma}}_0$. Note that $\widehat{\boldsymbol{\beta}}_{\lambda_l, 1}$ is the estimate

of the intercept term, and is exempted from the screening procedure in (14). The threshold value, α_l, controls the power of the low-rank screening. To ensure the selection consistency of the multistage variable selection procedure, a critical condition is to require $\lim_{n\to\infty} P(\mathcal{I}_0 \subseteq \mathcal{I}_{\mathrm{LR}}) = 1$, where $\mathcal{I}_0$ denotes the true active set of variables. However, this condition holds for any fixed $\alpha_l > 0$, and thus cannot be used to determine α_l. (That is, for any $j \in \mathcal{I}_0$, we have $\lim_{n\to\infty} |\widehat{\boldsymbol{\beta}}_{\lambda_l,j}|/(n^{-1}\widehat{\boldsymbol{\Sigma}}_{0,j})^{1/2} = \infty$ under the validity of model (6), and thus $j \in \mathcal{I}_{\mathrm{LR}}$ with probability tending to one.) To simplify the procedure, we suggest to set α_l such that

$$|\mathcal{I}_{\mathrm{LR}}| = n, \tag{15}$$

because the main purpose of low-rank screening is to reduce model size so that the subsequent Lasso screening can be more efficiently implemented, and because a linear model cannot identify more than n parameters. This rule shares the same idea as Fan and Lv (2008), which aims to reduce the model size from an ultrahigh scale to a relatively large scale. In practice, a smaller size, such as $n/3$, for $|\mathcal{I}_{\mathrm{LR}}|$ may also be considered; the choice can be made based on the size of n and the characteristics of the underlying study. The rule also implicitly assumes that $|\mathcal{I}_0| \le n$, which is commonly satisfied because $|\mathcal{I}_0|$ is usually small comparing to n.

The goal of Stage-(S2) in SLR screening is to enforce sparsity. Based on the selected index set, $\mathcal{I}_{\mathrm{LR}}$, we refit the model with a 1-norm penalty by minimizing

$$\frac{1}{2}\|\boldsymbol{Y} - b_0\mathbf{1}_n - \boldsymbol{X}_{\mathcal{I}_{\mathrm{LR}}}\boldsymbol{\beta}_{\mathcal{I}_{\mathrm{LR}}}\|^2 + \lambda_s\|\boldsymbol{\beta}_{\mathcal{I}_{\mathrm{LR}}}\|_1, \tag{16}$$

where $\boldsymbol{X}_{\mathcal{I}_{\mathrm{LR}}}$ and $\boldsymbol{\beta}_{\mathcal{I}_{\mathrm{LR}}}$ are the selected variables and parameters in $\mathcal{I}_{\mathrm{LR}}$, respectively, b_0 is the intercept, and $\lambda_s \ge 0$ is the penalty for the sparsity constraint. The fivefold cross-validation is applied to determine λ_s. Let the minimizer of (16) be $\widehat{\boldsymbol{\beta}}_{\mathcal{I}_{\mathrm{LR}}}$ with $\widehat{\boldsymbol{\beta}}_{\mathcal{I}_{\mathrm{LR}},j}$ being its jth element. The final identified main effects and interaction terms from SLR screening are

$$\mathcal{I}_{\mathrm{SLR}} = \left\{ j \in \mathcal{I}_{\mathrm{LR}} : \widehat{\boldsymbol{\beta}}_{\mathcal{I}_{\mathrm{LR}},j} \neq 0 \right\}. \tag{17}$$

Subsequent analysis is conducted on the variables in $\mathcal{I}_{\mathrm{SLR}}$.

We close this section by providing two comments for SLR screening. First, the conventional single-staged Lasso screening has to deal with m_p parameters. Typically, $n \ll m_p$ and the fitting may be inefficient. However, SLR screening is a two-staged implementation that first fits a rank-r interaction model and then fits a sparse model. Fitting all main and G×G effects using a rank-r model requires only d_r parameters. Because $d_r < n \ll m_p$ when r is small, the low-rank screening (14) is able to efficiently reduce the model size from m_p to a relatively small number, so that the subsequent Lasso screening (17) can achieve higher detecting power. See Figure 1 for a detailed illustration. Second, although in theory the proposed SLR screening assumes the validity of the low-rank model (6), we observed that its practical performance is robust to model misspecification. As will be shown in our numerical studies, a small value of r suffices to filter out most irrelevant variables while keeping relevant ones. This is an appealing observation, which increases the applicability of SLR screening.

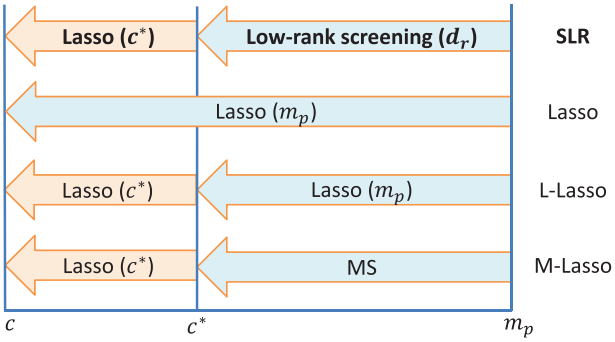

Figure 1. Illustration of SLR for identifying c candidate variables from m_p main and G×G effects. SLR first performs a low-rank screening using d_r parameters to reduce the model size from m_p to c^* (e.g., $c^* = n$), and then fits Lasso (with c^* parameters) to select c candidates. Since $d_r < n \ll m_p$, the number of required parameters in the two steps of SLR screening (i.e., d_r and c^*) can be smaller than n. In contrast, Lasso screening uses m_p ($\gg n$) parameters to select c candidates directly. L-Lasso and M-Lasso replace the low-rank screening by Lasso and marginal scan (MS), respectively.

4. Extensions of SLR Screening

4.1. *Extended Screen-and-Clean (ESC)*

We discuss how SLR screening can be incorporated into the well-received Screen-and-Clean approaches (Wasserman and Roeder, 2009), with an aim to address two common issues encountered in real practice. First, in many practices it is of interest to obtain the p-values of the identified variables. Second, the number of parameters required for low-rank model (6), say d_r, can still be very large when the original p is in the size of hundreds of thousand (i.e., $n \ll d_r \ll m_p$), which makes variable selection difficult. We tackle these issues in this section.

The Screen-and-Clean of Wasserman and Roeder (2009) is a novel variable selection procedure that is able to assign p-vales in high-dimensional model. First, the data $\mathcal{D}$ are randomly split into two parts, where $\mathcal{D}_1$ is for screening and $\mathcal{D}_2$ is for cleaning. In the screening stage, Lasso is used to fit all covariates, from which zero estimates are dropped. In the cleaning stage, the least squares estimate (LSE) is applied on those variables survived through the screening step to identify significant covariates based on the p-values assigned to them (with Bonferroni correction). The reduction of the model size by Lasso guarantees the success of using LSE to identify important covariates. Recently, Screen-and-Clean is modified by Wu et al. (2010) to detect G×G (referred to as SC in the rest of discussion), where the authors propose to fit Lasso to model (1) in the screening stage, and use LSE to assign p-values in the cleaning stage. SC has the advantage of being a joint-screening method. SC also suffers less penalty from the Bonferroni correction, since the number of tests needed to be adjusted is the number of covariates entering the cleaning stage only. Similarly, the data splitting idea can be incorpo-

rated into SLR screening to obtain p-values. That is, one can use $\mathcal{D}_1$ for SLR screening and use $\mathcal{D}_2$ for obtaining LSEs and p-values.

When the number of main effect terms is large (i.e., $n \ll d_r$) which makes SLR screening less efficient, we follow the idea of SC to start with a prescreening step using Lasso on main effects only before applying SLR screening (denote $\boldsymbol{g}_{\text{Lasso}}$ as the identified set). The underlying rationale is that interactive factors tend to exhibit marginal effects even when the interaction terms are not modeled (Cordell, 2002; Hirschhorn and Daly, 2005) and, hence, a Lasso prescreening is used to rule out loci with no effects. This prescreening, however, may miss interactions with weak marginal main effects. A remedy to this situation is to further include genes identified by MS. Specifically, we suggest to apply PLINK (based on $\mathcal{D}_1$) to produce a list $\boldsymbol{g}_{\text{MS}}$, where $g_i \in \boldsymbol{g}_{\text{MS}}$ if either g_i or $(g_i g_j)$ for some j are identified by PLINK. Then, take $\boldsymbol{g} = \boldsymbol{g}_{\text{Lasso}} \bigcup \boldsymbol{g}_{\text{MS}}$ to enter SLR screening.

The above two ideas are summarized in the Extended Screen-and-Clean (ESC), which is able to assign p-values for SLR screening. Let $\boldsymbol{g}^*$ be the set of whole genes under consideration, and let α be the family-wise error rate.

Extended Screen-and-Clean (ESC) (see Figure 2)

(E1) Based on $\mathcal{D}_1$, fit Lasso on $(Y, \boldsymbol{g}^*)$ to obtain $\boldsymbol{g}_{\text{Lasso}}$ (i.e., genes with nonzero estimates), and fit MS on $(Y, \boldsymbol{g}^*)$ to obtain $\boldsymbol{g}_{\text{MS}}$ (i.e., significant gene pairs). Set $\boldsymbol{g} = \boldsymbol{g}_{\text{Lasso}} \bigcup \boldsymbol{g}_{\text{MS}}$.

(E2) Based on $\mathcal{D}_1$, implement SLR screening on $(Y, \boldsymbol{g})$ to obtain $\mathcal{I}_{\text{SLR}}$. Let $\mathcal{S}$ consist of the main effects and interaction terms in $\mathcal{I}_{\text{SLR}}$.

(E3) Based on $\mathcal{D}_2$, fit LSE on $(Y, \mathcal{S})$ to obtain the p-values of main effect ξ_j and interaction η_{kl} as $p_j^{(\xi)}$ and $p_{kl}^{(\eta)}$. The chosen model is $\mathcal{M} = \left\{ g_j, (g_k g_l) \in \mathcal{S} : p_j^{(\xi)} < \alpha/|\mathcal{S}|, p_{kl}^{(\eta)} < \alpha/|\mathcal{S}| \right\}$.

To determine the 1-norm penalty λ_m of Lasso in Step-(E1), the StARS (Stability Approach to Regularization Selection) of Liu, Roeder, and Wasserman (2010) is applied, and this criterion is adopted in the R code of *Screen & Clean* (`http://wpicr.wpic.pitt.edu/WPICCompGen/`). We remind the readers that Step-(E1) of ESC is only required when d_r is large. For moderate size of p, we can start from Step-(E2) directly. Moreover, when p-values are not of interest, Steps-(E1)–(E2) (if d_r is large) or Step-(E2) (if d_r is moderate) can be directly applied on the entire data, $\mathcal{D}$, without Step-(E3). In practical implementation, we suggest to include in $\boldsymbol{g}^*$ only the top, say 5000, covariates identified by the sure independence screening (SIS) of Fan and Lv (2008). It is our experience that this preselection procedure will save computational cost without sacrificing the detecting power. Note that the main difference between ESC and SC is that in Step-(E2), ESC fits SLR screening on model (6), while SC fits Lasso screening on model (1). See Figure 2 for the flowchart of ESC.

We close this section by emphasizing the differences between the multistage joint screening-based G×G detection methods (e.g., ESC and SC) and MS. The MS, as mentioned

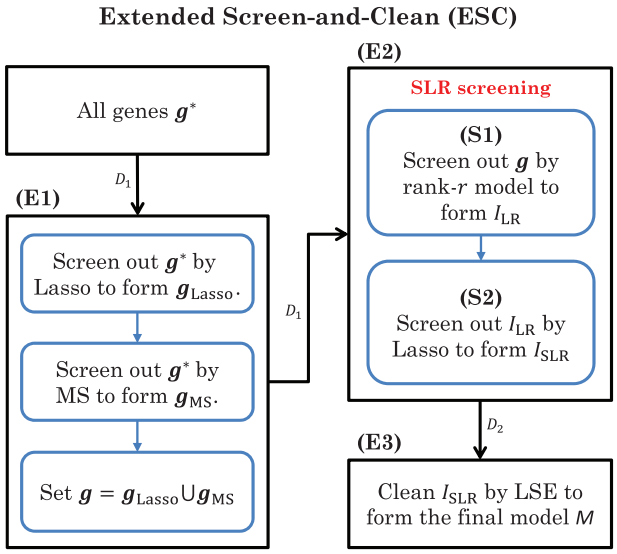

Figure 2. Flowchart of ESC ((E1)–(E3)) to detecting G×G, where SLR screening ((S1)–(S2)) is implemented in (E2). The arrows indicate which part of the data is used, where $\mathcal{D}_1$ is for screening and $\mathcal{D}_2$ is for cleaning. The SC procedure ignores the low-rank screening (S1). The LSC and MSC procedures replace the low-rank screening (S1) by Lasso and marginal scan (MS), respectively.

in Section 1, has the drawbacks of ignoring joint effects and being too conservative due to Bonferroni correction. Multistage joint-screening method is thus more preferable, since the number of variables entering the cleaning stage can be rather small and, hence, suffers less penalty from Bonferroni correction. SC, however, will suffer the problem of inefficiency when fitting model (1). The proposed ESC is an intermediate method between MS and SC. On one hand, ESC is a multistage joint-screening method which avoids the drawbacks of MS. On the other hand, ESC overcomes the inefficiency problem of SC by fitting the low-rank model (6). The superiority of ESC will be demonstrated by simulations in Section 5.

4.2. *ESC with Aggregated p-Values via Multisplit Technique*

The data partition in ESC allows the calculations of p-values and the control of the error rates, but the resulting performance is sensitive to the random partition of $\mathcal{D}$. To reduce the impact of random partition, one possible solution is to apply the multisplit technique of Meinshausen et al. (2009) to obtain an aggregated p-value as summarized below:

(i) For $b = 1, \ldots, B$, obtain a random partition, $\mathcal{D} = \mathcal{D}_1^{(b)} \bigcup \mathcal{D}_2^{(b)}$ and perform ESC. For the variables selected in the final model for the bth replicate, output the p-values, $\widetilde{p}_j^{(b)}$, using the LSEs in the cleaning stage. Otherwise, set $\widetilde{p}_j^{(b)} = 1$.

(ii) Calculate $p_j^{(b)} = \min\{1, m^{(b)} \widetilde{p}_j^{(b)}\}$, where $m^{(b)}$ is the number of variables entering LSE cleaning in the bth replicate.

(iii) Fix $s_{\min} \in (0,1)$. Calculate $p_j = \min\left\{1, (1-\log s_{\min}) \inf_{s\in(s_{\min},1)} Q_j(s)\right\}$, where $Q_j(s) = \min\left\{1, q_s(\{p_j^{(b)}/s : b = 1,\ldots,B\})\right\}$ and $q_s(\cdot)$ is the s-quantile function.

Meinshausen et al. (2009) showed that selecting variables by $\{j : p_j \le \alpha\}$ controls the family-wise error rate at level α, and can achieve higher detecting power than the single-split method. They also showed that a direct method (e.g., apply SLR screening on the whole data $\mathcal{D}$ to identify variables directly) usually achieves a higher true positive rate than its multisplit version, but at the cost of having a higher false positive rate. Thus, either the direct method or the multisplit method has its own merits, and the choice depends on the purpose of the underlying study.

5. Simulation Studies

5.1. *Simulation Settings*

Our simulation models are based on the design considered in Wu et al. (2010) with some extensions, where the genotypes $\boldsymbol{g}$ is generated using the CoLaus dataset (see Section 6.2 for detailed description). Given $\boldsymbol{g}$, Y is generated from two different models, M1 and M2, where $\beta \in \{0.25, 0.5, \ldots, 1.5\}$ is the effect size and $\varepsilon \sim N(0,1)$:

M1: $Y = \beta(g_5g_6 + 0.8g_5g_{10} + 0.6g_6g_{10} + g_{11}g_{16} + g_{11}g_{21} + 2g_{10} + 2g_{11}) + \varepsilon$.

M2: $Y = \beta \mathrm{vecp}(\boldsymbol{\eta})^T \mathrm{vecp}(\boldsymbol{J}) + \varepsilon$, where we randomly generate $\eta_{jk} = \mathrm{sign}(u_1)\cdot u_2$ with $u_1 \sim U(-1,9)$ and $u_2 \sim U(0.5,1)$ for $1 \le j \ne k \le 7$, and $\eta_{jk} = 0$ for $j,k > 7$.

M1 contains both main and interaction effects, and M2 only contains interaction effects among $(g_1, g_2, \ldots, g_7)$. Additional simulation results under other model can be found in Web Appendix C. In each simulation, we generated two independent datasets $\mathcal{D}_1$ and $\mathcal{D}_2$, each with sample size n. Set $\mathcal{D}_1$ is used to evaluate SLR, and set $\{\mathcal{D}_1, \mathcal{D}_2\}$ is used to evaluate ESC. We choose α_l such that $|\mathcal{I}_{\mathrm{LR}}| = n$ for SLR screening, and use the significance level $\alpha = 0.05$ for all procedures. Let $\mathcal{M}_0$ be the index set of nonzero G×G coefficients of the true model and let $\mathcal{M}$ be the index set of nonzero G×G coefficients of the estimated model. For SLR screening, we report the numbers of *true positive* (TP), $E(|\mathcal{M} \cap \mathcal{M}_0|)$, and *false positive* (FP), $E(|\mathcal{M} \cap \mathcal{M}_0^c|)$. For ESC, we report the *true positive rate* (TPR), $E(|\mathcal{M} \cap \mathcal{M}_0|/|\mathcal{M}_0|)$, and the *false positive rate* (FPR), $E(|\mathcal{M} \cap \mathcal{M}_0^c|/|\mathcal{M}_0^c|)$. Simulation results are reported based on 100 replications.

Because SLR contains two steps for variable selection (i.e., low-rank screening and Lasso), we consider three other strategies as benchmarks as illustrated in Figure 1: (1) Lasso (i.e., skipping the low-rank screening), (2) L-Lasso (i.e., replacing low-rank screening by Lasso), which is also known as relaxed Lasso (Meinshausen, 2007), and (3) M-Lasso (i.e., replacing low-rank screening by MS, where MS is based on the PLINK procedure with modifications to accommodate the continuous traits). Following this naming system, SLR is in essence "Low-Rank Lasso", although we stay with the term "SLR". Finally, to compare with ESC, Lasso, L-Lasso, and M-Lasso are also combined with an independent cleaning stage (denoted by SC, LSC, and MSC, respectively; see also Figure 2 for details).

5.2. *Simulations with m_p in the Size of* 10^6

Chromosome 11 (with 1465 SNPs) is used to generate $\boldsymbol{g}$, which gives about 10^6 pairs of candidate G×G for each of $n = 300$ subjects. The number of SNPs entering the screening stage ranges from 55 to 70, suggesting (by Remark 2) that a low-rank model can be properly fitted (using $n = 300$) with $r \le 2$. We thus implement SLR with $r = 1, 2$, which is denoted by SLR(r). We also implement SLR with r being selected by cross-validation, which is denoted by SLR(cv). The corresponding ESC procedures are denoted by ESC(r) and ESC(cv). Simulation results are placed in Figure 3.

Under M1 (the left panel of Figure 3), we see while SLR screening yields similar TPs with Lasso, the FPs of SLR are largely reduced. Recall the only difference between ESC and SC is the method used in the screening stage. Comparing to SC, the variable set of ESC entered into the cleaning stage had fewer variables and a higher signal-to-noise ratio; consequently, ESC had higher TPRs than SC while the FPRs were adequately controlled (subplots (c)–(d)). The same phenomenon is observed when comparing SLR with the two-staged screening method L-Lasso. Although M-Lasso also has two-staged nature in screening, it has the lowest TPs (and larger FPs than SLR); consequently MSC has a poor performance. This result also reveals the drawback of MS which ignores the joint effects in screening. The gain of SLR is more obvious under M2 (the right panel of Figure 3), which contains a substantial amount of interaction effects. Under M2, SLR screening, which uses a low-rank model, can better identify significant interactions in $\boldsymbol{\eta}$ (subplot (a)). In contrast, conventional methods (Lasso, L-Lasso, and M-Lasso), which do not utilize the low-rank structure of $\boldsymbol{\eta}$, tend to incorrectly filter out significant interactions (e.g., fewer TPs than SLR in subplot (a)) and retain many insignificant terms (e.g., more FPs than SLR in subplot (b)). Therefore, the subsequent LSEs were obtained from a model that was further deviated away from the correct model, and had smaller TPRs than ESC (subplot (c)). Generally, ESC has the best performance, followed by LSC, SC, and MSC.

Note that the rank of $\boldsymbol{\eta}$ in models M1–M2 ranges from 5 to 7, and using SLR with $r \le 2$ is sufficient to achieve good performance. This result indicates the robustness and applicability of low-rank model (6), even with an under-specified rank, r. Moreover, we observed that ESC(cv) and ESC(1) have comparable performance, and ESC(1) outperforms ESC(2) in most of the settings. The observations suggest that the key component for the performance gain of SLR and ESC is the efficiency of the low-rank model fitting (as opposed to the precision in model approximation). Because the goal of the low-rank screening is to reduce the model size so as to stabilize the subsequent Lasso, an approximation of $\boldsymbol{\eta}$, e.g., the rank-1 model, was able to remove nonimportant terms. In contrast, although the rank-2 model approximates $\boldsymbol{\eta}$ more precisely, it also requires more parameters in the model fitting. With limited sample size, the improvement in approximation accuracy cannot compensate for the loss in estimation efficiency, and thus ESC(2) did not have a better performance than ESC(1)

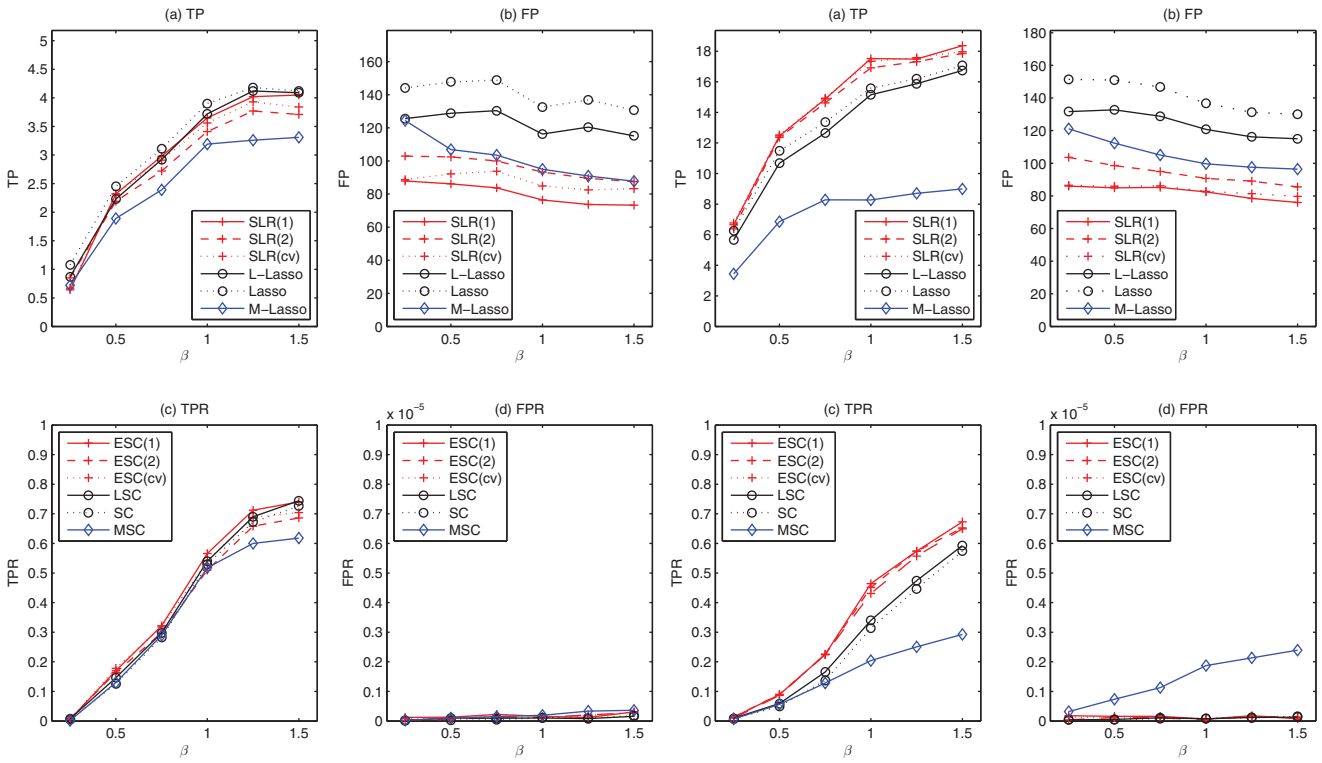

Figure 3. Simulation results with $(n, m_p) = (300, 10^6)$ under model M1 (left panel; the true model size is $|\mathcal{M}_0| = 5$) and model M2 (right panel; the true model size is $|\mathcal{M}_0| = 21$). (a)–(b): results of SLR screening using $\mathcal{D}_1$; (c)–(d): results of ESC using $\{\mathcal{D}_1, \mathcal{D}_2\}$.

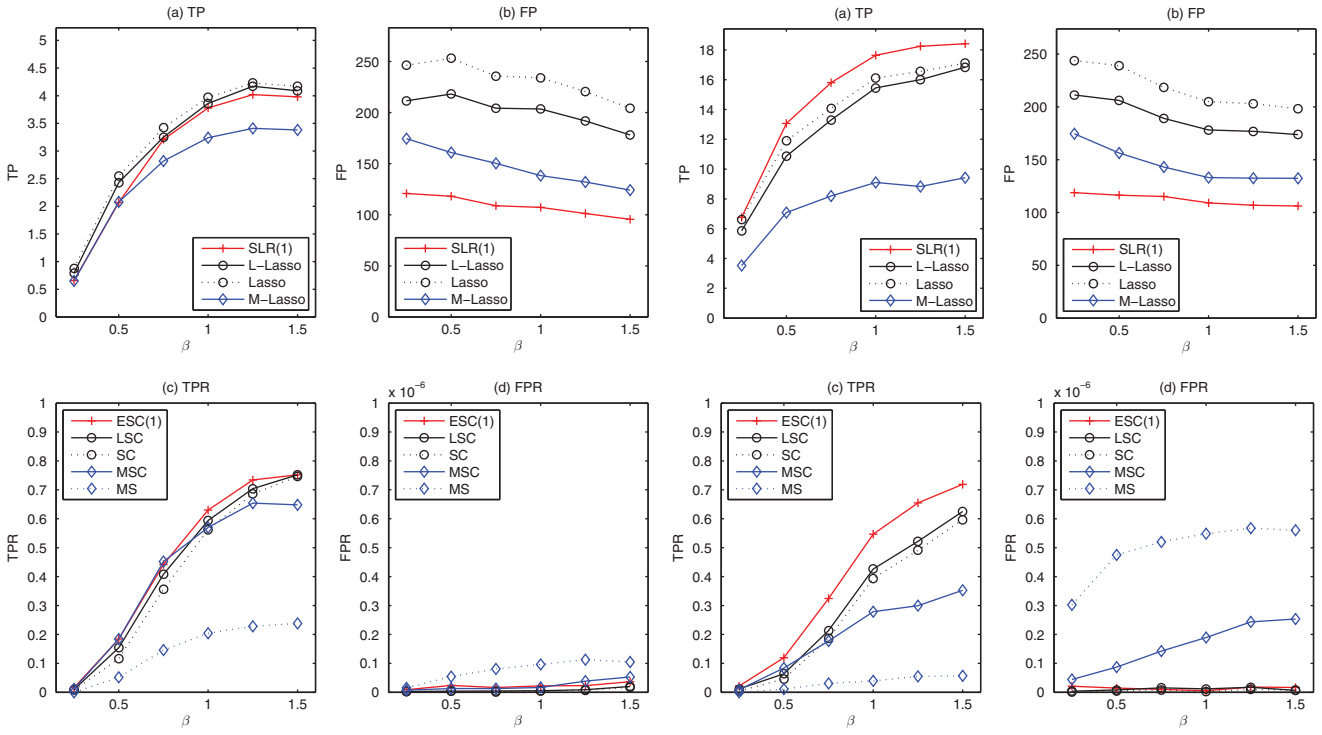

Figure 4. Simulation results with $(n, m_p) = (400, 10^7)$ under model M1 (left panel; the true model size is $|\mathcal{M}_0| = 5$) and model M2 (right panel; the true model size is $|\mathcal{M}_0| = 21$). (a)–(b): results of SLR screening using $\mathcal{D}_1$; (c)–(d): results of ESC using $\{\mathcal{D}_1, \mathcal{D}_2\}$.

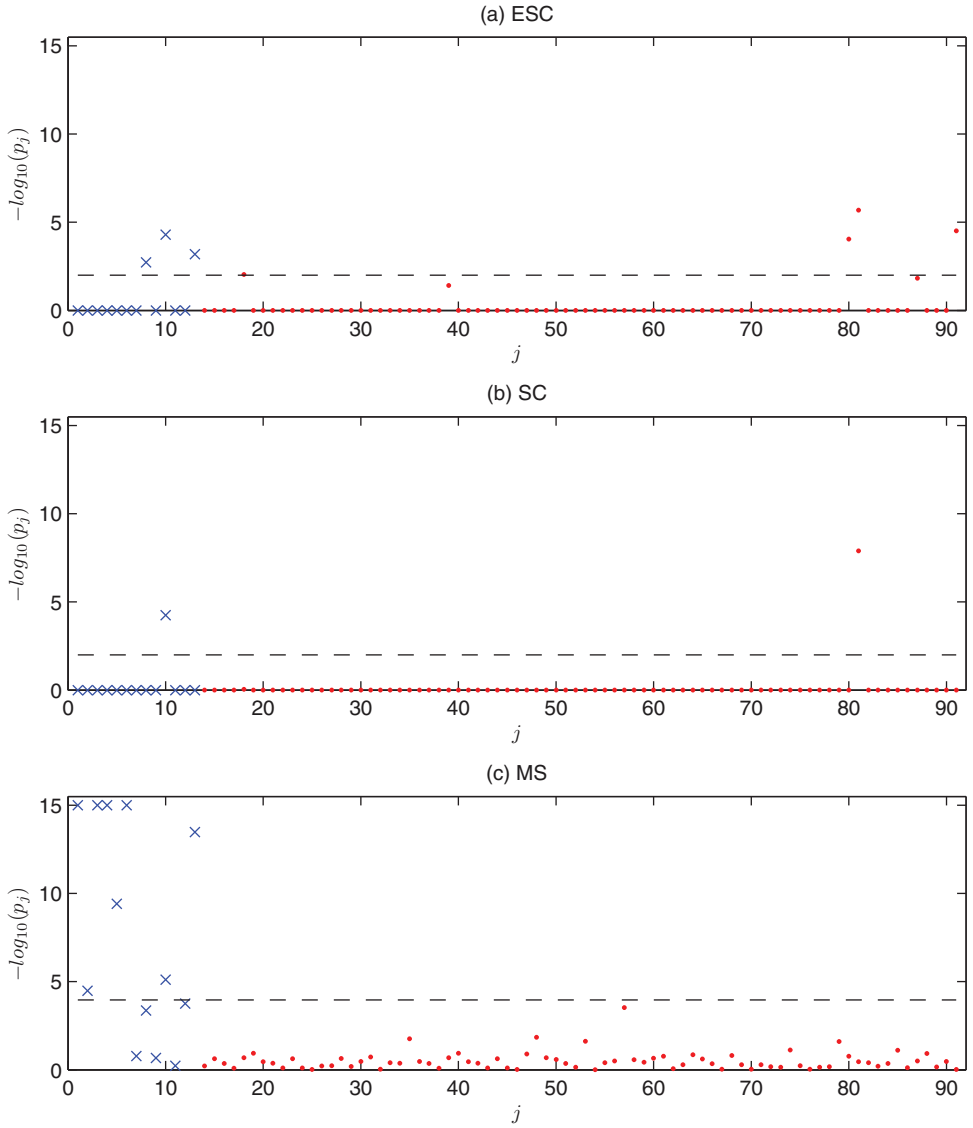

Figure 5. The p-values (in the scale of $-\log_{10}$) of the variables in the Warfarin data from ESC, SC, and MS. The x-axis represents the variable number, j, where $j = 1, \ldots, 13$ are for main effects (the $\times$), and $j = 14, \ldots, 91$ are for interactions (the $\bullet$). The horizontal dash line indicates the critical value under 0.01 family-wise error rate. The identified variables by each approach are listed in Web Appendix D.

when the sample size is not adequate. See Remark 2 for a discussion of selecting r.

5.3. *Simulations with m_p in the Size of* 10^7

In this simulation, chromosomes 10–12 (with 4666 SNPs) are used to generate $\boldsymbol{g}$, which gives about 10^7 pairs of candidate G×G. We consider the sample size to be $n = 400$. Since the available sample size is small in comparison with the number of SNPs, we only implement SLR with $r = 1$. For comparison, we also conduct MS (using the whole data $\{\mathcal{D}_1, \mathcal{D}_2\}$ directly). Simulation results are placed in Figure 4.

Comparing SLR with other screening methods (subplots (a)–(b)), we observed a similar pattern of TPs and FPs as in Figure 3 under both M1 and M2. That is, comparing to SLR, screening methods without using the low-rank nature of $\boldsymbol{\eta}$ identified a similar number of TPs but at the price of incorporating many additional FPs. The number of FPs of Lasso and L-Lasso was often twice of that of SLR in the high-dimensional setting. Consequently, it is not surprising to see that ESC outperformed SC and LSC (subplots (c)–(d)). Comparing to the case of $(n, m_p) = (300, 10^6)$, ESC exhibited comparable or better performance, which suggested the potential scalability of the proposed methods.

MS yielded lowest TPRs and highest FPRs among all methods. Unlike the joint screening-based ESC, MS examined variables one at a time and ignored the joint effects of variables. In addition, the multiple testing correction conducted in MS also led to severe loss in detecting power. Specifically, the total number of tests to be adjusted in ESC is the number of positive findings obtained from the $\mathcal{D}_1$ analysis, while the total

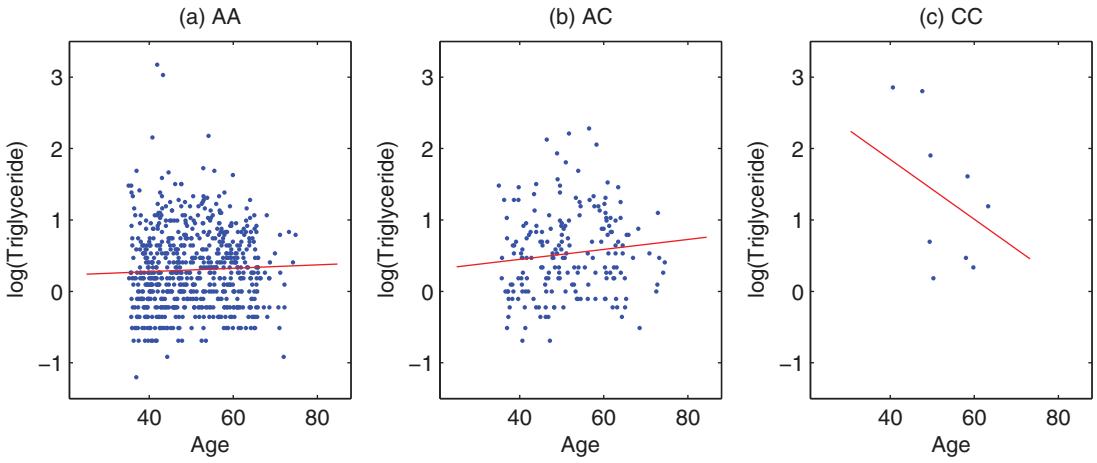

Figure 6. The scatter plots of log-triglyceride and age within different categories of rs6589567 in the CoLaus study. The line represents the fitted regression line that regresses log-triglyceride on age.

number of tests to be adjusted in MS is the number of all main effects and pairwise interactions from 4644 SNPs, i.e., about 10^7. As a result, MS cannot have better performance (i.e., low TPRs and high FPRs), even when the effect size β is moderate or large. In summary, the proposed ESC procedure improved upon SC, avoided the drawbacks of MS, and was able to perform satisfactorily in the case when m_p was large in size.

6. Data Analysis

6.1. *Application to the Warfarin Study (Moderate m_p)*

Warfarin is one of the most widely used oral anticoagulant agents, and the appropriate dose of Warfarin varies substantially among individuals. There is great interest in identifying the genetic and nongenetic factors that contribute to the variation in dosages, and in using these factors to determine appropriate dosages. The Warfarin data were compiled by the International Warfarin Pharmacogenetics Consortium (2009), and can be applied from the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB, `www.pharmgkb.org`). We focused on a homogeneous subset of the data comprising 265 Caucasian individuals who had complete information on two candidate genes (*VKORC1* and *CYP2C9*) and basic demographic variables. There were 13 variables in the analysis: variables 1–7 are from *VKORC1*, variables 8–9 are from *CYP2C9*, and variables 10–13 are Age, Sex, Height, and Weight.

Figure 5 shows the p-values for each variable obtained using ESC (with $r = 1$ being determined by fivefold cross-validation and $|\mathcal{I}_{\mathrm{LR}}| = n/3$), SC, and MS. The p-values of ESC and SC are obtained as discussed in Section 4.2; the p-values of MS are obtained as described in Section 1. A family-wise error rate of 0.01 was used to identify important variables (see Web Appendix D for detailed results). Comparing ESC to SC, both methods identified the main effect of Age and an interaction effect between CYP2C9*1 and Weight. In addition, ESC also identified interaction effects among CYP2C9*1, Height, and Weight, an interaction between VKORC1*1 and VKORC1*6 (i.e., two SNPs within VKORC1), and two main-effect terms (CYP2C9*1 and Weight). In contrast to the findings of ESC and SC, MS identified no interaction terms but eight main effect terms, among which Age and Weight overlap with the findings of ESC. While the influence of VKORC1 and CYP2C9 variations on warfarin dosages has been well established in the literature (e.g., the International Warfarin Pharmacogenetics Consortium, 2009), SC missed VKORC1 effects, and MS missed CYP2C9 effects. In contrast, ESC identified effects from both genes. Though typical Warfarin dosage research focuses on main effects, our analysis suggests that interaction effects may play a role in Warfarin dosage.

6.2. *Application to the CoLaus Study (Large m_p)*

Triglyceride concentration is an important risk factor for cardiovascular diseases. Understanding how the genetic variants modulate triglyceride would facilitate the development of therapeutic interventions to cardiovascular disease risk. In this data application, we used data from the Cohorte Lausannoise (CoLaus) study (Firmann et al., 2008). In our analysis, we focused on the subsamples of 883 male subjects with available information on ages, triglycerides, and GWAS data (genotyped with the Affymetrix 500 K SNP chip). As did in Wu et al. (2008), we restricted the interaction analysis on those SNPs with "promising" univariate effects on log-triglycerides. Specifically, we performed a single-SNP screening on the GWAS SNPs with minor allele frequencies (MAF) >0.01 and selected 1780 SNPs with p-values < 0.005 after adjusting for age and population substructures. We performed an interaction analysis using ESC, SC, and MS to identify important predictors among the main age effect, the main SNPs effects, and the pairwise interactions among age and SNPs.

Under 0.05 family-wise error rate, ESC identified a significant interaction between rs6589567 and age; SC did not identify any significant main or interaction effects; MS identified a significant main effect from rs6589567. Figure 6 shows the relationships between age and log-triglyceride by different genotype groups of rs6589567. While the triglyceride levels typically increase with age (Miller et al., 2011), we observed

a decreasing trend for the CC genotype group, suggesting a potential rs6589567×age effect as identified by ESC.

7. Discussion

The proposed method can be extended. First, in this article, we only consider the case of a normal error model with continuous response. It is important to extend our ESC procedure to the case of GLM. Although the idea is straightforward, the extension is not trivial due to the over-parameterization when modeling G×G as $\boldsymbol{\eta} = \boldsymbol{AUA}^T$, which further complicates the derivation of the asymptotic properties. Second, it is of interest to extend our method to a more complicated situation to approximate biological interaction. For example, consider $g_j \in \{\text{aa}, \text{Aa}, \text{AA}\}$ that represents a categorical random variable (instead of the numeric 0/1/2), which gives 3×3 possible interactions. To fit our model, we can encode g_j as $g_j^* = (g_{j1}, g_{j2})^T$, where g_{jk}'s are binary random variables: $g_j^* = (0,0)^T$ if $g_j = \{\text{aa}\}$, $g_j^* = (1,0)^T$ if $g_j = \{\text{Aa}\}$, and $g_j^* = (0,1)^T$ if $g_j = \{\text{AA}\}$. Let $\boldsymbol{g} = (g_1^{*T}, \ldots, g_p^{*T})^T$ be a $(2p)$-vector. The similar inference procedure for the low-rank model (6) developed in this article can be applied by using the newly defined $\boldsymbol{g}$, where the definition of vecp$(\cdot)$ and the selection criterion of ESC need some modifications accordingly. Finally, the efficiency gain of our method comes from treating G×G as a matrix, $\boldsymbol{\eta}$, and imposing a low-rank constraint on it. In view of low-rank, we can alternatively use the trace norm penalty $\|\boldsymbol{\eta}\|_*$ on $\boldsymbol{\eta}$ to identify G×G. Trace norm is a convex surrogate of the low-rank constraint and has deserved many advantages. It is of interest to investigate these extensions in future studies.

8. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2, 5, and 6, and a Matlab code to implementing the low-rank model fitting are available with this paper at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors thank the International Warfarin Pharmacogenetics Consortium and the PharmGKB resources for supplying the Warfarin data, thank Drs. Peter Vollenweider and Gerard Waeber, PIs of the CoLaus study, and Drs. Meg Ehm and Matthew Nelson, collaborators at GlaxoSmithKline, for providing the CoLaus phenotype and genetic data, and thank Dr. Shannon Holloway for input to improve the manuscript. This work was partially supported by National Science Council of Taiwan (to H.H. and S.Y.H.), by National Institutes of Health grant U01-HL-114494 (to P.C.), and by National Institutes of Health grants R01-MH-084022 and P01-CA-142538 (to J.Y.T.).

References

Cordell, H. J. (2002). Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics* **11**, 2463–2468.

Cordell, H. J. (2009). Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics* **10**, 392–404.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B* **70**, 849–911.

Firmann, M., Mayor, V., Vidal, P. M., Bochud, M., Pécoud, A., Hayoz, D., et al. (2008). The CoLaus study: A population-based study to investigate the epidemiology and genetic determinants of cardiovascular risk factors and metabolic syndrome. *BMC Cardiovascular Disorders* **8**, 6.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics* **6**, 95–108.

International Warfarin Pharmacogenetics Consortium, Klein, T.E., Altman, R. B., Eriksson, N., Gage, B. F., Kimmel, S. E., et al. (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. *The New England Journal of Medicine* **360**, 753–764.

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability approach to regularization selection (StARS) for high dimensional graphical models. *Neural Information Processing Systems*, 23.

Magnus, J. R. and Neudecker, H. (1979). The commutation matrix: Some properties and applications. *Annals of Statistics* **7**, 381–394.

Meinshausen, N. (2009). Relaxed Lasso. *Computational Statistics & Data Analysis* **52**, 374–393.

Meinshausen, N., Meier L., and Bühlmann, P. (2009). p-values for high-dimensional regression. *Journal of the American Statistical Association* **104**, 1671–1681.

Miller, M., Stone, N. J., Ballantyne, C., Bittner, V., Criqui, M. H., Ginsberg, H. N., et al. (2011). Triglycerides and cardiovascular disease: A scientific statement from the American Heart Association. *Circulation* **123**, 2292–2333.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559–575.

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N.L., et al. (2010). BOOST: A fast approach to detecting gene–gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics* **10**, 325–340.

Wasserman, L. and Roeder, K. (2009). High-dimensional variable selection. *Annals of Statistics* **37**, 5A, 2178–2201.

Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. (2010). Screen and clean: A tool for identifying interactions in genome-wide association studies. *Genetic Epidemiology* **34**, 275–285.

Received July 2014. *Revised June* 2015. *Accepted July* 2015.