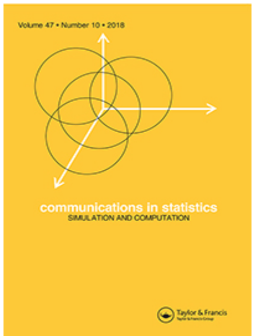
Volume 47 • Number 10 • 2018
communications in statistics
SIMULATION AND COMPUTATION
Taylor & Francis

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Group feature screening via the F statistic

Won Chul Song & Jun Xie

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2019.1691223

Taylor & Francis
Taylor & Francis Group

Check for updates

Group feature screening via the F statistic

Won Chul Song[a] and Jun Xie[b]

[a]Milwaukee School of Engineering, Milwaukee, WI, USA; [b]Department of Statistics, Purdue University, West Lafayette, IN, USA

ABSTRACT

Feature screening is crucial in the analysis of ultrahigh dimensional data, where the number of variables (features) is in an exponential order of the number of observations. In various ultrahigh dimensional data, variables are naturally grouped, giving us a good rationale to develop a screening method using joint effect of multiple variables. In this article, we propose a group screening procedure via the F-test statistic. The proposed method is a direct extension of the original sure independence screening procedure, when the group information is known, for example, from prior knowledge. Under certain regularity conditions, we prove that the proposed group screening procedure possesses the sure screening property that selects all effective groups with a probability approaching one at an exponential rate. We use simulations to demonstrate the advantages of the proposed method and show its application in a genome-wide association study. We conclude that the grouping method is very useful in the analysis of ultrahigh dimensional data, as the optimal F-test can detect true signals with desired properties.

ARTICLE HISTORY
Received 28 May 2018
Accepted 5 November 2019

KEYWORDS
Feature screening; Multiple regression; Sure screening property;
Ultrahigh dimension

1. Introduction

High dimensional data are commonly available in many areas of scientific research, where the number of variables, denoted as p, is higher than the number of samples, denoted as n. To examine relationships between these variables (covariates) and a response variable, a variety of variable selection methods have been developed, including Lasso (Tibshirani 1996), Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li 2001), elastic net (Zou and Hastie 2005), and the Dantzig selector (Candes and Tao 2007), etc. However, when the number of variables is much larger than the number of samples, these variable selection methods suffer in computational expediency, statistical accuracy, and algorithmic stability (Fan, Samworth, and Wu 2009). Fan and Lv (2008) introduced the term ultrahigh dimensional data, which refers to the situation of the number of variables being an exponential order of the number of samples, e.g., $p = O(\exp(n^a))$, for some constant $a > 0$. For the ultrahigh dimensional data, Fan and Lv (2008) recommended conducting a procedure called sure independence screening (SIS)

CONTACT Jun Xie junxie@purdue.edu Department of Statistics, Purdue University, 250 N. University Street, West Lafayette, IN 47907, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lssp.

Supplemental data for this article can be accessed here.

before variable selection. The SIS procedure ranks and selects variables by their Pearson's correlations with the response. It was proven that the SIS procedure had a sure screening property, that all influential covariates are selected with a probability going to 1 as n approaches to infinity. Since then, several feature screening procedures have been developed that are based on SIS. Fan and Song (2010) extended the SIS procedure to a generalized linear model using the maximum marginal likelihood estimator. Fan, Feng, and Song (2011) introduced nonparametric independence screening in sparse ultrahigh dimensional additive models. Zhao and Li (2012) proposed principled SIS for the Cox proportional hazards model. Li, Peng, Zhang, and Zhu (2012) suggested a robust rank correlation screening, which assorted covariates based on the Kendall's τ rank correlation coefficient. Li, Zhong, and Zhu (2012) extended the method to more general models using distance correlation learning. A limitation in the literature of all these feature screening methods is that they are based on the marginal effects of individual covariates. It is not uncommon that individual effects may be weak. Moreover, weak effects become extremely difficult to detect with the ultrahigh dimensional data, because the multiple testing would produce a great amount of false significances from the random chances. In these situations, the existing screening methods may not work. Motivated by group structures of real data, we propose a group screening method, named Group Screening, based on the standard multiple regression model. We aim to improve marginal methods by aggregating individual effects, as well as by reducing the burden of multiple testing.

???此外，弱效应在超高维数据中变得极难检测，因为多重测试将从随机通道中产生大量虚假意义

There are many examples of ultrahigh dimensional data in which covariates are grouped. Our motivating example is the genome-wide association study (GWAS), which contains up to millions of genetic variants, e.g., single nucleotide polymorphisms (SNPs). SNPs can be grouped into genes or other functional genetic segments based on known information, e.g., from a reference genome map, and these SNPs within a group would be highly correlated due to linkage disequilibrium. In a GWAS, we often screen millions of SNPs to examine whether a quantitative phenotype has a genetic background. We can improve the analysis by considering the joint effect of multiple SNPs in a gene group (Liu and Xie 2018). Our proposed Group Screening method will be based on the multiple regression model, as a direct extension of the original sure independence screening procedure, which is based on simple regression. Group Screening scans variables by groups and selects the covariate groups whose F-test statistics are greater than a threshold value, which is defined by a given false positive number. By employing F-tests to ultrahigh dimensional data with grouped structure, we will show that Group Screening possesses the sure screening property while controlling false positive rates.

This paper is organized as follows. In Sec. 2, we describe the Group Screening procedure. In Sec. 3, we establish the sure screening property for Group Screening. In Sec. 4, numerical simulations and a real data example are presented. In the applications of genome-wide association studies, Group Screening is a SNP-set analysis approach. We therefore compare it with several SNP-set methods, namely, the minimum p-value method (minP) (Chen et al. 2006), higher criticism (HC) (Wu et al. 2014), and the Sequence Kernel Association Test (SKAT) (Wu et al. 2011). Section 5 provides a discussion of some of our experiences and insights on screening procedures. The proofs of Theorem 1 and Lemmas, and additional simulations, are given in the Supplementary material.

2. Group screening procedure

Suppose the data consists of a random sample of a response and a set of ultrahigh dimensional covariates, represented in a matrix form $(\mathbf{Y}, \mathbf{X}_1, \ldots, \mathbf{X}_p)$, where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ denotes the responses of n data points and $\mathbf{X}_i = (X_{i1}, \ldots, X_{in})^T$ is the ith covariate in the sample, $1 \leq i \leq p$, with p up to an exponential order of n. The ultrahigh dimensional covariates $(\mathbf{X}_1, \ldots, \mathbf{X}_p)$ are divided into r groups based on known information. For each group $j = 1, \ldots, r$, let q_j be the number of covariates in the jth group, then $q_1 + \cdots + q_r = p$. For each group $j = 1, \ldots, r$, we consider a multiple regression model

$$\mathbf{Y} = \beta_{0j} 1 + \mathbf{X}_{n \times q_j} \boldsymbol{\beta}_{q_j} + \boldsymbol{\epsilon}_j$$

where 1 is a n-dimensional vector of 1, $\mathbf{X}_{n \times q_j}$ is a $n \times q_j$ covariate matrix for the jth group, β_{0j} is the intercept, $\boldsymbol{\beta}_{q_j}$ is the q_j-dimensional vector of coefficients, and $\boldsymbol{\epsilon}_j$ is a n-dimensional vector of errors that are iid normal with zero mean and variance σ_j^2. This multiple regression model allows correlated covariates within a group. Besides, we do not assume the groups are independent of each other. In other words, different groups may or may not be independent. Without a loss of generality, we consider q_j less than the sample size but at the same scale of n, e.g., $q_j < n - 1$ and $q_j = a_j n$ for some constant $0 < a_j < 1$. On the other hand, the total number of groups, r, increases exponentially with n, more specifically as $r = O(\exp(n^{1-2\kappa}))$, for some $\kappa < 1/2$.

Our proposed Group Screening studies a global hypothesis for each group $j = 1, \ldots, r$,

$$H_{0j} : \boldsymbol{\beta}_{q_j} = 0 \quad \text{versus} \quad H_{1j} : \boldsymbol{\beta}_{q_j} \neq 0.$$

For this hypothesis of group effect, the F-test is an optimal test, in the sense that it maximizes the minimum power when there is an overall regression effect (Arias-Castro, Candès, and Plan 2011). Let F_j denote the F statistic

$$F_j = \frac{\sum_{k=1}^{n} (\hat{Y}_k - \bar{Y})^2 / q_j}{\sum_{k=1}^{n} (Y_k - \hat{Y}_k)^2 / (n - q_j - 1)},$$

where $\hat{Y}_k$ is the predicted response based on the regression model, $k = 1, \ldots, n$. The Group Screening procedure consists of the following steps:

(1) Choose the number of false positives, denoted as m, and calculate a threshold value $\delta = \Phi^{-1}(1 - m/r)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function.

(2) Calculate the F-statistics for all groups, $F_1, \ldots, F_r$.

(3) Compute a score $\nu_j = \sqrt{q_j}(F_j - 1)/\tau_j$ for each group, $j = 1 \ldots r$, and select the groups whose scores are greater than δ, where $\tau_j = \sqrt{2\left(1 + \frac{q_j}{n - q_j - 1}\right)}$.

We define the screening threshold, $\delta = \Phi^{-1}(1 - m/r)$, from the inverse of the normal distribution function, where m corresponds to a pre-specified number of false positives. This threshold is derived from a normal approximation of the F statistic under the null hypothesis H_{0j}. We leave the specific derivation of the normal approximation in Lemma 1 in the next section. The number of false positives, m, is typically chosen as 1% to 5% of the total number of groups, being comparable to 1% to 5% of Type I error.

Our screening procedure is based on the F-tests of groups one by one. As a remark, we notice that if the true model consists of more than one group of covariates, then fitting a regression model using only one group of covariates would be a type of model misspecification. In other words, significance in the test of $H_{0j} : \boldsymbol{\beta}_{q_j} = 0$ versus $H_{1j} : \boldsymbol{\beta}_{q_j} \neq 0$ might not correspond to significance of the true model with more than one group of covariates. However, per discussion in Fan and Song (2010), testing groups one by one can be equivalent to testing against a joint regression of the true model with any number of nonzero $\boldsymbol{\beta}$'s. This is especially the case when we assume a fixed design matrix for the true model. Therefore, the validity of the F-test screening procedure is guaranteed.

3. Property of group screening

We denote $\mathcal{D}^c = \{j : \boldsymbol{\beta}_{q_j} = 0\}$ and $\mathcal{D} = \{1, \ldots, r\} - \mathcal{D}^c$, as the inactive and active groups of covariates, respectively. The Group Screening method selects groups, whose ν_j scores are greater than a threshold value, and produces a set of selected covariates, denoted as $\hat{\mathcal{D}} = \{j : \nu_j \geq \delta\}$. We first show that the screening score, as defined by $\nu_j = \sqrt{q_j}(F_j - 1)/\tau_j$, converges to the standard normal distribution under the null hypotheses and a regularity condition:

(C1) Consider a $n \times q$ design matrix $\mathbf{X}$ with n samples and q covariates. There exist constants $B > b > 0$ such that the eigenvalues of $\mathbf{X}^T\mathbf{X}/n$ are between b and B, and there exist constants $r > 4$ and $B' > 0$ such that $E|\epsilon_i|^r < B'$ and $E(\epsilon_i^2|\mathbf{X}) = \sigma^2 > 0$ for all $i = 1, \ldots, n$.

Lemma 1. *Assume that (C1) holds and that the number of covariates in a group is less than the sample size, i.e., $q_j = a_j n$, for a constant $0 < a_j < 1$. Under the null hypothesis $\boldsymbol{\beta}_{q_j} = 0$, the following equation holds for all $j = 1, \ldots, r$ and for any given x,*

$$P(\nu_j < x) = \Phi(x) + O(n^{-1/2}),$$

where $\nu_j = \sqrt{q}_j(F_j - 1)/\tau_j$ and $\tau_j = \sqrt{2\left(1 + \frac{q_j}{n - q_j - 1}\right)}$.

Condition (C1) is typical, requiring the design matrix to be bounded and the model error homoscedastic. The lemma is derived from an asymptotic expansion of the characteristic function of the transformed F statistic by Tonda and Fujikoshi (2004). The specific proof is provided in the Supplementary material. By Lemma 1, for any $j \in \mathcal{D}^c$, which corresponds to the null hypothesis, ν_j has an asymptotic standard normal distribution. Suppose the true model $\mathcal{D}$ has size $|\mathcal{D}| = s$ and then $|\mathcal{D}^c| = r - s$. We obtain the expected false positive rate as

$$E\left(\frac{|\hat{\mathcal{D}} \cap \mathcal{D}^c|}{|\mathcal{D}^c|}\right) = \frac{1}{r - s} \sum_{\{j \in \mathcal{D}^c\}} P(\nu_j \geq \delta).$$

When we set $\delta = \Phi^{-1}(1 - m/r)$, each probability on the right hand side of the equation is approximately m/r, so is the average over $(r - s)$ terms for $j \in \mathcal{D}^c$. Therefore, the false positive rate is controlled at m/r in the limit as $n \to \infty$. A simulation on the control of false positive rate through the normal approximation is reported in the Supplementary material Sec. 3.4.

For an active group $j \in \mathcal{D}$, there is a noncentral parameter, denoted as

$$\lambda_j = (X_{q_j}\boldsymbol{\beta}_j)^T (P_{F_j} - P_1)(X_{q_j}\boldsymbol{\beta}_j),$$

where $\boldsymbol{\beta}_j$ is the coefficient vector for the jth group ($\boldsymbol{\beta}_{q_j}$ in the previous notation), P_{F_j} is the projection matrix based on all covariates in the jth group, and P_1 is the projection matrix of the null model (all coefficients are zero except an intercept). We further assume the following condition:

(C2) There exist positive constants C_1, C_2 and $\kappa < 1/2$ such that

$$\min_{j\in\mathcal{D}} \lambda_j > C_1 n^{1-\kappa} \quad \text{and} \quad \max_{j\in\mathcal{D}} \lambda_j < C_2 n^{5/4-\kappa/2}.$$

Condition (C2) requires lower and upper bounds for the effect of active groups. The lower bound guarantees that the noncentral parameter is large enough, i.e., $\lambda_j/\sqrt{q_j} \to \infty$ for all $j \in \mathcal{D}$, which is the condition for the F-test to be asymptotically powerful (Arias-Castro, Candès, and Plan 2011). It is easy to show that this lower bound can be rewritten as $\min_{j\in\mathcal{D}} E(F_j - 1) > 2C_1' n^{-\kappa}$ for a positive constant C_1', which is analogous to Condition 3 of the sure independent screening (SIS) method (Fan and Lv 2008). The upper bound guarantees that the variance of the noncentral F statistic converges to zero when n increases, as demonstrated in the Supplementary material, which is a sufficient condition for the following sure screening property that Group Screening selects all active groups of covariates with a probability approaching to 1 at an exponential rate.

Theorem 1. *(Sure Screening Property) Assume Condition (C1) and (C2) and the threshold value $\delta = \Phi^{-1}(1 - m/r) < cn^{1/2-\kappa}$, for $0 < \kappa < 1/2$. Then, there exist constants c_1, c_2, c_3 such that*

$$P\left(\mathcal{D} \subseteq \hat{\mathcal{D}}\right) > 1 - O\left(s\left[\exp\left(-c_1 n^{1-2\kappa}\right) + \exp\left(-c_2 n^{3/4-(3/2)\kappa}\right) + \exp\left(-c_3 n^{1/2-\kappa}\right)\right]\right).$$

where s is the size of active groups $\mathcal{D}$ and m is the number of false positives.

The upper bound of the threshold value δ will be satisfied when r/m is an exponential order of n (details in the Supplementary material). The fact that we have a huge number of groups, i.e., $r = O(\exp(n^{1-2\kappa}))$, is a special case to guarantee that.

4. Application examples

The first simulation example compares Group Screening with the original SIS (Fan and Lv 2008). The second simulation example compares Group Screening with several other SNP-set methods, where the other SNP-set methods are considered as improvements to the single-SNP (marginal SNP effect) method in practice. At the end, we apply Group Screening on a real data example and show its performance through prediction errors.

Table 1. The means of true and false positive rates (TPR and FPR) out of 500 simulations for SIS and group screening.

Sparsity level = 1, m = 4					Sparsity level = 1, m = 6				
	Group	SIS	SIS10	SIS20		Group	SIS	SIS10	SIS20
FPR	0.03	0.33	0.00	0.00	FPR	0.04	0.44	0.01	0.00
TPR	1.00	0.99	0.74	0.31	TPR	1.00	0.99	0.84	0.46
Sparsity level = 3/4, m = 4					Sparsity level = 3/4, m = 6				
	Group	SIS	SIS10	SIS20		Group	SIS	SIS10	SIS20
FPR	0.03	0.33	0.00	0.00	FPR	0.04	0.44	0.01	0.00
TPR	1.00	1.00	0.87	0.59	TPR	1.00	1.00	0.93	0.70

4.1. Simulation study 1

We consider 250 groups of covariates where the number of covariates in each group is randomly generated from a uniform distribution, Uniform(10, 50). The total number of covariates is 7335. A group of covariates $\mathbf{X}_{q_j}$ is generated from a multivariate normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{kl}) = (0.5^{|k-l|})$ for $j = 1, \ldots, 250,$ where k and l are the row and column indexes of the matrix. We have considered two scenarios: 1) the groups are independent of each other; 2) the groups are correlated with a covariance matrix $\Sigma = (\sigma_{kl}) = (0.5^{|k-l|})$ for the entire list of covariates. The response variable is generated from the model $Y = \boldsymbol{\beta}_1 \mathbf{X}_{q_1} + \cdots + \boldsymbol{\beta}_6 \mathbf{X}_{q_6} + \epsilon,$ where the error ϵ is generated from the standard normal distribution and the number of true (active) group is $s = 6$. Some components of $\boldsymbol{\beta}$ may be zero, corresponding to different sparsity levels. By convention, q_j^{γ} number of components are assumed nonzero for $j = 1, \ldots, 6,$ where γ denotes the sparsity level. We consider two situations with $\gamma = 1$ or 3/4. The nonzero values of $\boldsymbol{\beta}$ are generated from Uniform(0,1). We randomly assign negative signs to half of the nonzero coefficients. The sample size is $n = 500$.

To compare with SIS, we need to modify the original SIS for group screening, since SIS was designed to select individual variables instead of groups. We consider three approaches: (1) SIS: An entire group is included in the selection $\hat{\mathcal{D}},$ as long as the original SIS selects one covariate in the group; (2) SIS10: An entire group is included in $\hat{\mathcal{D}},$ if the original SIS selects at least 10% of the covariates in the group; (3) SIS20: An entire group is included in $\hat{\mathcal{D}},$ if the original SIS selects at least 20% of the covariates in the group. The threshold of the SIS procedure is decided by permuting the response Y and using the $(1 - m/r)$ quantile of the permuted correlation coefficients of Y and individual covariates, where $m = 4$ or 6 is the number of false positives. For Group Screening, the threshold is $\delta = \Phi^{-1}(1 - m/r)$.

We repeat the simulation 500 times and report the average true positive rate (TPR) and false positive rate (FPR) of each method in Table 1 for the scenario of independent groups. More specifically, using the notations $\hat{\mathcal{D}}$ and $\mathcal{D}$ for the selected groups and the true active groups, we have TPR $= |\hat{\mathcal{D}} \cap \mathcal{D}|/|\mathcal{D}|$ and FPR $= |\hat{\mathcal{D}} \cap \mathcal{D}^c|/|\mathcal{D}^c|$. Because Group Screening is a group selection method, based on the global test from the F statistic, the TPR and FPR are calculated on the group level. For SIS and its modified versions, with higher percent of covariates required for the group selection,

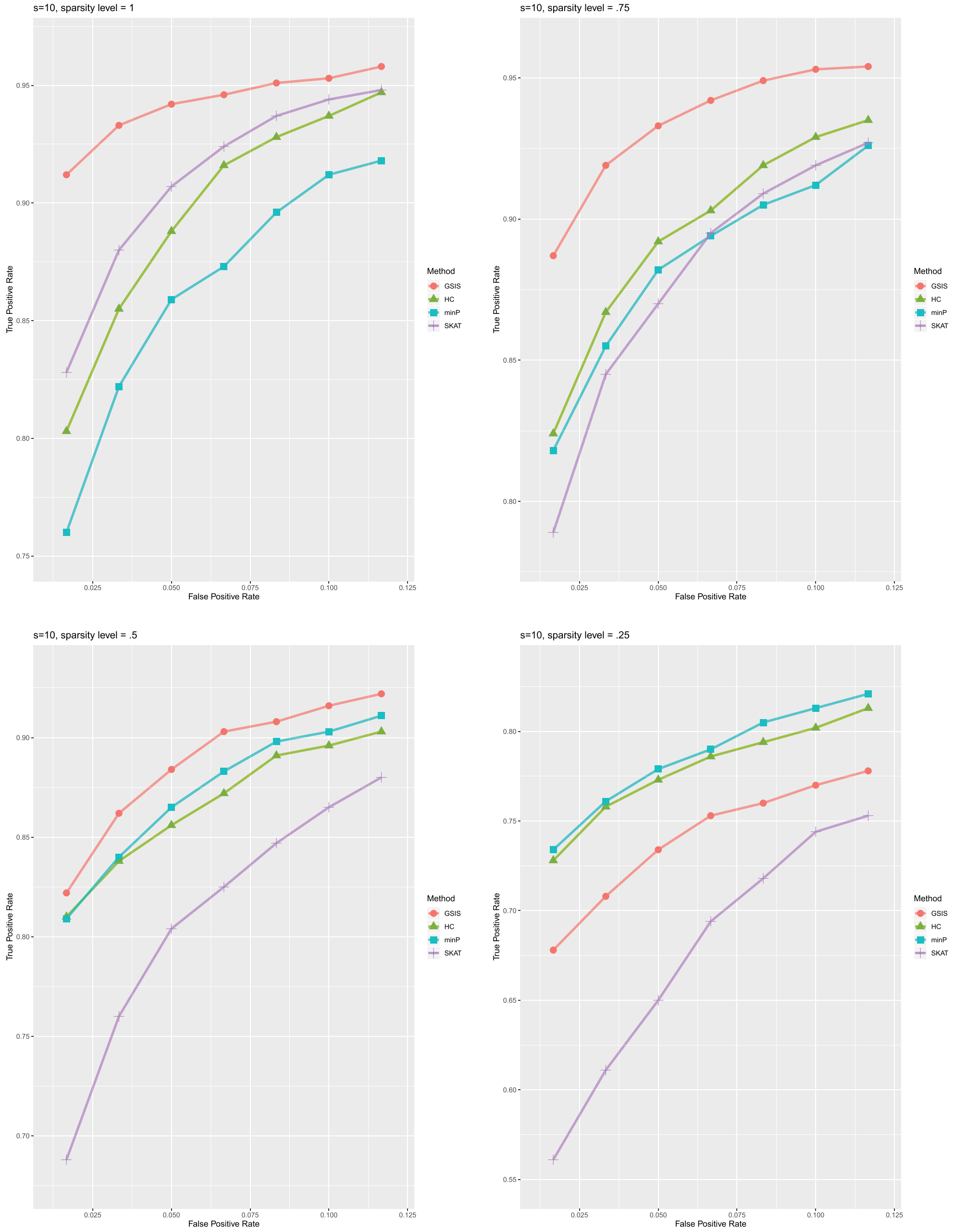

Figure 1. ROC curves of Group Screening and the other SNP set methods. The four plots show the results at different sparsity levels of 1, 0.75, 0.5, and 0.25. When the model is not very sparse, with sparsity levels of 0.5 or bigger, Group Screening is better than the other methods, giving a higher true positive rate at the given false positive level.

we observe lower FPR but also lower TPR. The proposed Group Screening method has a higher true positive rate than SIS10 and SIS20 and a comparable true positive rate with SIS. On the other hand, Group Screening has a lower false positive rate than SIS and a comparable false positive rate with SIS10 and SIS20. Overall, Group

Screening performs better than SIS and its modified approaches. We have also obtained similar results when conducting simulations with several different active group sizes s, or when the number of covariates in a group is from other distributions, e.g., Poisson, Gamma. In addition, the result for the scenario when groups are correlated is analogous with a similar conclusion. These simulation results are reported in the Supplementary material.

4.2. Simulation study 2

This simulation compares Group Screening with other SNP-set methods used in genome-wide association studies, including the minimum p-value method (minP) (Chen et al. 2006), Higher Criticism (HC) (Wu et al. 2014), and the Sequence Kernel Association Test (SKAT) (Wu et al. 2011). Each of these methods defines its specific test statistic from a group of SNPs (Liu and Xie 2018), while Group Screening is based on the classic F-test statistic. To simulate SNP genotypes, we consider discrete covariates, $X = 0$, 1, 2, denoting the genotype of a SNP variant with values of 0, 1, 2 corresponding to the number of copies of the minor allele. Assume $X \sim Binomial(2, 0.4)$ with the minor allele frequency 0.4. A R package *bindata* is used to generate correlated genotype data. The correlation matrices of SNPs have $\rho_{ij} = 0.75^{|i-j|}$, where i and j are the row and column indexes of the matrix. The total number of groups is again $r = 250$, and the sample size $n = 500$. The number of covariates for each group is randomly generated from Uniform(4, 36), resulting in a total number of covariates $p = 4901$ out of 250 groups. Similar to the previous simulation study, the response is generated from the regression model, $\mathbf{Y} = \boldsymbol{\beta}_1 \mathbf{X}_{q_1} + \cdots + \boldsymbol{\beta}_{10} \mathbf{X}_{q_{10}} + \epsilon$, with the number of true (active) groups $s = 10$. The coefficients, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{10}$ are from Uniform(0.5, 1) but at four different sparsity levels, $\gamma = 0.25, 0.5, 0.75, 1$. A very sparse model with $\gamma = 0.25$ would only have 1-2 covariates with nonzero coefficients. The error ϵ is generated from the standard normal distribution.

We compare Group Screening with the other methods via the ROC curve in Figure 1, plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The number of false positive groups are controlled at a variety of values, $m = 4, 8, 12, 16, 20, 24, 28$. We rank groups of covariates by p-values from the smallest to the largest for HC, minP, and SKAT, and by the score ν_j from the largest to the smallest for Group Screening. For each of the four methods, we count the number of true positives until we have m false positives. Then, the true positive rates are obtained at a given false positive rate. We repeat the simulation 100 times and use the mean values of true positive rates for the ROC curve. Figure 1 demonstrates that when the model is not very sparse, with $\gamma = 0.5$ or bigger, Group Screening is better than the other methods, giving a higher true positive rate at the given false positive level. On the other hand, minP and HC perform better than Group Screening when the model is very sparse. We have obtained similar performance results when conducting simulations with several different active group sizes other than $s = 10$, as well as a non-Gaussian error, e.g., a t-distribution with df $= 4$, in the regression model.

Table 2. Average prediction error and R^2 using the selected SNP sets from group screening and SIS in the dog GWAS example with 100 simulations from random splits of training and testing data.

	Prediction error	
	Group SIS	SIS
m = 1	0.0397 (0.007807)	0.0371 (0.007149)
m = 2	0.0396 (0.007829)	0.0401 (0.006063)
m = 3	0.0393 (0.008003)	0.0479 (0.006836)
	R^2	
	Group SIS	SIS
m = 1	0.794 (0.0456)	0.802 (0.0381)
m = 2	0.794 (0.0455)	0.796 (0.0338)
m = 3	0.794 (0.0468)	0.772 (0.0351)

The numbers in the parentheses are the sample standard deviation of the prediction error and R^2.

4.3. A real data example

We apply Group Screening to a genome-wide association study of dogs (Olsson et al. 2015), which aims to search for the most influential genes to Immunoglobulin A (IgA). We consider German Shepherd Dogs for the analysis. After a quality control, the final data contains 91,866 SNPs and 504 samples.

We use the dog genome assembly (CanFam3, http://genome.ucsc.edu/cgi-bin/hgGateway?db = canFam3) to group SNPs according to their locations in a gene or an intergenic region, resulting in a total of 1,292 gene groups in the data. The group size (number of SNPs) has a highly right-skewed distribution ranging between 1 and 1,275, and the first quantile, median, and third quantile of the group size are 2, 7, 76, respectively. There are only 28 gene groups with sizes larger than the sample size 504. We exclude them in the analysis. We apply Group Screening and the original SIS procedure on the data and consider three threshold settings, corresponding to different numbers of false positive groups, $m = 1, 2, 3$. To evaluate the screening methods, we randomly choose 80 percent of the sample to be training data and 20 percent of the sample to be testing data. We use the selected SNP sets from the training data to fit Lasso regression on the testing data and then calculate the prediction errors and R^2 values, where the prediction error is the mean squared error of the observed and fitted responses, $\sum_{k=1}^{n}(Y_k - \hat{Y}_k)^2/n$, with n as the sample size in the testing data. We repeat the process 100 times, randomly splitting 80% of the sample as training and 20% as testing data.

The selected SNP sets on the training data show that Group Screening and SIS mostly select different sets of SNPs, implying distinctive effects between the individual SNPs and the combination of multiple SNPs. Table 2 presents the prediction error and R^2 on the testing data using the SNPs selected from either Group Screening or SIS, and Lasso regression afterwards. The average and standard deviation of the prediction errors and R^2 in 100 simulations from random splits of training and testing data are reported. The prediction performances of Group Screening and SIS are comparable. Group Screening can give smaller prediction errors than SIS but the difference is not significant. On the other hand, Group Screening selects groups of SNPs based on joint effects of multiple SNPs and would be useful if we are interested in gene level effects instead of individual SNP level effects.

5. Discussion

Most screening procedures for ultrahigh dimensional data analysis, including SIS, are based on the marginal effects of individual variables. However, the issue of multiple testing and weak marginal effects may limit the application of the traditional statistical methods. This paper attempts to improve statistical analysis of ultrahigh dimensional data through a grouping method. The F-test is not only simple but also optimal under certain conditions for testing an overall group effect. The sure screening property established here further supports the use of the grouping method. The strength of the group effect, as indicated by the noncentral parameter in Condition (C2), is the most important factor for the sure screening property. Moreover, instead of considering a false discovery rate, as in the conventional setting of multiple testing, the sure screening property studies another aspect of the analysis method, i.e., the relationship between the selected group set $\hat{\mathcal{D}}$ and the true group set $\mathcal{D}$. Our Theorem 1 ensures the detection of true signals with a desired property.

The performance of Group Screening mainly depends on the model sparsity level. When the regression model is very sparse, with few nonzero coefficients, Group Screening would not be optimal. This is demonstrated in our simulations and also supported by the literature on optimality of the F-test (Liu and Xie 2018), which is asymptotically powerful for non-sparse models. On the other hand, as shown in the additional simulations in the Supplementary material, we observe robust performance of Group Screening when the groups are not independent, and even when the groups are misspecified, or the regression error is from a non-Gaussian distribution.

We consider Group Screening as the first step in ultrahigh dimensional data analysis, to narrow down to a subset of variables for further analysis. As described in the real data example, after the screening procedure, we would then run Lasso for variable selection and model fitting. Group Screening would be particularly useful in applications when we know variables are grouped and are interested in group effects, e.g., studying gene level effects rather than SNP level effects in a GWAS.

Acknowledgments

The authors thank the editor and the three anonymous referees for their comments and suggestions that have helped greatly improve the paper.

Funding

Research reported in this manuscript is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R21GM101504.

References

Arias-Castro, E., E. J. Candès, and Y. Plan. 2011. Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics* 39 (5):2533–56. doi:10.1214/11-AOS910.

Candes, E., and T. Tao. 2007. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* 35 (6):2313–51. doi:10.1214/009053606000001523.

Chen, B. E., L. C. Sakoda, A. W. Hsing, and P. S. Rosenberg. 2006. Resampling-based multiple hypothesis testing procedures for genetic case-control association studies. *Genetic Epidemiology* 30 (6):495–507. doi:10.1002/gepi.20162.

Fan, J., Y. Feng, and R. Song. 2011. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association* 106 (494):544–57. doi:10.1198/jasa.2011.tm09779.

Fan, J., and R. Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96 (456):1348–60. doi:10.1198/016214501753382273.

Fan, J., and J. Lv. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70 (5):849–911. doi:10.1111/j.1467-9868.2008.00674.x.

Fan, J., R. Samworth, and Y. Wu. 2009. Ultrahigh dimensional feature selection: Beyond the linear model. *Journal of Machine Learning Research* 10 (Sep):2013–38.

Fan, J., and R. Song. 2010. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics* 38 (6):3567–604. doi:10.1214/10-AOS798.

Li, G., H. Peng, J. Zhang, and L. Zhu. 2012. Robust rank correlation based screening. *The Annals of Statistics* 40 (3):1846–77. doi:10.1214/12-AOS1024.

Li, R., W. Zhong, and L. Zhu. 2012. Feature screening via distance correlation learning. *Journal of the American Statistical Association* 107 (499):1129–39. doi:10.1080/01621459.2012.695654.

Liu, Y., and J. Xie. 2018. Powerful test based on conditional effects for genome-wide screening. *The Annals of Applied Statistics* 12 (1):567–85. doi:10.1214/17-AOAS1103.

Olsson, M., K. Tengvall, M. Frankowiack, M. Kierczak, K. Bergvall, E. Axelsson, L. Tintle, E. Marti, P. Roosje, T. Leeb, et al. 2015. Genome-wide analyses suggest mechanisms involving early b-cell development in canine iga deficiency. *PLoS One* 10 (7):e0133844. doi:10.1371/journal.pone.0133844.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1):267–88. doi:10.1111/j.2517-6161.1996.tb02080.x.

Tonda, T., and Y. Fujikoshi. 2004. Asymptotic expansion of the null distribution of LR statistic for multivariate linear hypothesis when the dimension is large. *Communications in Statistics - Theory and Methods* 33 (5):1205–20. doi:10.1081/STA-120029835.

Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* 89 (1):82–93. doi:10.1016/j.ajhg.2011.05.029.

Wu, Z., Y. Sun, S. He, J. Cho, H. Zhao, and J. Jin. 2014. Detection boundary and higher criticism approach for rare and weak genetic effects. *The Annals of Applied Statistics* 8 (2):824–51. doi:10.1214/14-AOAS724.

Zhao, S. D., and Y. Li. 2012. Principled sure independence screening for cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis* 105 (1):397–411. doi:10.1016/j.jmva.2011.08.002.

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2):301–20. doi:10.1111/j.1467-9868.2005.00503.x.