Feature Screening for Ultrahigh Dimensional Categorical Data With Applications

Danyang Huang
Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University,
Beijing 100871, P.R. China (*dyhuang89@gmail.com*)

Runze Li
Department of Statistics and the Methodology Center, The Pennsylvania State University, University Park,
PA 16802 (*rzli@psu.edu*)

Hansheng Wang
Department of Business Statistics and Econometrics, Guanghua School of Management, Peking University,
Beijing 100871, P.R. China (*hansheng@gsm.pku.edu.cn*)

Ultrahigh dimensional data with both categorical responses and categorical covariates are frequently encountered in the analysis of big data, for which feature screening has become an indispensable statistical tool. We propose a Pearson chi-square based feature screening procedure for categorical response with ultrahigh dimensional categorical covariates. The proposed procedure can be directly applied for detection of important interaction effects. We further show that the proposed procedure possesses screening consistency property in the terminology of Fan and Lv (2008). We investigate the finite sample performance of the proposed procedure by Monte Carlo simulation studies and illustrate the proposed method by two empirical datasets.

KEY WORDS: Pearson's chi-square test; Screening consistency; Search engine marketing; Text classification.

1. INTRODUCTION

Since the seminal work of Fan and Lv (2008), feature screening for ultrahigh dimensional data has received considerable attention in the recent literature. Wang (2009) proposed the forward regression method for feature screening in ultrahigh dimensional linear models. Fan, Samworth, and Wu (2009) and Fan and Song (2010) developed sure independence screening (SIS) procedures for generalized linear models and robust linear models. Fan, Feng, and Song (2001) developed nonparametric SIS procedure for additive models. Li et al. (2012) developed a rank correlation based SIS procedure for linear models. Liu, Li, and Wu (2014) developed an SIS procedure for varying coefficient model based on conditional Pearson's correlation. Procedures aforementioned are all model-based methods. In the analysis of ultrahigh dimensional data, it would be very challenging in specifying a correct model in the initial stage. Thus, Zhu et al. (2011) advocated model-free procedures and proposed a sure independence and ranking screening procedure based on multi-index models. Li, Zhong, and Zhu (2012) proposed a model-free SIS procedure based on distance correlation (Szekely, Rizzo, and Bakirov 2007). He, Wang, and Hong (2013) proposed a quantile-adaptive model-free SIS for ultrahigh dimensional heterogeneous data. Mai and Zou (2013) proposed an SIS procedure for binary classification with ultrahigh dimensional covariates based on Kolmogorov's statistic. The aforementioned methods implicitly assume that predictor variables are continuous. Ultrahigh dimensional data with categorical predictors and categorical responses are frequently encountered in practice. This work aims to develop a new SIS-type procedure for this particular situation.

This work was partially motivated by an empirical analysis of data related to search engine marketing (SEM), which is also referred to as paid search advertising. It has been a standard practice to make textual advertisements on search engines such as Google in USA and Baidu in China. Keyword management plays a critical role in textual advertisements and therefore is of particular importance in SEM practice. Specifically, to maximize the amount of potential customers, the SEM practitioner typically maintains a large number of relevant keywords. Depending on the business scale, the total number of keywords ranges from thousands to millions. Practically managing so many keywords is a challenging task. For an easy management, the keywords need to be classified into fine groups. This is a requirement enforced by all major search engines (e.g., Google and Baidu). Ideally, the keywords belong to the same group should bear similar textual formulation and semantic meaning. This is a nontrivial task demanding tremendous efforts and expertise. The current industry practice largely relies on human forces, which is expensive and inaccurate. This is particularly true in China, which has the largest emerging SEM market in the world. Then, how to automatically classify Chinese keywords into prespecified groups becomes a problem of great importance. Such a problem indeed is how to handle high-dimensional categorical feature construction and how to identify important features.

Journal of Business & Economic Statistics
April 2014, Vol. 32, No. 2
DOI: 10.1080/07350015.2013.863158

From statistical point of view, we can formulate the problem as follows. We treat each keyword as a sample and index it by i with $1 \leq i \leq n$. Next, let $Y_i \in \{1, 2, \dots, K\}$ be the class label. We next convert the textual message contained in each keyword to a high-dimensional binary indicator. Specifically, we collect a set of most frequently used Chinese characters and index them by j with $1 \leq j \leq p$. Define a binary indicator X_{ij} as $X_{ij} = 1$ if the jth Chinese character appears in the ith keyword and $X_{ij} = 0$ otherwise. Collect all those binary indicators by a vector $X_i = (X_{i1}, \dots, X_{ip})^\top \in \mathbb{R}^p$. Because the total number of Chinese characters is huge, the dimension of X_i (i.e., p) is ultrahigh. Subsequently, the original problem about keyword management becomes an ultrahigh dimensional classification problem from X_i to Y_i. Many existing methods, including k-nearest neighbors (Hastie, Tibshirani, and Friedman 2001, kNN), random forest (Breiman 2001, RF), and support vector machine (Tong and Koller 2001; Kim, Howland, and Park 2005, SVM) can be used for high-dimensional binary classification. However, these methods become unstable if the problem is ultrahigh dimensional. As a result, feature screening becomes indispensable.

This article aims to develop a feature screening procedure for multiclass classification with ultrahigh dimensional categorical predictors. To this end, we propose using Pearson's chi-square (PC) test statistic to measure the dependence between categorical response and categorical predictors. We develop a screening procedure based on the Pearson chi-square test statistic. Since the Pearson chi-square test can be directly calculated using most statistical software packages. Thus, the proposed procedure can be easily implemented in practice. We further study the theoretical property of the proposed procedure. We rigorously prove that, with overwhelming probability, the proposed procedure can retain all important features, which implies the sure independence screening (SIS) property in the terminology of Fan and Lv (2008). In fact, under certain conditions, the proposed method can correctly identify the true model consistently. For convenience, the proposed procedure is referred to as PC-SIS, which possesses the following virtues.

The PC-SIS is a model-free screening procedure because the implementation of PC-SIS does not require one to specify a model for the response and predictors. This is an appealing property since it is challenging to specify a model in the initial stage of analyzing ultrahigh dimensional data. The PC-SIS can be directly applied for multicategorical response and multicategorical predictors. The PC-SIS has excellent capability in detecting important interaction effects by creating new categorical predictors for interactions between predictors. Furthermore, the PC-SIS is also applicable for multiple response and grouped or multivariate predictors by defining a new univariate categorical variable for the multiple response or the grouped predictors. Finally, by appropriate categorization, PC-SIS can handle the situation with both categorical and continuous predictors. In summary, the PC-SIS provides a unified approach for feature screening in ultrahigh dimensional categorical data analysis. We conduct Monte Carlo simulation to empirically verify our theoretical findings and illustrate the proposed methodology by two empirical datasets.

The rest of this article is organized as follows. Section 2 describes the detailed procedure of PC-SIS and establishes its theoretical property. Section 3 presents some numerical studies. Section 4 presents two real-world applications. The conclusion remark is given in Section 5. Technical proofs are given in the Appendix.

2. THE PEARSON CHI-SQUARE TEST BASED SCREENING PROCEDURE

2.1 Sure Independence Screening

Let $Y_i \in \{1, \dots, K\}$ be the corresponding class label, and $X_i = (X_{i1}, \dots, X_{ip})^\top \in \mathbb{R}^p$ be the associated categorical predictor. Since the predictors involved in our intended SEM application are binary, we assume thereafter that X_{ij} is binary. This allows us to slightly simplify our notation and technical proofs. However, the developed method and theory can be readily applied to general categorical predictors. Define a generic notation $\mathcal{S} = \{j_1, \dots, j_d\}$ to be a model with $X_{ij_1}, \dots, X_{ij_d}$ included as relevant features. Let $|\mathcal{S}| = d$ be the model size. Let $X_{i(\mathcal{S})} = (X_{ij} : j \in \mathcal{S}) \in \mathbb{R}^{|\mathcal{S}|}$ be the subvector of X_i according to $\mathcal{S}$. Define $\mathcal{D}(Y_i|X_{i(\mathcal{S})})$ to be the conditional distribution of Y_i given $X_{i(\mathcal{S})}$. Then a candidate model $\mathcal{S}$ is called sufficient, if

$$\mathcal{D}(Y_i|X_i) = \mathcal{D}\big(Y_i|X_{i(\mathcal{S})}\big). \tag{2.1}$$

Obviously, the full model $\mathcal{S}_F = \{1, 2, \dots, p\}$ is sufficient. Thus, we are only interested in the smallest sufficient model. Theoretically, we can consider the intersection of all sufficient models. If the intersection is still sufficient, it must be the smallest. We call it the true model and denote it by $\mathcal{S}_T$. Throughout the rest of this article, we assume $\mathcal{S}_T$ exists with $|\mathcal{S}_T| = d_0$.

The objective of feature screening is to find a model estimate $\widehat{\mathcal{S}}$ such that: (1) $\widehat{\mathcal{S}} \supset \mathcal{S}_T$; and (2) the size of $|\widehat{\mathcal{S}}|$ is as small as possible. To this end, we follow the marginal screening idea of Fan and Lv (2008) and propose the Pearson chi-square type statistic as follows. Define $P(Y_i = k) = \pi_{yk}$, $P(X_{ij} = k) = \pi_{jk}$, and $P(Y_i = k_1, X_{ij} = k_2) = \pi_{yj,k_1k_2}$. Those quantities can be estimated by $\hat{\pi}_{yk} = n^{-1}\sum I(Y_i = k)$, $\hat{\pi}_{jk} = n^{-1}\sum I(X_{ij} = k)$, and $\hat{\pi}_{yj,k_1k_2} = n^{-1}\sum I(Y_i = k_1)I(X_{ij} = k_2)$. Subsequently, a chi-square type statistic can be defined as

$$\widehat{\Delta}_j = \sum_{k_1=1}^{K}\sum_{k_2=1}^{2} \frac{\big(\hat{\pi}_{yk_1}\hat{\pi}_{jk_2} - \hat{\pi}_{yj,k_1k_2}\big)^2}{\hat{\pi}_{yk_1}\hat{\pi}_{jk_2}}, \tag{2.2}$$

which is a natural estimator of

$$\Delta_j = \sum_{k_1=1}^{K}\sum_{k_2=1}^{2} \frac{\big(\pi_{yk_1}\pi_{jk_2} - \pi_{yj,k_1k_2}\big)^2}{\pi_{yk_1}\pi_{jk_2}}. \tag{2.3}$$

Obviously, those predictors with larger $\widehat{\Delta}_j$ values are more likely to be relevant. As a result, we can estimate the true model by $\widehat{\mathcal{S}} = \{j : \widehat{\Delta}_j > c\}$, where $c > 0$ is some prespecified constant. For convenience, we refer to $\widehat{\mathcal{S}}$ as a PC-SIS estimator.

Remark 1. As one can see, $\widehat{\mathcal{S}}$ can be equivalently defined in terms of p-value. Specifically, define $\widehat{P}_j = P(\chi^2_{K-1} > n\hat{\Delta}_j)$, where χ^2_{K-1} stands for a chi-squared distribution with $K-1$ degrees of freedom. Because $\widehat{P}_j$ is a monotonically decreasing function in $\widehat{\Delta}_j$, $\widehat{\mathcal{S}}$ can be equivalently expressed as $\widehat{\mathcal{S}} = \{j : \widehat{P}_j < p_c\}$ for some constant $0 < p_c < 1$. For situations in which the number of categories involved by each predictor is different, the predictor involving more categories is likely to be associated with larger Δ_j values, regardless of whether the predictor is important or not. In such cases, directly using Δ_j for variable screening is less accurate. Instead, it is more appropriate to use p-value $\widehat{P}_j$ obtained from the Pearson chi-squared test of independence with degrees of freedom $(K-1)(R_j-1)$ for an R_j-level categorical predictor.

2.2 Theoretical Properties

We next investigate the theoretical properties of $\widehat{\mathcal{S}}$. Define $\omega_j^{k_1k_2} = \mathrm{cov}\{I(Y_i=k_1), I(X_{ij}=k_2)\}$. We then assume the following conditions.

- (C1) (*Response Probability*) Assume that there exist two positive constants $0<\pi_{\min}<\pi_{\max}<1$ such that $\pi_{\min}<\pi_{yk}<\pi_{\max}$ for every $1\le k\le K$ and $\pi_{\min}<\pi_{jk}<\pi_{\max}$ for every $1\le j\le p$ and $1\le k\le K$.
- (C2) (*Marginal Covariance*) Assume $\Delta_j=0$ for any $j\notin\mathcal{S}_T$. We further assume that there exists positive constant $\omega_{\min}$, such that $\min_{j\in\mathcal{S}_T}\max_{k_1k_2}(\omega_j^{k_1k_2})^2>\omega_{\min}$.
- (C3) (*Divergence Rate*) Assume $\log p\le\nu n^{\xi}$ for some constants $\nu>0$ and $0<\xi<1$.

Condition (C1) excludes those features with one particular category's response probability extremely small (i.e., $\pi_{yk}\approx 0$) or extremely large (i.e., $\pi_{yk}\approx 1$). Condition (C2) requires that, for every relevant categorical feature $j\in\mathcal{S}_T$, there exists at least one response category (i.e., k_1) and one feature category (i.e., k_2), which are marginally correlated (i.e., $\omega_j^{k_1k_2}>\omega_{\min}$). Under a linear regression setup, similar condition was also used by Fan and Lv (2008) but in terms of the marginal covariance. Condition (C2) also assumes that $\Delta_j=0$ for every $j\notin\mathcal{S}_T$. With the help of this condition, we can rigorously show that $\widehat{\mathcal{S}}$ is selection consistent for $\mathcal{S}_T$, that is $P(\widehat{\mathcal{S}}=\mathcal{S}_T)\to 1$ as $n\to\infty$ in Theorem 1. If this condition is removed, the conclusion becomes screening consistent (Fan and Lv 2008), that is $P(\widehat{\mathcal{S}}\supset\mathcal{S}_T)\to 1$ as $n\to\infty$. Finally, condition (C3) allows the feature dimension p to diverge at an exponentially fast speed in terms of the sample size n. Accordingly, the feature dimension could be much larger than sample size n. Then, we have the following theorem.

Theorem 1. (Strong Screening Consistency) Under Conditions (C1)–(C3), there exists a positive constant c such that $P(\widehat{\mathcal{S}}=\mathcal{S}_T)\to 1$.

2.3 Interaction Screening

Interaction detection is important for the intended SEM application. Consider two arbitrary feature X_{ij_1} and X_{ij_2}. We say they are free of interaction effect if conditioning on Y_i, they are independent with each other. Otherwise, we say they have nontrivial interaction effect. Theoretically, such an interaction effect can be conveniently measured by

$$\Omega_{j_1j_2}=\sum_k\sum_{k_1=1}^{2}\sum_{k_2=1}^{2}\frac{(\pi_{k,j_1,k_1}\pi_{k,j_2,k_2}-\pi_{k,j_1j_2,k_1k_2})^2}{\pi_{k,j_1,k_1}\pi_{k,j_2,k_2}},$$

where $\pi_{k,j_1j_2,k_1k_2}=P(X_{ij_1}=k_1,X_{ij_2}=k_2|Y_i=k)$ and $\pi_{k,j,k^*}=P(X_{ij}=k^*|Y_i=k)$. They can be estimated, respectively, by $\hat\pi_{k,j_1j_2,k_1k_2}=\{\sum I(Y_i=k)\}^{-1}\sum I(X_{ij_1}=k_1,X_{ij_2}=k_2,Y_i=k)$, and $\hat\pi_{k,j,k^*}=\{\sum I(Y_i=k)\}^{-1}\sum I(X_{ij}=k^*,Y_i=k)$. Subsequently, $\Omega_{j_1j_2}$ can be estimated by

$$\widehat{\Omega}_{j_1j_2}=\sum_k\sum_{k_1=1}^{2}\sum_{k_2=1}^{2}\frac{(\hat\pi_{k,j_1,k_1}\hat\pi_{k,j_2,k_2}-\hat\pi_{k,j_1j_2,k_1k_2})^2}{\hat\pi_{k,j_1,k_1}\hat\pi_{k,j_2,k_2}}.$$

Accordingly, those interaction terms with large $\widehat{\Omega}_{j_1j_2}$ values should be considered as promising ones. As a result, it is natural to select important interaction effects by $\widetilde{\mathcal{I}}=\{(j_1,j_2):\widehat{\Omega}_{j_1j_2}>c\}$ for some critical value $c>0$. It is remarkable that the critical value c used here is typically different from that of $\widehat{\mathcal{S}}$. As one can imagine, searching for important interaction effects over every possible feature pair is computationally expensive. To save computational cost, we suggest to focus on those features in $\widehat{\mathcal{S}}$. This leads to the following practical solution:

$$\widehat{\mathcal{I}}=\big\{(j_1,j_2):\widehat{\Omega}_{j_1j_2}>c\text{ and }j_1,j_2\in\widehat{\mathcal{S}}\big\}. \tag{2.4}$$

Under appropriate conditions, we can also show that $I(\widehat{\mathcal{I}}=\mathcal{I}_T)\to\infty$ as $n\to\infty$, where $\mathcal{I}_T=\{(j_1,j_2):\Omega_{j_1j_2}>0\}$.

2.4 Tuning Parameter Selection

We first consider tuning parameter selection for $\widehat{\mathcal{S}}$. To this end, various nonnegative values can be considered for c. This leads to a set of candidate models, which are collected by a solution path $\mathcal{F}=\{\mathcal{S}_j:1\le j\le p\}$, where $\mathcal{S}_j=\{k_1,\ldots,k_j\}$. Here $\{k_1,\ldots,k_p\}$ is a permutation of $\{1,\ldots,p\}$ such that $\hat\Delta_{k_1}\ge\hat\Delta_{k_2}\ge\cdots\ge\hat\Delta_{k_p}$. As a result, the original problem about tuning parameter selection for c is converted into a problem about model selection for $\mathcal{F}$. To solve the problem, we propose the following maximum ratio criterion. To illustrate the idea, we temporarily assume that $\mathcal{S}_T\in\mathcal{F}$. Recall that the true model size is $|\mathcal{S}_T|=d_0$. We then should have $\hat\Delta_{k_j}/\hat\Delta_{k_{j+1}}\to_p c_{jj+1}$ for some positive constant $c_{jj+1}>0$, as long as $j+1\le d_0$. One the other side, if $j>d_0$, we should have both $\hat\Delta_j$ and $\hat\Delta_{j+1}$ converges in probability toward 0. If their convergence rates are comparable, we should have $\hat\Delta_{k_j}/\hat\Delta_{k_{j+1}}=O_p(1)$. However, if $j=d_0$, we should have $\hat\Delta_j\to_p c_j$ for some positive constant $c_j>0$ but $\hat\Delta_{j+1}\to_p 0$. This makes the ratio $\hat\Delta_{k_j}/\hat\Delta_{k_{j+1}}\to_p\infty$. This suggests that d_0 can be estimated by

$$\hat d=\mathrm{argmax}_{0\le j\le p-1}\hat\Delta_{k_j}\big/\hat\Delta_{k_{j+1}},$$

where $\hat\Delta_0$ is defined to be $\hat\Delta_0=1$ for the sake of completeness. Accordingly, the final model estimate is given by $\widehat{\mathcal{S}}=\{j_1,j_2,\ldots,j_{\hat d}\}\in\mathcal{F}$. Similar idea also can be used to estimate the interaction model $\widehat{\mathcal{I}}$ and get the interaction model size $\hat d_{\mathcal{I}}$. Our numerical experiments suggest that it works fairly well.

3. SIMULATION STUDIES

3.1 Example 1: A Model Without Interaction

We first consider a simple example without any interaction effect. We generate $Y_i\in\{1,2,\ldots,K\}$ with $K=4$ and $P(Y_i=k)=1/K$ for every $1\le k\le K$. Define the true model to be $\mathcal{S}_T=\{1,2,\ldots,10\}$ with $|\mathcal{S}_T|=10$. Next, conditional on Y_i,

Table 1. Probability specification for Example 1

	j									
θ_{kj}	1	2	3	4	5	6	7	8	9	10
$k = 1$	0.2	0.8	0.7	0.2	0.2	0.9	0.1	0.1	0.7	0.7
$k = 2$	0.9	0.3	0.3	0.7	0.8	0.4	0.7	0.6	0.4	0.1
$k =3$	0.7	0.2	0.1	0.6	0.7	0.6	0.8	0.9	0.1	0.8
$k = 4$	0.1	0.9	0.6	0.1	0.3	0.1	0.4	0.3	0.6	0.4

we generate relevant features as $P(X_{ij} = 1|Y_i = k) = \theta_{kj}$ for every $1 \le k \le K$ and $j \in \mathcal{S}_T$. Their detailed values are given in Table 1. Then, for any $1 \le k \le K$ and $j \notin \mathcal{S}_T$, we define $\theta_{kj} = 0.5$. For a comprehensive evaluation, various feature dimensions ($p = 1000, 5000$) and sample sizes ($n = 200, 500, 1000$) are considered.

For each random replication, the proposed maximum ratio method is used to select both $\widehat{\mathcal{S}}$ and $\widehat{\mathcal{I}}$. Subsequently, the number of correctly identified main effects CME $= |\widehat{\mathcal{S}} \bigcap \mathcal{S}_T|$ and incorrectly identified main effects IME $= |\widehat{\mathcal{S}} \bigcap \mathcal{S}_T^c|$ with $\mathcal{S}_T^c = \mathcal{S}_F \backslash \mathcal{S}_T$ are computed. The interaction effects are similarly summarized. This leads to the number of correctly and incorrectly identified interaction effects, which are denoted by CIE and IIE, respectively. Moreover, the final model size, that is MS $= |\widehat{\mathcal{S}}| + |\widehat{\mathcal{I}}|$, is computed. The coverage percentage, defined by CP $= (|\widehat{\mathcal{S}} \bigcap \mathcal{S}_T| + |\widehat{\mathcal{I}} \bigcap \mathcal{I}_T|)/(|\mathcal{S}_T| + |\mathcal{I}_T|)$, is recorded. Finally, all those summarizing measures are averaged across the 200 simulation iterations and then reported in Table 2. They correspond to the rows with the screening method flagged by $\widehat{\mathcal{S}} + \widehat{\mathcal{I}}$. For comparison purpose, the full main effect model $\mathcal{S}_F$ (i.e., the model with all the main effect without interaction) and also the selected main effect model $\widehat{\mathcal{S}}$ (i.e., the model with all the main effect in $\widehat{\mathcal{S}}$ without interaction) are also included.

The detailed results are given in Table 2. For a given simulation model, a fixed feature dimension p, and a diverging sample size n, we find that the CME increases toward $|\mathcal{S}_T| = 10$ and IME decreases toward 0, and there is no overfitting effect. This result corroborates the theoretical result of Theorem 1 very well. In the meanwhile, since there is no interaction in this particular model, CIE is 0 and IIE converges toward 0 as n goes to infinity.

3.2 Example 2: A Model With Interaction

We next investigate an example with genuine interaction effects. Specifically, the class label is generated in the same way as the previous example with $K = 4$. Conditional on $Y_i = k$, we generate X_{ij} with $j \in \{1, 3, 5, 7\}$ according to probability $P(X_{ij} = 1|Y_i = k) = \theta_{kj}$, whose detailed values are given in Table 3. Conditional on Y_i and $X_{i,2m-1}$, we generate $X_{i,2m}$ according to

$$P(X_{i,2m} = 1|Y_i = k, X_{i,2m-1} = 0) = 0.05I(\theta_{k,2m-1} \ge 0.5) + 0.4I(\theta_{k,2m-1} < 0.5)$$

$$P(X_{i,2m} = 1|Y_i = k, X_{i,2m-1} = 1) = 0.95I(\theta_{k,2m-1} \ge 0.5) + 0.4I(\theta_{k,2m-1} < 0.5),$$

Table 2. Example 1 detailed simulation results

			Main effect		Interaction effect			
p	n	Method	CME	IME	CIE	IIE	MS	CP%
1000	200	$\mathcal{S}_F$	10.0	990.0	0.0	0.0	1000.0	100.0
		$\widehat{\mathcal{S}}$	9.8	0.0	0.0	0.0	9.9	98.6
		$\widehat{\mathcal{S}} + \widehat{\mathcal{I}}$	9.8	0.0	0.0	1.1	11.0	98.6
	500	$\mathcal{S}_F$	10.0	990.0	0.0	0.0	1000.0	100.0
		$\widehat{\mathcal{S}}$	10.0	0.0	0.0	0.0	10.0	100.0
		$\widehat{\mathcal{S}} + \widehat{\mathcal{I}}$	10.0	0.0	0.0	0.2	10.2	100.0
	1000	$\mathcal{S}_F$	10.0	990.0	0.0	0.0	1000.0	100.0
		$\widehat{\mathcal{S}}$	10.0	0.0	0.0	0.0	10.0	100.0
		$\widehat{\mathcal{S}} + \widehat{\mathcal{I}}$	10.0	0.0	0.0	0.0	10.0	100.0
5000	200	$\mathcal{S}_F$	10.0	4990.0	0.0	0.0	5000.0	100.0
		$\widehat{\mathcal{S}}$	9.6	0.0	0.0	0.0	9.6	96.6
		$\widehat{\mathcal{S}} + \widehat{\mathcal{I}}$	9.6	0.0	0.0	1.1	10.7	96.6
	500	$\mathcal{S}_F$	10.0	4990.0	0.0	0.0	5000.0	100.0
		$\widehat{\mathcal{S}}$	10.0	0.0	0.0	0.0	10.0	100.0
		$\widehat{\mathcal{S}} + \widehat{\mathcal{I}}$	10.0	0.0	0.0	0.8	10.8	100.0
	1000	$\mathcal{S}_F$	10.0	4990.0	0.0	0.0	5000.0	100.0
		$\widehat{\mathcal{S}}$	10.0	0.0	0.0	0.0	10.0	100.0
		$\widehat{\mathcal{S}} + \widehat{\mathcal{I}}$	10.0	0.0	0.0	0.0	10.0	100.0

Table 3. Probability specification for Example 2

	j			
θ_{kj}	1	3	5	7
$k = 1$	0.8	0.8	0.7	0.9
$k = 2$	0.1	0.3	0.2	0.3
$k = 3$	0.7	0.9	0.1	0.1
$k = 4$	0.2	0.1	0.9	0.7

for every $1 \le k \le K$ and $m \in \{1, 2, 3, 4\}$. Finally, we define $\theta_{kj} = 0.4$ for any $1 \le k \le K$ and $j > 8$. Accordingly, we should have $\mathcal{S}_T = \{1, 2, \ldots, 8\}$ and $\mathcal{I}_T = \{(1, 2), (3, 4), (5, 6), (7, 8)\}$. The detailed results are given in Table 4. The basic findings are qualitatively similar to those in Table 2. The only difference is that the CIE value no longer converges toward 0. Instead, it converges toward $|\mathcal{I}_T| = 4$ as $n \to \infty$ and p fixed. Also, CP values for $\widehat{\mathcal{S}}$ are no longer near 100% since $\widehat{\mathcal{S}}$ only takes main effect into consideration. Instead, the CP value for $\widehat{\mathcal{S}} + \widehat{\mathcal{I}}$ converges toward 100% as n increases and p is fixed.

3.3 Example 3: A Model With Both Categorical and Continuous Variables

We consider here an example with both categorical and continuous variables. Fix $|\mathcal{S}_T| = 20$. Here, $Y_i \in \{1, 2\}$ is generated according to $P(Y_i = 1) = P(Y_i = 2) = 1/2$. Given $Y_i = k$, we generate latent variable $Z_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ip})^\top \in \mathbb{R}^p$ with Z_{ij} independently distributed as $N(\mu_{kj}, 1)$, where $\mu_{kj} = 0$ for any $j > d_0$, $\mu_{kj} = -0.5$ if $Y_i = 1$ and $j \le d_0$, and $\mu_{kj} = 0.5$ if $Y_i = 2$ and $j \le d_0$. Finally, we construct observed feature X_{ij} as follows. If j is an odd number, we then define $X_{ij} = Z_{ij}$. Otherwise, define $X_{ij} = I(Z_{ij} > 0)$. As a result, this example involves a total of $d_0 = 20$ features are relevant. Half of them are continuous and half of them are categorical. To apply our method, we need to first discretize the continuous variables to be categorical. Specifically, let z_α stands for the αth quantile of a standard normal distribution. We then redefine those continuous predictors as $X_{ij} = 1$ if $X_{ij} < z_{0.25}$, $X_{ij} = 2$ if $z_{0.25} < X_{ij} < z_{0.50}$, $X_{ij} = 3$ if $z_{0.50} < X_{ij} < z_{0.75}$, and $X_{ij} = 4$ if $X_{ij} > z_{0.75}$. By doing so all the features become categorical. We next apply our method to the converted datasets by using p-values as described in the Remark 1. The experiment is replicated in a similar manner as before with detailed results summarized in Table 5. The results are qualitatively similar to those in Example 1.

4. REAL DATA ANALYSIS

4.1 A Chinese Keyword Dataset

The data contain a total of 639 keywords (i.e., samples), which are classified into $K = 13$ categories. The total number of Chinese characters involved is $p = 341$. For each class, we randomly split the sample into two parts with equal sizes. One part is used for training and the other for testing. The sample size of the training data is $n = 320$. Based on the training data, models are selected by the proposed PC-SIS method and various classification methods (i.e., kNN, SVM, and RF) are applied. Their forecasting accuracies are examined on the testing data. For a reliable evaluation, such an experiment is randomly replicated 200 times. The detailed results are given in Table 6. As seen, the PC-SIS estimated main effect model $\widehat{\mathcal{S}}$, with size 14.6 on average, consistently outperforms the full model $\mathcal{S}_F$, regardless of the classification method. The relative improvement margin could be as high as $87.2\% - 51.1\% = 36.1\%$ for SVM. Such an outstanding performance can be further improved by including about 22.3 interaction effects. The maximum improvement margin is $78.0\% - 67.6\% = 10.4\%$ for RF.

Table 4. Example 2 detailed simulation results

			Main effect		Interaction effect			
p	n	Method	CME	IME	CIE	IIE	MS	CP%
1000	200	$\mathcal{S}_F$	8.0	992.0	0.0	0.0	1000.0	66.6
		$\widehat{\mathcal{S}}$	5.4	0.0	0.0	0.0	5.4	45.7
		$\widehat{\mathcal{S}}+\widehat{\mathcal{I}}$	5.4	0.0	1.4	5.0	12.0	58.2
	500	$\mathcal{S}_F$	8.0	992.0	0.0	0.0	1000.0	66.6
		$\widehat{\mathcal{S}}$	7.8	0.0	0.0	0.0	7.8	65.5
		$\widehat{\mathcal{S}}+\widehat{\mathcal{I}}$	7.8	0.0	3.8	1.1	12.8	97.8
	1000	$\mathcal{S}_F$	8.0	992.0	0.0	0.0	1000.0	66.6
		$\widehat{\mathcal{S}}$	8.0	0.0	0.0	0.0	8.0	66.6
		$\widehat{\mathcal{S}}+\widehat{\mathcal{I}}$	8.0	0.0	4.0	0.2	12.2	100.0
5000	200	$\mathcal{S}_F$	8.0	4992.0	0.0	0.0	5000.0	66.6
		$\widehat{\mathcal{S}}$	4.9	0.0	0.0	0.0	4.9	41.2
		$\widehat{\mathcal{S}}+\widehat{\mathcal{I}}$	4.9	0.0	0.9	4.0	9.9	49.5
	500	$\mathcal{S}_F$	8.0	4992.0	0.0	0.0	5000.0	66.6
		$\widehat{\mathcal{S}}$	7.5	0.0	0.0	0.0	7.5	63.1
		$\widehat{\mathcal{S}}+\widehat{\mathcal{I}}$	7.5	0.0	3.5	1.7	12.8	92.9
	1000	$\mathcal{S}_F$	8.0	4992.0	0.0	0.0	5000.0	66.6
		$\widehat{\mathcal{S}}$	7.9	0.0	0.0	0.0	7.9	66.6
		$\widehat{\mathcal{S}}+\widehat{\mathcal{I}}$	7.9	0.0	3.9	0.2	12.2	99.9

Table 5. Example 3 detailed simulation results

			Main effect		Interaction effect			
p	n	Method	CME	IME	CIE	IIE	MS	CP%
1000	200	$\mathcal{S}_F$	20.0	980.0	0.0	0.0	1000.0	100.0
		$\widehat{\mathcal{S}}$	17.9	0.2	0.0	0.0	18.2	89.6
		$\widehat{\mathcal{S}}+\widehat{\mathcal{I}}$	17.9	0.2	0.0	0.3	18.5	89.6
1000	500	$\mathcal{S}_F$	20.0	980.0	0.0	0.0	1000.0	100.0
		$\widehat{\mathcal{S}}$	19.9	0.0	0.0	0.0	19.9	99.9
		$\widehat{\mathcal{S}}+\widehat{\mathcal{I}}$	19.9	0.0	0.0	0.0	19.9	99.9
1000	1000	$\mathcal{S}_F$	20.0	980.0	0.0	0.0	1000.0	100.0
		$\widehat{\mathcal{S}}$	20.0	0.0	0.0	0.0	20.0	100.0
		$\widehat{\mathcal{S}}+\widehat{\mathcal{I}}$	20.0	0.0	0.0	0.0	20.0	100.0
5000	200	$\mathcal{S}_F$	20.0	4980.0	0.0	0.0	5000.0	100.0
		$\widehat{\mathcal{S}}$	15.7	0.2	0.0	0.0	16.0	78.9
		$\widehat{\mathcal{S}}+\widehat{\mathcal{I}}$	15.7	0.2	0.0	1.0	17.1	79.1
5000	500	$\mathcal{S}_F$	20.0	4980.0	0.0	0.0	5000.0	100.0
		$\widehat{\mathcal{S}}$	19.9	0.0	0.0	0.0	19.9	99.9
		$\widehat{\mathcal{S}}+\widehat{\mathcal{I}}$	19.9	0.0	0.0	0.0	19.9	99.9
5000	1000	$\mathcal{S}_F$	20.0	4980.0	0.0	0.0	5000.0	100.0
		$\widehat{\mathcal{S}}$	20.0	0.0	0.0	0.0	20.0	100.0
		$\widehat{\mathcal{S}}+\widehat{\mathcal{I}}$	20.0	0.0	0.0	0.0	20.0	100.0

4.2 Labor Supply Dataset

We next consider a dataset about labor supply. This is an important dataset generously donated by Mroz (1987) and was discussed by Wooldridge (2002). It contains a total of 753 married white women aged between 30 and 60 in 1975. For illustration purpose, we take a binary variable $Y_i \in \{0, 1\}$ as the response of interest, which indicates whether the woman participated to the labor market or not. The dataset contains a total of 77 predictive variables with interaction terms included. These variables were observed for both participated and nonparticipated women. They are recorded by X_i. Understanding the regression relationship between X_i and Y_i is useful for calculating the propensity score for a woman's employment decision (Rosenbaum and Rubin 1983). However, due to its high dimensionality, directly using all the predictors for propensity score estimation is suboptimal. Thus, we are motivated to apply our method for variable screening.

Following similar strategy, we randomly split the dataset into two parts with equal sizes. One part is used for training and the other for testing. We then apply PC-SIS method to the training dataset. Because this dataset involves both continuous and categorical predictors, the method of discretization (as given in simulation Example 3) is used. We then apply PC-SIS to the discretized dataset, which leads to estimated model $\widehat{\mathcal{S}}$. Because the interaction terms with good economical meanings are already included in X_i (Mroz 1987), we did not further pursue the interaction model $\widehat{\mathcal{I}}$. An usual logistic regression model is then estimated based on the training dataset, and the resulting model's forecasting accuracy is evaluated on the testing data in terms of AUC, which is area under the ROC curve (Wang 2007). The definition is given as follows. Let $\hat{\beta}$ be the maximum likelihood estimator, which is obtained by conducting a logistic regression model for Y_i and X_i but based on the training data. Denote the testing dataset, which can be further decomposed as $\mathcal{T} = \mathcal{T}_0 \bigcup \mathcal{T}_1$ with $\mathcal{T}_0 = \{i \in \mathcal{T} : Y_i = 0\}$ and $\mathcal{T}_1 = \{i \in \mathcal{T} : Y_i = 1\}$. Simply speaking, $\mathcal{T}_0$ and $\mathcal{T}_1$ collect indices of those testing samples with response being 0 and 1, respectively. Then, AUC in Wang (2007) is defined as

$$\mathrm{AUC} = \frac{1}{n_0 n_1} \sum_{\{i_1 \in \mathcal{T}_1\}} \sum_{\{i_2 \in \mathcal{T}_0\}} I(X_{i_1}^\top \hat{\beta} > X_{i_2}^\top \hat{\beta}), \tag{4.1}$$

where n_0 and n_1 are the sample sizes of $\mathcal{T}_0$ and $\mathcal{T}_1$, respectively.

For comparison purpose, the full model $\mathcal{S}_F$ is also evaluated. For a reliable evaluation, the experiment is randomly replicated

Table 6. Detailed results for search engine marketing dataset

	Model	Main	Interaction	Forecasting accuracy %		
Method	Size	Effect	Effect	kNN	SVM	RF
$\mathcal{S}_F$	341.00	341.00	0.00	76.89	51.13	60.57
$\widehat{\mathcal{S}}$	14.60	14.60	0.00	85.20	87.19	67.55
$\widehat{\mathcal{S}}+\widehat{\mathcal{I}}$	36.85	14.60	22.25	86.96	88.66	78.01

200 times. We find that a total of 10.20 features are selected on average with AUC = 98.03%, which is extremely comparable to that of the full model (i.e., AUC=98.00%) but with substantially reduced features. Lastly, we apply our method to the whole dataset, with 10 important main effects identified and no interaction is included. The 10 selected main effects are, respectively, family income, after tax full income, wife's weeks worked last year, wife's usual hours of work per week last year, actual wife experience, salary, hourly wage, overtime wage, hourly wage from the previous year, and a variable indicating whose hourly wage from the previous year is not 0.

Remark 2. One can also evaluate AUC according to (4.1) but based on the whole sample and then optimize it with respect to an arbitrary regression coefficient β. This leads to the Maximum Rank Correlation (MRC) estimator, which has been well studied by Han (1987), Sherman (1993), and Baker (2003).

5. CONCLUDING REMARKS

To conclude this article, we discuss here two interesting topics for future study. First, as we discussed before, the proposed method and theory can be readily extended to the situation with general categorical predictors. Second, we assume here the number of response classes (i.e., K) is finite. How to conduct variable selection and screening with a diverging K is theoretically challenging.

APPENDIX: PROOF OF THEOREM 1

The proof of Theorem 1 consists of five steps. First, we show that there exists a lower bound on Δ_j for every $j \in \mathcal{S}_T$. Second, we establish $\widehat{\Delta}_j$ as a uniformly consistent estimator of Δ_j which is over $1 \le j \le p$. Last, we argue that there exists a positive constant c such that $\widehat{\mathcal{S}} = \mathcal{S}_T$ with probability tending to 1.

Step 1. By definition, we have $\omega_j^{k_1k_2} = \pi_{yj,k_1k_2} - \pi_{yk_1}\pi_{jk_2}$. Then for every $j \in \mathcal{S}_T$, by Condition (C1), π_{yk} and π_{jk} are both upper bounded by $\pi_{\max}$. We then have $\Delta_j = \sum_{k_1k_2}\{(\omega^{k_1k_2})^2(\pi_{yk_1}\pi_{jk_2})^{-1}\} \ge \pi_{\max}^{-2}\sum_{k_1k_2}(\omega_j^{k_1k_2})^2$. Next, by Condition (C2), if $j \in \mathcal{S}_T$, $\sum_{k_1k_2}(\omega_j^{k_1k_2})^2 \ge \max_{k_1k_2}(\omega_{k_1k_2}^2) \ge \omega_{min}$. These results together make Δ_j lower bounded by $\omega_{\min}\pi_{\max}^{-2}$. We can then define $\Delta_{\min} = 0.5\omega_{\min}\pi_{\max}^{-2}$, which is a positive constant resulting in $\min_{j\in S_T}\Delta_j > \Delta_{\min}$.

Step 2. The proof of uniform consistency for $\hat{\pi}_{jk}$ and $\hat{\pi}_{yj,k_1k_2}$ is similar. As a result, we omit the details of $\hat{\pi}_{yj,k_1k_2}$. Also, based on the uniform consistency of $\hat{\pi}_{jk}$ and $\hat{\pi}_{yj,k_1k_2}$, the uniform consistency of $\widehat{\Delta}_j$ needs only some standard argument using Taylor's expansion. The technical details of $\widehat{\Delta}_j$'s uniform consistency are also omitted. We focus on $\hat{\pi}_{jk}$ only.

To this end, we define $Z_{i,jk} = I(X_{ij} = k) - \pi_{jk}$. By that we know $EZ_{ij,k} = 0$, $EZ_{ij,k}^2 = \pi_{jk} - \pi_{jk}^2$, and $|Z_{ij,k}| \le M$ with $M = 1$. Also, for a fixed pair of (j, k), we know that $Z_{ij,k}$ are independent for i. All those conditions remind us of Bernstein's inequality, by which we have

$$P\left(\sum_i Z_{ij,k} > \varepsilon\right) \le \exp\left\{\frac{-3\varepsilon^2}{2M\varepsilon + 6n(\pi_{jk} - \pi_{jk}^2)}\right\},$$

where $\varepsilon > 0$ is an arbitrary positive constant. Since $M = 1$ and $\pi_{jk} - \pi_{jk}^2 \le 1/4$, the right-hand side of the inequality can further bounded above by $\exp\{-6\varepsilon^2/(4\varepsilon + 3n)\}$, Thus,

$$P\left(\left|\frac{1}{n}\sum_i Z_{ij,k}\right| > \varepsilon\right) \le 2\exp\left\{\frac{-6n^2\varepsilon^2}{4n\varepsilon + 3n}\right\} = 2\exp\left\{\frac{-6n\varepsilon^2}{4\varepsilon + 3}\right\}.$$

With $\hat{\pi}_{k,j} - \pi_{k,j} = n^{-1}\sum_i Z_{ij,k}$, we have

$$\begin{aligned} &P\left(\max_k \max_{1\le j\le p} |\hat{\pi}_{k,j} - \pi_{k,j}| > \varepsilon\right) \\ &\quad = P\left(\max_k \max_{1\le j\le p} \left|\frac{1}{n}\sum_i Z_{ij,k}\right| > \varepsilon\right) \\ &\quad \le \sum_{jk} P\left(\frac{1}{n}\left|\sum_i Z_{ij,k}\right| > \varepsilon\right) \\ &\quad \le 2K\exp\left\{\log p + \frac{-6n\varepsilon^2}{4\varepsilon + 3}\right\} \to 0, \end{aligned} \qquad \text{(A.1)}$$

where the first inequality is due to Bonferonni's inequality. By Condition (C3), the right-hand side of the final inequality goes to 0 as $n \to \infty$. Then we have, under Conditions (C1)–(C3), $\max_k \max_{1\le j\le p} |\hat{\pi}_{jk} - \pi_{jk}| = o_p(1)$.

Step 3. Recall that $\Delta_{\min} = 0.5\omega_{\min}\pi_{\max}^{-2}$. Define $c = (2/3)\Delta_{\min}$ and we should have $\widehat{\mathcal{S}} \supset \mathcal{S}_T$. Otherwise, there must exist a $j^* \in \mathcal{S}_T$ but $j^* \notin \widehat{\mathcal{S}}$. Accordingly, we must have $\widehat{\Delta}_{j^*} \le (2/3)\Delta_{\min}$ and $\Delta_{j^*} > \Delta_{\min}$. Thus $|\widehat{\Delta}_{j^*} - \Delta_{j^*}| > (1/3)\Delta_{\min}$, which implies, if $\widehat{\mathcal{S}} \not\supset \mathcal{S}_T$ then $\max|\widehat{\Delta}_j - \Delta_j| > (1/3)\Delta_{\min}$. On the other hand, we know by $\widehat{\Delta}_j$'s uniform consistency, with $\varepsilon = (1/3)\Delta_{\min}$, $P(\widehat{\mathcal{S}} \not\supset \mathcal{S}_T) \le P(\max|\widehat{\Delta}_j - \Delta_j| > (1/3)\Delta_{\min}) \to 0$, as $n \to \infty$.

Similarly, we have $\mathcal{S}_T \supset \widehat{\mathcal{S}}$. Or else there should be a $j^* \in \widehat{\mathcal{S}}$ but $j^* \notin \mathcal{S}_T$. Thus $\widehat{\Delta}_{j^*} \ge (2/3)\Delta_{\min}$ and $\Delta_{j^*} = 0$. We should have $|\widehat{\Delta}_{j^*} - \Delta_{j^*}| > (2/3)\Delta_{\min}$. Let $\varepsilon = (2/3)\Delta_{\min}$, and by uniform consistency again, we have $P(\mathcal{S}_T \not\supset \widehat{\mathcal{S}}) \le P(\max|\widehat{\Delta}_j - \Delta_j| > (2/3)\Delta_{\min}) \to 0$, as $n \to \infty$. As a result, we know that $P(\widehat{\mathcal{S}} = \mathcal{S}_T) \to 1$ with $c = (2/3)\Delta_{\min}$, as $n \to \infty$. This completes the proof.

ACKNOWLEDGMENTS

Their research was supported in part by National Natural Science Foundation of China (NSFC, 11131002, 11271032), Fox Ying Tong Education Foundation, the Business Intelligence Research Center at Peking University, and the Center for Statistical Science at Peking University. Wang's research was also supported in part by the Methodology Center at Pennsylvania State University. His research was supported by National Institute on Drug Abuse (NIDA) grant P50-DA10075. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or NIDA.

[Received April 2013. Revised September 2013.]

REFERENCES

Baker, S. G. (2003), "The Central Role of Receiver Operating Characteristics (ROC) Curves in Evaluating Tests for the Early Detection of Cancer," *Journal of the National Cancer Institute*, 95, 511–515. [243]

Breiman, L. (2001), "Random Forest," *Machine Learning*, 45, 5–32. [238]

Fan, J., Feng, Y., and Song, R. (2011), "Nonparametric Independence Screening in Sparse Ultra-High Dimensional Additive Models," *Journal of the American Statistical Association*, 116, 544–557. [237]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultra-High Dimensional Feature Space" (with discussion), *Journal of the Royal Statistical Society,* Series B, 70, 849–911. [237,238,239]

Fan, J., Samworth, R., and Wu, Y. (2009), "Ultrahigh Dimensional Feature Selection: Beyond the Linear Model," *Journal of Machine Learning Research*, 10, 1829–1853. [237]

Fan, J., and Song, R. (2010), "Sure Independent Screening in Generalized Linear Models With NP-Dimensionality," *The Annals of Statistics*, 38, 3567–3604. [237]

Han, A. K. (1987), "Nonparametric Analysis of a Generalized Regression Model," *Journal of Econometrics*, 35, 303–316. [243]

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, New York: Springer. [238]

He, X., Wang, L., and Hong, H. G. (2013), "Quantile-Adaptive Model-Free Variable Screening for High-Dimensional Heterogeneous Data," *The Annals of Statistics*, 41, 342–369. [237]

Kim, H., Howland, P., and Park, H. (2005), "Dimension Reduction in Text Classification With Support Vector Machines," *Journal of Machine Learning Research*, 6, 37–53. [238]

Li, G. R., Peng, H., Zhang, J., and Zhu, L. X. (2012), "Robust Rank Correlation Based Screening," *The Annals of Statistics*, 40, 1846–1877. [237]

Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening Via Distance Correlation Learning," *Journal of American Statistical Association*, 107, 1129–1139. [237]

Liu, J., Li, R., and Wu, R. (2014), "Feature Selection for Varying Coefficient Models With Ultrahigh Dimensional Covariates," *Journal of American Statistical Association*, 109, DOI: 10.1080/01621459.2013.850086. [237]

Mai, Q., and Zou, H. (2013), "The Kolmogorov Filter for Variable Screening in High-Dimensional Binary Classification," *Biometrika*, 100, 229–234. [237]

Mroz, T. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55, 765–799. [242]

Rosenbaum, P., and Rubin, D. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [242]

Sherman, R. P. (1993), "The Limiting Distribution of the Maximum Rank Correlation Estimator," *Econometrica*, 61, 123–137. [243]

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and Testing Dependence by Correlation of Distances," *The Annals of Statistics*, 35, 2769–2794. [237]

Tong, S., and Koller, D. (2001), "Support Vector Machine Active Learning With Application to Text Classification," *Journal of Machine Learning Research*, 2, 45–66. [238]

Wang, H. (2007), "A Note on Iterative Marginal Optimization: A Simple Algorithm for Maximum Rank Correlation Estimation," *Computational Statistics & Data Analysis*, 51, 2803–2812. [242]

——— (2009), "Forward Regression for Ultra-High Dimensional Variable Screening," *Journal of the American Statistical Association*, 104, 1512–1524. [237]

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge, MA: MIT Press. [242]

Zhu, L. P., Li, L., Li, R., and Zhu, L. X. (2011), "Model-Free Feature Screening for Ultrahigh Dimensional Data," *Journal of the American Statistical Association*, 106, 1464–1475. [237]