*The Annals of Statistics*
2015, Vol. 43, No. 4, 1471–1497
DOI: 10.1214/14-AOS1303


# THE FUSED KOLMOGOROV FILTER: A NONPARAMETRIC MODEL-FREE SCREENING METHOD

BY QING MAI[1] AND HUI ZOU[2]

*Florida State University and University of Minnesota*

A new model-free screening method called the fused Kolmogorov filter is proposed for high-dimensional data analysis. This new method is fully nonparametric and can work with many types of covariates and response variables, including continuous, discrete and categorical variables. We apply the fused Kolmogorov filter to deal with variable screening problems emerging from a wide range of applications, such as multiclass classification, nonparametric regression and Poisson regression, among others. It is shown that the fused Kolmogorov filter enjoys the sure screening property under weak regularity conditions that are much milder than those required for many existing nonparametric screening methods. In particular, the fused Kolmogorov filter can still be powerful when covariates are strongly dependent on each other. We further demonstrate the superior performance of the fused Kolmogorov filter over existing screening methods by simulations and real data examples.

**1. Introduction.** Consider a statistical problem with a response variable $Y$ and covariates $\mathbf{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$. When $p$ is very large, a popular assumption is the sparsity assumption that only a small subset of variables are actually responsible for modeling $Y$. To be specific, following Li, Zhong and Zhu (2012), define

$$\mathbf{D} = \{j : F(y \mid \mathbf{X}) \text{ functionally depends on } X_j \text{ for some } y\},$$

where $F(y \mid \mathbf{X})$ is the conditional cumulative probability function of $Y$. Then the sparsity assumption states that $|\mathbf{D}| \ll p$.

Variable selection aims to discover $\mathbf{D}$ exactly. Variable screening is less ambitious in that it only aims to discover a majority of $\mathbf{D}^c$. In other words, a good variable screening method tries to find a subset $\mathbf{S}$ such that $\mathbf{D} \subset \mathbf{S}$, which is referred to as the sure screening property [Fan and Lv (2008)] in the literature. Consistent variable selection is a very challenging task. It requires sophisticated estimation techniques, strong model assumptions and often advanced computing algorithms [Fan and Li (2001), Lv and Fan (2009), Tibshirani (1996), Zhang (2010)]. Because variable screening deals with a much less ambitious goal, it is possible that sure

Received October 2014.
[1]Supported by the FYAP grant from Florida State University.
[2]Supported an NSF grant and an ONR grant.
*MSC2010 subject classifications.* 62G99.
*Key words and phrases.* Variable screening, high-dimensional data, sure screening property.

screening could be achieved by using some simple (both conceptually and computationally) method. This idea was first successfully demonstrated in Fan and Lv (2008) where marginal correlation screening is shown to lead to sure screening results in high-dimensional linear regression under certain regularity conditions. Since the sure independence screening paper by Fan and Lv (2008), variable screening has received a lot of attention in the literature and many variable screening techniques, both parametric and nonparametric, have been proposed and studied in recent years [Chang, Tang and Wu (2013), Fan and Fan (2008), Fan, Feng and Song (2011), Fan, Samworth and Wu (2009), Fan and Song (2010), He, Wang and Hong (2013), Li, Zhong and Zhu (2012), Li et al. (2012), Mai and Zou (2013), Zhu et al. (2011)]. Variable screening is naturally appealing to practitioners, because if sure screening is achieved before doing a thorough analysis, the analysis part would become much easier with the screening subset. At least, the computational cost can be greatly reduced.

The main message in Fan and Lv (2008) is that although we should not do variable selection based on marginal correlations alone, marginal correlations can be used to filter out many noise variables and keep all important variables. Many new screening methods have been proposed with the aim of improving the marginal correlation screening method. Fan and Song (2010) propose a screening method based on the marginal maximum likelihood for generalized linear models. Chang, Tang and Wu (2013) propose using marginal empirical likelihood ratios to rank variables and demonstrate their good performance. The nonparametric independence screening (NIS) [Fan, Feng and Song (2011)] starts with a generalized additive model for modeling the regression response variable $Y$. For each variable $X_j$, NIS uses nonparametric smoothing, for example, B-spline regression, to obtain $\hat{m}_j = \arg\min_{m_j} \|Y - m_j(X_j)\|_n^2$. NIS then selects the variables with large $\|\hat{m}_j(X_j)\|_n^2$. Compared to marginal correlation learning, NIS is more robust because it captures nonlinear dependence between $Y$ and $X_j$. The quantile–adaptive screening (QA) [He, Wang and Hong (2013)] further improves the robustness of NIS by allowing heteroscedasticity in the model. Under such models, QA minimizes the check function instead of the squared error loss function to identify the important predictors. Li et al. (2012) propose using Kendall tau correlation to replace the usual Pearson correlation in marginal correlation screening so that the resulting screening method is more robust and can be useful under a semiparametric single-index model with a monotone link function. The distance correlation screening (DCS) [Li, Zhong and Zhu (2012)] is a model-free screening method that uses the distance correlation to replace Pearson correlation in marginal correlation screening. The distance correlation [Székely, Rizzo and Bakirov (2007)] between two random variables is zero if and only if they are independent. The Kolmogorov filter [Mai and Zou (2013)] is a fully nonparametric robust screening method. It deals with binary classification problems and uses the Kolmogorov–Smirnov test

statistic to screen covariates. The Kolmogorov filter has several unique, nice properties. First, it significantly outperforms other existing screening methods for binary classification problems. Second, it works with all types of covariates and is invariant under univariate monotone transformations of the covariates. Third, it can have the sure screening property even when the covariates are strongly dependent on each other. This result is very promising because it was commonly believed before Mai and Zou (2013) that marginal screening methods tend to work well if and only if the noise variables are weakly correlated with the important variables.

Fan and Lv (2008) suggest an iterative screening and model fitting procedure to deal with the strong correlation issue in model-based screening methods. Although this idea has been empirically demonstrated [Fan, Feng and Song (2011), Fan and Lv (2008), Fan and Song (2010), He, Wang and Hong (2013)], its theoretical justification still remains unknown. Furthermore, its theoretical justification heavily depends on model assumptions and hence may not be very robust. It is now clear that variable screening can be separated from the model fitting part. Both DCS [Li, Zhong and Zhu (2012)] and the Kolmogorov filter [Mai and Zou (2013)] have demonstrated that sure screening can be achieved without resorting to a particular form of model for the data. Moreover, we advocate the use of model-free screening methods in practice. The reasons are twofold. First, the model-free screening results are much more robust in the sense that the sure screening property can hold under much weaker conditions. The second reason is related to the choice of the statistical analysis tool in the modeling stage. Note that after the screening we have a low-dimensional dataset, and one may want to apply modern nonparametric learning methods such as boosting and random forest for further analysis [Hastie, Tibshirani and Friedman (2009)]. Yet model-based screening methods typically eliminate such choices because one has to stick with the model used in the first stage. For example, if we apply marginal correlation screening or marginal maximum likelihood screening, we have to use a linear regression model or generalized linear model in the second stage, although we can do penalized model fitting by using a penalty such as lasso [Tibshirani (1996)] or SCAD [Fan and Li (2001)]. If the underlying model for the data is highly nonlinear, then boosting or random forest is expected to be a better choice than linear models.

Our goal here is to develop a new fully nonparametric model-free variable screening method that can provide a unified solution to variable screening problems emerging from a wide variety of applications such as binary classification, multiclass classification, regression and Poisson regression, among others. The new method should also work with discrete, categorical or continuous covariates. Moreover, it is desirable to have the new method be invariant under univariate monotone transformations of response variable or covariates or both, because variable transformation models have wide applications in practice. Imagine that a variable transformation model is determined to be the best fit in the second modeling stage, we do wish to see that variable screening results should remain unchanged

if we would repeat the screening procedure by working with the transformed variables. DCS and the Kolmogorov filter are the two existing, fully nonparametric, model-free screening methods in the literature. Neither of them completely meets our expectations. DCS does not have the invariance property under monotone variable transformation, and its sure screening property heavily depends on a distribution assumption on covariates that they should have sub-exponential tails. In many applications, the covariates are heavy-tailed, and DCS may not be ideal in such cases. The limitation of the Kolmogorov filter is obvious as well: it is designed for binary classification problems and is inapplicable when the response variable can take more than two values.

To this end, we propose the fused Kolmogorov filter and study its theoretical and numerical properties. As the name suggests, the fused Kolmogorov filter is built upon two main ideas, the Kolmogorov–Smirnov test statistic, as used in Mai and Zou (2013), and fusion. When the response variable is binary, the fused Kolmogorov filter is exactly the Kolmogorov filter proposed in Mai and Zou (2013), and fusion is not needed. The fusion part becomes critically important when the response variable is continuous. We introduce two levels of fusion. In the first level, we slice the response variables into multiple slices, compute a Kolmogorov–Smirnov test statistic for each pair of slices and then take the supreme of all pairwise Kolmogorov–Smirnov test statistics. To make the method insensitive to the slicing scheme, we conduct the second level of fusion, where we repeat the first level for different ways of slicing and then take the sum of their outcomes as the final screening statistic, which we call the fused Kolmogorov statistic. The second level of fusion is important when the response variable is continuous or ordinal. The fused Kolmogorov filter ranks each covariate by its fused Kolmogorov statistic and screens out those covariates at the bottom of the rank list. By definition, the fused Kolmogorov filter is intuitively appealing, computationally convenient and automatically has the invariance property under monotone variable transformation.

The rest of the paper is organized as follows. The methodological details of the fused Kolmogorov filter are given in Section 2. In Section 3 we establish the sure screening property of the fused Kolmogorov filter under weak regularity conditions. We discuss these regularity conditions and find that they can hold, even when important variables and noise variables are strongly dependent. This promising result suggests that marginal variable screening could be more useful than we expected. Sections 4 and 5 contain simulated and real data examples. Technical proofs are presented in the Appendix.

## 2. Method.

2.1. *Motivation*. To see why the Kolmogorov–Smirnov statistic is very useful for variable screening, let us first revisit the binary Kolmogorov filter. When the response variable is binary, say $Y = 1, 2$, a variable $X$ is independent of $Y$ if

and only if the conditional distributions of $X$ given $Y = 1$ or $Y = 2$ are identical. Motivated by this simple fact, Mai and Zou (2013) propose using

$$K_j = \sup_x |F_j(x \mid Y = 1) - F_j(x \mid Y = 2)|$$

to measure the dependence between $X_j$ and $Y$, where $F_j$ denotes the generic cumulative distribution function (CDF) for $X_j$. Given the observed data, an empirical version of $K_j$ is defined as

$$\hat{K}_j = \sup_x |\hat{F}_j(x \mid Y = 1) - \hat{F}_j(x \mid Y = 2)|,$$

where $\hat{F}_j$ denotes the generic empirical CDF. Mai and Zou (2013) demonstrate the strong theoretical and numerical performance of the binary Kolmogorov filter.

Given the success of the binary Kolmogorov filter, it is natural to ask what its counterpart is for a continuous response variable or a general discrete variable (like counts data in Poisson regression). First, it seems straightforward to consider

$$K_j^* = \sup_{y_1, y_2} \sup_x |F_j(x \mid Y = y_1) - F_j(x \mid Y = y_2)| \tag{1}$$

because $K_j^* = 0$ if and only if $X_j$ is independent of $Y$. Thus $K_j^*$ is a natural generalization of $K_j$. In order to use $K_j^*$, we must have an empirical version of $K_j^*$. This step is trivial for the binary response case, but it is much more difficult when $Y$ takes infinite values because it requires the knowledge of $F_j(x \mid y)$ for all possible values $y$. On the other hand, we can find an approximation of $K_j^*$ by slicing the response. Define a partition

$$\mathbf{G} = \left\{ [a_l, a_{l+1}) : a_l < a_{l+1}, l = 0, \ldots, G-1 \text{ and } \bigcup_{j=1}^{G-1} [a_l, a_{l+1}) \setminus \{a_0\} = \mathbb{R} \right\},$$

where $a_0 = -\infty$ and $a_G = \infty$. Note that the interval $(a_0, a_1)$ is open, but we abuse the notation a little by writing the intervals $[a_l, a_{l+1})$ for all $l$. Each $[a_l, a_{l+1})$ is called a slice. We then define a random variable $H \in \{1, \ldots, G\}$ such that $H = l + 1$ if and only if $Y$ is in the $l$th slice. In particular, if $Y$ is discrete as in a multiclass problem, that is, $Y = 1, \ldots, G$, we can set $H = Y$. Now let

$$K_j^{\mathbf{G}} = \max_{l,m} \sup_x |F_j(x \mid H = l) - F_j(x \mid H = m)|,$$

where $F_j(x \mid H = l) = \Pr(X_j \le x \mid H = l)$.

The idea of slicing is very natural. First, If $Y$ is binary, $K_j^{\mathbf{G}}$ and $K_j$ are the same. If $Y$ is multiclass, the slicing breaks the multiclass problem into pairwise binary problems. This strategy has been proven successful as a method for generalizing a binary classifier to its multiclass counterpart [Hastie and Tibshirani (1998)]. Yet $K_j^{\mathbf{G}}$ can be still be computed when $Y$ is a count that takes infinite discrete values, such as in the Poisson regression model. When $Y$ is continuous, slicing is widely

used in the field of sufficient dimension reduction [Cook and Weisberg (1991), Li (1991)] to infer about the conditional means and/or variances of predictors. However, these sufficient dimension reduction methods generally deal with problems with large sample sizes compared to the dimension. To the best of our knowledge, this paper is the first to utilize slicing for variable screening for large $p$ and small $n$ problems.

It is obvious that $X_j$ is independent of $Y$ if and only if $K_j^{\mathbf{G}} = 0$ when $Y$ takes finite values and each possible value forms a slice. In what follows, we assume that $Y$ is continuous, as it is the more challenging case. The following lemma shows that $K_j^{\mathbf{G}}$ sheds light on the dependence between $Y$ and $X_j$ as well when $Y$ is continuous.

LEMMA 1. (a) *$X_j$ is independent of $Y$ if and only if $K_j^{\mathbf{G}} = 0$ for all possible choices of $\mathbf{G}$.*

(b) *Assume that $X_j$ is not independent of $Y$ and for any fixed $y \in \mathbb{R}$, $\Pr(Y \le y \mid X_j = x)$ is not a constant in $x$; then $K_j^{\mathbf{G}} \neq 0$ for any $\mathbf{G}$.*

(c) *Assume that $F_j(x \mid y)$ is continuous in $y$. If $\max_{l=1,\ldots,G} \Pr(H = l) \to 0$ as $G \to \infty$, then $K_j^{\mathbf{G}} \to K_j^*$ as $G \to \infty$, where $K_j^*$ is defined in (1). Therefore, for $X_j$ not independent of $Y$, $K_j^{\mathbf{G}} > 0$ for sufficiently large $G$.*

Although we initially proposed $K_j^{\mathbf{G}}$ as a surrogate of $K_j^*$ and Lemma 1 part (c) indicates this as well, it turns out that $K_j^{\mathbf{G}}$ could be a better measure for variable screening than $K_j^*$. To see this interesting point, we present the following lemma.

LEMMA 2. *If $(X_j, Y)$ has a bivariate Gaussian copula distribution such that, after transformation via two monotone functions $g_1, g_2$, $(g_1(X_j), g_2(Y))$ is jointly normal with correlations $\rho_j = \operatorname{Cor}(g_1(X_j), g_2(Y))$ and $g_1(X_j), g_2(Y)$ are marginally standard normal. Then we have the following two conclusions*:

(a) *$K_j^* = 1$ if $\rho_j \neq 0$ and $K_j^* = 0$ otherwise.*

(b) *Suppose $Y$ is sliced at $\frac{l}{G}$'th quantile of $Y$ for $l = 1, \ldots, G-1$. Then $K_j^{\mathbf{G}}$ can be expressed as*

$$K_j^{\mathbf{G}} = G \int_{-\infty}^{\Phi^{-1}(1/G)} \left( 2\Phi\left( \frac{-|\rho_j| y}{\sqrt{1-\rho_j^2}} \right) - 1 \right) \frac{e^{-y^2/2}}{\sqrt{2\pi}} \, \mathrm{d}y,$$

*where $\Phi$ is the CDF for the standard normal distribution. Consequently, for any $G$, $K_j^{\mathbf{G}}$ is a strictly increasing function in $|\rho_j|$.*

With Lemma 2 in mind, we revisit the variable screening problem under a high-dimensional linear regression model as examined in [Fan and Lv (2008)]. For simplicity, assume that the model is

$$Y = X_1 + X_2 + Z$$

and

$$X_j = aX_1 + Z_j, \qquad j \geq 3,$$

where $X_1, X_2, Z, Z_j$ are independent $N(0,1)$ variables. Then we have

$$\operatorname{Cor}(X_1, Y) = \operatorname{Cor}(X_2, Y) = \frac{1}{\sqrt{3}},$$

$$\operatorname{Cor}(X_j, Y) = \frac{a}{\sqrt{3(1+a^2)}} \qquad \text{for } j = 3, \ldots, p.$$

So this is a perfect case for using the marginal correlation screening of Fan and Lv (2008). By Lemma 2 we have the following results:

$$K_j^* = 1, \qquad j = 1, 2, 3, \ldots,$$

$$K_1^{\mathbf{G}} = K_2^{\mathbf{G}} > K_j^{\mathbf{G}}, \qquad j = 3, \ldots.$$

Thus $K_j^*$ cannot separate $(X_j, j \geq 3)$ from $X_1, X_2$ no matter how small $a$ is. On the other hand, $K_j^{\mathbf{G}}$ works perfectly in this example, just like the marginal correlations. Of course, $K_j^{\mathbf{G}}$ in general works much better than $\operatorname{Cor}(X_j, Y)$, which will be clearly demonstrated in the later sections.

2.2. *The fused Kolmogorov filter.* In this subsection we show how to use $K_j^{\mathbf{G}}$ for variable screening based on a random sample $(\mathbf{X}^i, Y^i)_{i=1}^n$. We first need to estimate $K_j^{\mathbf{G}}$ accurately for all $p$ variables. Given a partition $\mathbf{G}$, we estimate $K_j^{\mathbf{G}}$ by

$$\hat{K}_j^{\mathbf{G}} = \max_{(l,m)} \sup_y \bigl|\hat{F}_j(x \mid H_j = l) - \hat{F}_j(x \mid H_j = m)\bigr|,$$

where

$$\hat{F}_j(x \mid H = l) = \frac{1}{n_l} \sum_{H^i = l} 1\bigl(X_j^i \leq x\bigr),$$

and $n_l$ is the sample size within the $l$th slice, and $H^i = l$ if $Y^i$ is in the $l$th slice.

If $Y$ is a multi-level categorical variable, then the partition is simply done according to $Y$'s value. When $Y$ has infinitely many possible values, the partition/slicing scheme can be important. With finite sample size, it is important to have enough sample sizes within each slice to control the estimation variance. As mentioned in the Introduction, the idea of slicing response variable has been used by researchers in sufficient dimension reduction. Early researchers proved that the sliced inverse regression (SIR) can be consistent even when there are only two observations in each slice [Hsing and Carroll (1992), Li (1991)], which implies that SIR is reasonably insensitive to the slicing scheme. Yet Zhu and Ng (1995) later observed that, even though SIR can be consistent for all slicing schemes

with the same number of observations in each slice, there is a loss of efficiency when there are too many slices. Based on our experience, the choice of slices does not affect variable screening results very much. However, significant improvement can be achieved by fusion. Suppose that we have $N$ different partitions, $\mathbf{G}_i$ for $i = 1, \ldots, N$, where each partition $\mathbf{G}_i$ contains $G_i$ intervals. Then we let

$$\hat{K}_j = \sum_{i=1}^{N} \hat{K}_j^{\mathbf{G}_i}.$$

By doing so, we combine the information from all $\mathbf{G}_i$. This fusion step is motivated by Cook and Zhang (2014), who showed that in sufficient dimension reduction, combining several slicing schemes works better than the usual practice relying on a single slicing scheme. As shown in Section 4, fusion does yield variable screening results that are superior to using a single slicing scheme.

We suggest an intuitive uniform slicing to partition data into $G$ slices. If $Y$ is categorical with levels $1, \ldots, G$, or $Y$ is discrete with finite possible values $1, \ldots, G$, we set $H = Y$. If $Y$ is discrete and can take infinite values as in a Poisson regression model, we set $H = Y + 1$ if $Y < G - 1$ and $H = G$ if $Y \geq G - 1$. For the case where $Y$ is continuous, we let the partition $\mathbf{G}$ contain the intervals bounded by the $\frac{l}{G}$th sample quantiles of $Y$ for $l = 0, \ldots, G$. From now on, we always write $\hat{K}_j(G) = \hat{K}_j^{\mathbf{G}}$ when $\mathbf{G}$ is a uniform partition with $G$ slices. By fusion, we consider multiple uniform slicing $\mathbf{G}_i, 1 \leq i \leq N$ where $\mathbf{G}_i$ has $G_i$ many slices. In practice, we suggest choosing $G_i \leq \lceil \log n \rceil$ for all $i$ so that there is a decent sample size within each slice for all slicing schemes. This is important because the fused Kolmogorov filter is a fully nonparametric method and sample size plays a central role in nonparametric statistics. Then the final fused Kolmogorov filter statistic is

$$\hat{K}_j = \sum_{i=1}^{N} \hat{K}_j^{\mathbf{G}_i}, \tag{2}$$

and the fused Kolmogorov filter screening set is defined as

$$\hat{\mathbf{D}} = \{j : \hat{K}_j \text{ is among the } d_n\text{'th largest}\}. \tag{3}$$

**3. Theory.** In this section we establish the sure screening property of the fused Kolmogorov filter.

3.1. *Main theorem.* We first introduce a concept called the oracle fused Kolmogorov filter. If we know the distribution of $Y$, then we can use an oracle uniform slicing such that the partition $\mathbf{G}_i$ contains the intervals bounded by the $\frac{l}{G_i}$th theoretical quantiles of $Y$ for $l = 0, \ldots, G_i$. For this special slicing, write $K_j^{(o)}(G_i) = K_j^{\mathbf{G}_i}$ and $K_j^{(o)} = \sum_i K_j^{(o)}(G_i)$. Then we can obtain a screening set as $\hat{\mathbf{D}}(\text{oracle}) = \{j : \hat{K}_j^{(o)} \text{ is among the } d_n\text{'th largest}\}$, where $d_n$ is a predefined positive integer. Throughout this section, $C$ denotes a generic positive constant.

To show the sure screening property of the fused Kolmogorov filter, we consider the following two regularity conditions:

*Regularity conditions*.

(C1) There exists a set $\mathbf{S}$ such that $\mathbf{D} \subset \mathbf{S}$ and

$$\Delta_{\mathbf{S}} = \min_i \Big( \min_{j \in \mathbf{S}} K_j^{(o)}(G_i) - \max_{j \notin \mathbf{S}} K_j^{(o)}(G_i) \Big) > 0.$$

(C2) Let $G_{\min} = \min_i \{G_i\}$. Then for any $b_1, b_2$ such that $\Pr(Y \in [b_1, b_2)) \leq 2/G_{\min}$, we have

$$\big| F_j(x \mid y_1) - F_j(x \mid y_2) \big| \leq \frac{\Delta_{\mathbf{S}}}{8} \tag{4}$$

for all $x$, $j$ and $y_1, y_2 \in [b_1, b_2)$.

THEOREM 1. *Assume conditions* (C1) *and* (C2). *Define*

$$\eta = CNp(\log^2 n) \exp\left( -C \frac{n \Delta_{\mathbf{S}}^2}{\log n} \right) + CN(\log^2 n) \exp\left( -C \frac{n}{\log^2 n} \right).$$

*If* $G_i \leq \lceil \log n \rceil$ *for all* $i$ *and* $d_n \geq |S|$, *we have the following conclusions*:

(1) *For the oracle fused Kolmogorov filter, we have*

$$\Pr\big( \mathbf{D} \subset \hat{\mathbf{D}}(\text{oracle}) \big) \geq 1 - \eta. \tag{5}$$

*Therefore, the oracle fused Kolmogorov filter enjoys the sure screening property with a probability tending to one if* $\Delta_{\mathbf{S}} \gg \sqrt{\frac{\log n \cdot \log(pN \log n)}{n}}$.

(2) *For the fused Kolmogorov filter defined in* (2) *and its screening set defined in* (3), *we have*

$$\Pr(\mathbf{D} \subset \hat{\mathbf{D}}) \geq 1 - \eta. \tag{6}$$

*Therefore, the fused Kolmogorov filter enjoys the sure screening property with a probability tending to one if*

$$\Delta_{\mathbf{S}} \gg \sqrt{\frac{\log n \log(pN \log n)}{n}}. \tag{7}$$

REMARK 1. By comparing (5) and (6), we see that the fused Kolmogorov filter can handle the same order of dimensions as the oracle fused Kolmogorov filter. Therefore, slicing at the sample quantiles results in a method that is as powerful as one utilizing oracle information about the theoretical quantiles. Also, Theorem 1 sheds light on the choice of $\mathbf{G}_i$. The minimum number of slices was 3 in Cook and Zhang (2014). Then Theorem 1 requires that $G_i \leq \lceil \log n \rceil$, with each $\mathbf{G}_i$ containing $G_i$ intervals bounded by sample quantiles. Therefore, in practice, we suggest setting $G_i = 3, \ldots, \lceil \log n \rceil$, with each $\mathbf{G}_i$ containing $G_i$ intervals bounded by sample quantiles.

REMARK 2. One could obtain a limit on the dimension for the fused Kolmogorov filter from Theorem 1. Suppose we choose the slicing scheme recommended in Remark 1. It follows that $N \leq \log n$. Then if there exists $0 < \kappa < 1$ such that $\Delta_{\mathbf{S}} \gg n^{-\kappa}$, (7) reduces to

$$\log p \ll n^{\xi},$$

for any $\xi \in (0, 1 - 2\kappa)$. Note that this restriction on $p$ is the same as that for SIS; see Conditions 1 and 3 in Fan and Lv (2008). Therefore, the fused Kolmogorov filter can handle the same order of dimensions as SIS without imposing any parametric assumptions.

REMARK 3. Theorem 1 shows that the fused Kolmogorov filter enjoys the sure screening property with a probability tending to one as long as we choose a reasonably large $d_n$. One interesting fact is that (6) does not involve $d_n$ explicitly. It holds as long as $d_n \geq |\mathbf{S}|$. This insensitivity to $d_n$ leads to tremendous practical convenience, because we can always use a reasonably large $d_n$ to guarantee a high probability of enjoying the sure screening property. In particular, when performing variable selection, one often assumes that the number of important variables is less than $n$. For example, lasso can only produce up to $n$ nonzero coefficients. Therefore, when we apply the fused Kolmogorov filter, we can use $d_n = a\lceil \frac{n}{\log n} \rceil$ where $a$ is some constant. A more conservative choice could be $d_n = n$.

REMARK 4. With the regularity conditions (C1)–(C2), the sure screening property results from the fact that $\hat{K}_j$ are close to $K_j^{(o)}$, which is a consequence of the Dvoretzky–Kiefer–Wolfowitz inequality. In the following subsection, we further discuss the implications of the two regularity conditions.

3.2. *Comments on the regularity conditions.* The conditions for Theorem 1 are very mild. First, note that, in contrast to DCS [Li, Zhong and Zhu (2012)], we make no assumption on the distribution of $\mathbf{X}$. Therefore, the fused Kolmogorov filter is expected to be more powerful than DCS when the predictors are heavy-tailed. Moreover, we do not assume any form of the dependence of $Y$ on $\mathbf{X}$. So the fused Kolmogorov filter will be more flexible than NIS and QA. The only two conditions we require are conditions (C1) and (C2).

We first comment on condition (C2). This condition is slightly stronger than requiring $F_j(x \mid y)$ to be continuous in $y$, as in Conclusion (c) of Lemma 1. Such a condition guarantees that the sample quantiles of $Y$ are close enough to the population quantiles of $Y$. Obviously, this result is expected for many distributions of $Y$. A conseqence is that the actual slicing used in practice is very close (asymptotically) to the oracle slicing such that $\hat{K}_j$'s accurately approximate $\hat{K}_j^{(o)}$'s.

In order to establish the sure screening property, a nontrivial condition is needed. For example, the partial orthogonality condition, that is, $\mathbf{X}_{\mathbf{D}} \perp \mathbf{X}_{\mathbf{D}^C}$ [Fan and Song

(2010), Huang, Horowitz and Ma (2008)], has been considered in the literature. Clearly, the theory is more interesting when $\mathbf{X_D}$ and $\mathbf{X_{D^C}}$ are dependent. In our theory, condition (C1) is the core condition which is used to guarantee that jointly important predictors, that is, the predictors belonging to the set $\mathbf{D}$, should also be marginally important, which is more or less assumed in the theory for existing marginal screening methods in the literature. In the context of binary classification, it has been shown that the sure screening property of the Kolmogorov filter can be established even when $\mathbf{X_D}$ and $\mathbf{X_{D^C}}$ are strongly correlated [Mai and Zou (2013)]. This phenomenon can be generalized to the multiclass classification rather directly, whose derivation is omitted here for the sake of space. In what follows we focus on the case where $Y$ is continuous to show that condition (C1) can still be true even when $\mathbf{X_D}$ and $\mathbf{X_{D^C}}$ are strongly correlated, and hence the sure screening property can hold with high probability. We highlight this interesting point by considering the following variable-transformation linear normal model:

$$T_y(Y) = \mathbf{T}(\mathbf{X})^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon, \tag{8}$$

where $\mathbf{T} = (T_1, \ldots, T_p)$ and $T_y, T_1, \ldots, T_p$ are strictly monotone univariate transformations. It is also assumed that $\mathbf{T}(\mathbf{X}) \sim N(0, \boldsymbol{\Sigma})$ with $\Sigma_{jj} = 1$ for $j = 1, \ldots, p$, and $\varepsilon \sim N(0, \sigma^2)$ is independent of $\mathbf{X}$. Note that $(T_y, \mathbf{T})$ are unknown, and we do not assume any parametric forms for them. Therefore, (8) is a very flexible semiparametric regression model. The main idea in model (8) is that after whitening each variable in the dataset we could fit a linear regression model. This interesting model has close connections to many transformation models in the literature; for example, see Breiman and Friedman (1985), He and Shen (1997), Li et al. (2012).

LEMMA 3. *Consider the model in* (8). *Without loss of generality, assume that* $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathbf{D}}, 0)$. *Define* $\boldsymbol{\alpha} = \boldsymbol{\Sigma}\boldsymbol{\beta}$. *Then for any set of* $\mathbf{G}_i, i = 1, \ldots, N$, *we have*:

(1) *Condition* (C1) *is true if and only if there exists* $\mathbf{S}$ *such that* $\min_{j\in\mathbf{S}} |\alpha_j| > \max_{j\notin\mathbf{S}} |\alpha_j|$.

(2) *If* $\boldsymbol{\Sigma}$ *is blockwise diagonal, that is,* $\sigma_{ij} = 0$ *if* $i \in \mathbf{D}$, $j \notin \mathbf{D}$, *then* $\Delta_{\mathbf{D}} > 0$ *if and only if* $\min_{j\in\mathbf{D}} |\alpha_j| > 0$.

(3) *Suppose* $\Sigma_{ij} = \rho^{|i-j|}$. *If* $\min_{j\in\mathbf{D}} |\alpha_j| > 0$ *and we let*

$$\mathbf{S} = \left\{1, \ldots, d + \left\lceil \frac{\log\left(\min_{j\in\mathbf{D}} |\alpha_j|/|\alpha_d|\right)}{\log|\rho|} \right\rceil\right\},$$

*then* $\Delta_{\mathbf{S}} > 0$.

(4) *Suppose* $\Sigma_{ij} = \rho$ *and* $\Sigma_{jj} = 1$. *Define* $\mathbf{S} = \{j : \alpha_j \neq 0\}$. *Then* $\Delta_{\mathbf{S}} > 0$. *Moreover,* $\mathbf{D} \subset \mathbf{S}$ *if and only if* $1^{\mathrm{T}}\boldsymbol{\beta} = 0$.

(5) *Suppose* $\Sigma_{ij} = \rho$ *and* $\Sigma_{jj} = 1$. *Then* $\Delta_{\mathbf{D}} > 0$ *if* $\rho > 0$ *and* $\beta_j$ *has the same sign for all* $j \in \mathbf{D}$.

In the following we discuss the implications of Lemma 3.

REMARK 5. In part (3) where the covariance has an autoregressive structure, to ensure the sure screening property, we need $d_n \geq d + \lceil \frac{\log(\min_{j\in\mathbf{D}} |\alpha_j|/|\alpha_d|)}{\log|\rho|} \rceil$. It follows that

$$|\rho| \leq \exp\left(\frac{\log(\min_{j\in\mathbf{D}} |\alpha_j|/|\alpha_d|)}{d_n - d}\right).$$

With $d_n = \lceil \frac{n}{\log n} \rceil$, the upper bound of $|\rho|$ tends to 1. Therefore, there is little restriction on $\rho$. In parts (4) and (5) where $\mathbf{\Sigma}$ has the compound symmetry structure, $\rho$ can be arbitrary as well.

REMARK 6. A direct calculation shows that in the fused Kolmogorov filter, $K_j$ is monotone in $\alpha_j$, while the joint importance $X_j$ is measured by $\beta_j$. Part (2) of Lemma 3 corresponds to the partial orthogonality condition under which the important variables and noise variables are independent, so this is an expected result. Somewhat surprisingly, parts (3)–(5) of Lemma 3 show that even when the predictors are highly correlated, condition (C1) still holds. Then by Theorem 1, the fused Kolmogorov filter will enjoy the sure screening property with high probability.

REMARK 7. Let us consider the normal linear model where we further assume $T_y(Y) = Y$ and $T_j(X_j) = X_j$, that is, $Y = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon$, where $\mathbf{X} \sim N(0, \mathbf{\Sigma})$. Lemma 3 can be applied to marginal correlation screening (SIS) and distance correlation screening (DCS). However, the fused Kolmogorov filter is more flexible than SIS, DCS and many other screening methods because it is invariant under monotone transformations. Many existing screening methods, except rank correlation screening [Li et al. (2012)], do not have this nice invariance property. As a result, when the true mode is a transformation normal linear model, SIS and DCS can perform poorly, while the fused Kolmogorov filter's performance remains the same, regardless of the transformations. We will clearly demonstrate this point in the simulation study in Section 4.

## 4. Simulations.

4.1. *Simulation design.* In this section, we compare the fused Kolmogorov filter with existing screening methods on simulated datasets. In all the models, we set $n = 200$, $p = 5000$. We consider the fused Kolmogorov filter based on $K_j(G_i)$ for $G_i = 3, \ldots, 6$, because $\lceil \log n \rceil = 6$. When the response is continuous, we slice $Y$ at $\frac{l}{G_i}$th sample quantiles for $l = 1, \ldots, G_i - 1$. We further include six other successful screening methods in the literature for comparison, marginal correlation screening (SIS) [Fan and Lv (2008)], nonparametric independence screening (NIS) [Fan, Feng and Song (2011)], distance correlation screening (DCS) [Li, Zhong and Zhu (2012)], rank correlation screening (RCS) [Li et al. (2012)], empirical likelihood screening (ELS) [Chang, Tang and Wu (2013)] and the quantile–adaptive

screening (QA) [He, Wang and Hong (2013)]. In all the models, we use SIS to denote the linear screening method. For example, if the response is continuous, SIS is the original marginal correlation screening. For the generalized linear model we use SIS to denote the marginal maximum likelihood estimator (MMLE) [Fan and Song (2010)]. When $Y$ is a multi-level categorical variable, SIS fits $p$ multinomial models with the `R` package `nnet` [Venables and Ripley (2002)] and selects the predictors with the largest deviances. With a little abuse of notation, we refer to all these methods as SIS when it is clear from the context.

Following He, Wang and Hong (2013), we consider $\alpha = 0.5, 0.75$ for QA. We use the implementation of NIS and QA at http://users.stat.umn.edu/~wangx346/research/example1b.txt. The distance correlation is computed by the `R` package `energy`. For ELS, we use the implementation of ELS by the authors of Chang, Tang and Wu (2013). As in Fan and Lv (2008), we report the minimum number of predictors needed to keep all the useful predictors. The results are based on 500 replicates. We consider the following six models in this simulation study:

*Model* 1. $T_y(Y) = \mathbf{T}(\mathbf{X})^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon$, where $\boldsymbol{\beta} = 2.8 \times (1, -1, 0_{p-2})$, $\mathbf{T}(\mathbf{X}) \sim N(0, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \mathrm{CS}(0.7)$, $\varepsilon \sim N(0, 1)$ is independent of $\mathbf{X}$. We consider three sets of $(T_y, \mathbf{T})$:

(a) $T_y(Y) = Y, T_j(X_j) = X_j$;
(b) $T_y(Y) = Y, T_j(X_j) = X_j^{1/9}$;
(c) $T_y(Y) = Y^{1/9}, T_j(X_j) = X_j$.

Models 1(a), 1(b) and 1(c) are examples of model (8) with a compound symmetry correlation matrix of which the correlation coefficient is 0.7.

*Model* 2. $Y = \mathbf{T}(\mathbf{X})^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon$, where $\boldsymbol{\beta} = 0.8 \times (1_{10}, 0_{p-10})$. $\mathbf{T}(\mathbf{X}) \sim N(0, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \mathrm{AR}(0.7)$. Again, we consider three sets of $(T_y, \mathbf{T})$:

(a) $T_y(Y) = Y, T_j(X_j) = X_j$;
(b) $T_y(Y) = Y, T_j(X_j) = \frac{1}{2}\log X_j$;
(c) $T_y(Y) = \log(Y), T_j(X_j) = X_j$.

Models 2(a), 2(b) and 2(c) are examples of model (8) with an autoregressive correlation matrix of which the autoregressive correlation coefficient is 0.7.

*Model* 3 (*Single index regression model*). $Y = (X_1 + X_2 + 1)^3 + \varepsilon$, where $X_j$'s follow the Cauchy distribution independently and $\varepsilon \sim N(0, 1)$ is independent of $\mathbf{X}$.

*Model* 4 (*Additive model*). $Y = 4X_1 + 2\tan(\pi X_2/2) + 5X_3^2 + \varepsilon$, where $X_j$'s follow Unif(0, 1) independently and $\varepsilon \sim N(0, 1)$ is independent of $\mathbf{X}$.

*Model* 5 (*Heteroskedastic regression model*). $Y = 2(X_1 + 0.8X_2 + 0.6X_3 + 0.4X_4 + 0.2X_5) + \exp(X_{20} + X_{21} + X_{22})\varepsilon$, where $\varepsilon \sim N(0, 1)$, and $\mathbf{X} \sim N(0, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \mathrm{AR}(0.8)$. This model is adapted from He, Wang and Hong (2013). In He, Wang and Hong (2013), they report the minimum number of predictors to keep the

first five predictors for QA with $\alpha = 0.5$ because QA with $\alpha = 0.5$ can only detect the predictors affecting the median. However, it is difficult to use such information for other methods. Therefore, we report the minimum number of predictors we need to keep all the eight important predictors for QA with $\alpha = 0.5$ too, so that it is fair to other methods.

*Model* 6 (*Poisson regression model*). $Y \sim \text{Poisson}(\mu)$, where $\mu = \exp(\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta})$, $\boldsymbol{\beta} = (0.8, -0.8, 0_{p-2})$, $X_j \sim t_2$ independently. The counterpart for SIS for this model is the marginal maximum likelihood estimator (MMLE) [Fan and Song (2010)]. Note that the predictors are heavy-tailed in this model, and $Y$ may consequently have extreme outliers. Therefore, to resolve computational issues, we delete an observation whenever $Y > 1000$ in MMLE. In addition, we consider the Kolmogorov filter and DCS on this model because all other methods are inapplicable to such datasets. Now, for the Kolmogorov filter, we set $H = Y$ if $Y < 2$; otherwise, $H = 3$.

*Model* 7 (*Multiclass classification model*). $Y = 1, \ldots, 5$. For each $g$, if $Y = g$, $X_{2(g-1)+1}$ and $X_{2g}$ independently follow $0.5N(3, 0.3^2) + 0.5N(-3, 0.3^2)$, and $X_j$ follows the Cauchy distribution independently for all other $j$. The counterpart for SIS for this model is to screen the predictors by marginally performing multinomial regression. Other than SIS, only the Kolmogorov filter and DCS are applicable to this model. Because $Y$ is categorical, we directly take $H = Y$ for the Kolmogorov filter and apply no further fusion. For DCS, we create a dummy variable $Y^{\mathrm{dm}} \in \mathbb{R}^{n\times 5}$ and compute the distance correlation between $Y^{\mathrm{dm}}$ and $X_j$.

4.2. *Simulation results and conclusions*. The simulation results are reported in Table 1. There are two important conclusions.

- We see that the Kolmogorov filter using a single slicing works reasonably well, and its performance is rather insensitive to the choice of number of slices. Nevertheless, the Kolmogorov filters with fewer slices tend to be more efficient when the underlying model is simple, such as in Model 1 where the true model is a transformed linear model. On the other hand, the Kolmogorov filters with more slices tend to be more accurate when the model is complicated, such as in Model 5. However, by combining different slicing schemes, the fused Kolmogorov filter has the best overall performance. The fused Kolmogorov filter is at least as good as the best $\hat{K}_j(G_i)$ in Models 1–3. In Models 4 and 5, where the fused Kolmogorov filter is slightly worse than the $\hat{K}_j(G_i)$ with the best $G_i$, the difference is very small.
- Compared with SIS, DCS, NIS, ELS and QA, the fused Kolmogorov filter is either the best or one of the best, and outperforms the rest by a large margin. This clearly shows that the fused Kolmogorov filter is a superior screening technique.

This simulation also reveals some major drawbacks of the existing screening methods. Although SIS, DCS, NIS and ELS work well in Models 1(a) and 2(a),

TABLE 1

*Simulation results for Models* 1–7. *We report the minimum number of predictors needed to keep all the useful predictors. The numbers in the table are medians of* 500 *replicates. Standard errors are in parentheses. A cell is left empty if the corresponding method is not applied to the specific model*

| | **Model** 1 | | | **Model** 2 | | |
|---|---|---|---|---|---|---|
| | **(a)** $d = 2$ | **(b)** $d = 2$ | **(c)** $d = 2$ | **(a)** $d = 10$ | **(b)** $d = 10$ | **(c)** $d = 10$ |
| Kolmogorov | | | | | | |
| $G = 3$ | 4 (0.5) | 4 (0.5) | 4 (0.5) | 10 (0) | 10 (0) | 10 (0) |
| $G = 4$ | 6 (0.9) | 6 (0.9) | 6 (0.9) | 10 (0) | 10 (0) | 10 (0) |
| $G = 5$ | 12 (1.6) | 12 (1.6) | 12 (1.6) | 10 (0) | 10 (0) | 10 (0) |
| $G = 6$ | 21 (3.2) | 21 (3.2) | 21 (3.2) | 10 (0) | 10 (0) | 10 (0) |
| Fused | 2 (0.3) | 2 (0.3) | 2 (0.3) | 10 (0) | 10 (0) | 10 (0) |
| SIS | 2 (0) | 1636 (93.5) | 486.5 (100.5) | 10 (0) | 1552.5 (99.2) | 1084.5 (62.9) |
| DCS | 2 (0) | 354 (34.8) | 229 (54.0) | 10 (0) | 10 (0) | 543 (52.8) |
| RCS | 2 (0) | 2 (0) | 2 (0) | 10 (0) | 10 (0) | 10 (0) |
| NIS | 2 (0) | 2 (0.4) | 1214 (79.0) | 10 (0) | 10 (0) | 1462.5 (92.8) |
| ELS | 2 (0) | 2879 (103.4) | 2460.5 (87.7) | 10 (0) | 565 (287.8) | 4401 (36.9) |
| QA | | | | | | |
| $\tau = 0.5$ | 5 (0.6) | 30.5 (5.4) | 5 (0.6) | 10 (0) | 10 (0) | 12 (0.4) |
| $\tau = 0.75$ | 13.5 (1.9) | 84.5 (13.7) | 44 (7.6) | 10 (0) | 11 (0) | 36 (2.4) |

| | **Model 3** $d = 2$ | **Model 4** $d = 3$ | **Model 5** $d = 8$ | **Model 6** $d = 2$ | **Model 7** $d = 8$ |
|---|---|---|---|---|---|
| Kolmogorov | | | | | |
| $G = 3$ | 2 (0) | 6 (0.8) | 207.5 (27.1) | 2 (0) | |
| $G = 4$ | 2 (0) | 5 (0.4) | 54.5 (7.2) | | 15 (0.4) |
| $G = 5$ | 2 (0) | 5 (0.4) | 32 (3.0) | | |
| $G = 6$ | 2 (0) | 7 (0.7) | 25 (1.3) | | |
| Fused | 2 (0) | 3 (0) | 16 (0.9) | | |
| SIS | 439.5 (38.3) | 3177 (95.9) | 4094 (81.0) | 13 (1.7) | 4661.5 (25.6) |
| DCS | 260.5 (36.2) | 40.5 (6.5) | 22 (2.7) | 1002 (89.2) | 1038 (121.2) |
| RCS | 2 (0) | 3 (0) | 3430 (124.4) | | |
| NIS | 494 (96.4) | 3258.5 (114.5) | 4260.5 (55.3) | | |
| ELS | 3247.5 (94.7) | 3801 (69.1) | 4510 (26.6) | 3253 (96.2) | |
| QA | | | | | |
| $\tau = 0.5$ | 50 (2.3) | 17 (1.7) | 1193 (129.4) | | |
| $\tau = 0.75$ | 70 (3.7) | 1234.5 (75.4) | 32.5 (1.4) | | |

variable transformation as in Models 1(b)–1(c) and Models 2(b)–2(c) can easily destroy their performance. Models 3 and 4 are nonlinear with heavy-tailed covariates. Most screening methods other than the fused Kolmogorov filter have too many false discoveries, especially in Model 4. NIS, RCS and QA are not directly

applicable when we have a Poisson regression model in Model 6. Model 5 has heteroscedasticity, which impairs SIS, NIS, RCS and ELS.

**5. A real data example.** In this section, we demonstrate the fused Kolmogorov filter on the Tecator dataset. The Tecator dataset was collected by Tecator Infratec Food and Feed Analyzer working in the wavelength range 850–1050 nm by the Near Infrared Transmission (NIT) principle. The predictors are 100 channel spectrum of absorbances. The response is the percentage of fat in finely chopped meat. This dataset is available at http://lib.stat.cmu.edu/datasets/tecator. The provider of the dataset suggested using the first 215 samples to test the performance of a statistical method by treating 43 of them as the testing set. However, samples #103 and #105 appear to be outliers, so we deleted them. Then we standardized the response so that it has a standard deviation of 1. We randomly chose 41 samples as our testing set in each replicate. Also, in addition to the 100 predictors in the original dataset, we added 4900 independent noise variables following the Cauchy distribution.

We include the fused Kolmogorov filter, DCS, SIS, QA, NIS and ELS for comparison. First, we examine whether the screening methods can distinguish the useful predictors from the noise variables. In the fused Kolmogorov filter, we still consider the combination of $G_i = 3, \ldots, 6$, as in the simulation studies. For each screening method we keep the top 100 predictors, as the "truth" is there are 4900 pure noise variables. We report the number of the original 100 predictors captured by screening in Table 2. It is easy to see that the fused Kolmogorov filter, DCS and NIS have much better performance in preserving the true predictors. In particular, the fused Kolmogorov filter has a nearly perfect screening result.

We further examine how variable screening helps predict the response variable. Again, we start with the augmented dataset with the additional 4900 pure noise variables. For a nonparametric model-free method such as the fused Kolmogorov filter and DCS, the prediction is made by fitting a random forest after screening. Hence the resulting methods are called K-RF and DCS-RF, respectively. NIS is designed based on a generalized additive model. So when NIS is used for variable

TABLE 2

*Comparison of the screening methods on the tecator dataset. We report the number of true predictors that are preserved after the screening step. The numbers are averaged over* 100 *replicates. Standard errors are in parentheses*

| | | | | | QA | | |
|---|---|---|---|---|---|---|---|
| | **Kolmogorov** | **DCS** | **NIS** | **SIS** | $\boldsymbol{\alpha = 0.5}$ | $\boldsymbol{\alpha = 0.75}$ | **ELS** |
| True predictors | 99.6 (0.06) | 75.4 (0.44) | 77.3 (0.28) | 11.7 (0.27) | 45.4 (0.56) | 42.2 (0.43) | 6.24 (0.14) |

TABLE 3

*Comparison of the prediction performance on the tecator dataset. The numbers are averaged over* 100 *replicates. Standard errors are in parentheses. A paired t-test shows that K-RF is significantly better than DCS-RF and NIS-RF, with p-values less than* $1 \times 10^{-5}$

| | **K-RF** | **DCS-RF** | **NIS–GroupLasso** | **INIS–GroupLasso** | **NIS-RF** |
|---|---|---|---|---|---|
| Average MSE | 0.097 (0.009) | 0.102 (0.010) | 0.195 (0.019) | 0.187 (0.017) | 0.103 (0.010) |

screening, the prediction is made by a sparse generalized additive model. We denote this method by NIS–GroupLasso. In K-RF, DCS-RF and NIS–GroupLasso, we let $d_n = 100$.

Moreover, we include an iterative procedure that performs NIS and group-lasso penalized regression repeatedly. After the initial screening, we keep the top 100 predictors, and then we follow Fan, Feng and Song (2011) to iteratively conduct the following two-step procedure: first, we add the predictor with the most predictive power that is not in the selected set of predictors; second, we delete some predictors in the selected set of predictors via group-lasso. In the deletion step, the tuning parameter is chosen to be the largest tuning parameter that produces an error within one standard deviation of the minimum error. This resulting method is referred to as INIS–GroupLasso. We use the `R` package `gglasso` [Yang and Zou (2015)] to fit the group-lasso penalized additive model.

Finally, as suggested by a referee, we also include the prediction performance for NIS followed by random forest, which is denoted by NIS-RF. The average mean squared errors (MSE) on the testing sets are listed in Table 3. The method K-RF has significantly better performance than all the other methods.

**6. Discussion.** In this paper we have proposed the fused Kolmogorov filter and demonstrated its superior performance over the existing screening methods. Before concluding this work, we would like to further comment on two main messages delivered in this paper. First, we have proposed the slicing and fusion idea to deal with general response variables such as continuous response variable and counts (e.g., Poisson) response variable. In this general approach one may use a different test statistic for testing the equivalence of two distributions to replace the Kolmogorov–Smirnov statistic, and the resulting screening method would be different and likely effective as well. We prefer the Kolmogorov–Smirnov statistic because it is invariant under variable transformation and works naturally with many different types of covariates. Moreover, its sure screening property can be established without assuming any special distributional property of the covariates. Any future proposal for variable screening should possess all these nice properties of the fused Kolmogorov filter and some nontrivial new properties. The second message is about nonparametric screening versus model-based screening. The vibrant

research on variable screening started with a simple model-based method, marginal correlations screening. However, it is clear now that nonparametric model-free variable screening should be preferred in real data analysis, unless the user strongly believes that the data can be fit well by a parametric model. Otherwise, nonparametric screening methods are more robust, have wider applicability and when combined with nonparametric learning techniques, they can provide better prediction than a model-based method. On the other hand, an obvious advantage of model-based screening is that its performance can be boosted by an iterative screening and model-fitting procedure. It is unclear how to derive a similar iterative procedure for a nonparametric model-free screening method. It would be interesting and useful to do so, such that we could have an iterative way to combine the fused Kolmogorov filter or other nonparametric screening method and nonparametric learning methods. This is an open question left for future study. We do not expect an easy solution. Note that even for the model-based iterative screening methods, their theoretical properties still remain unknown.

## APPENDIX: TECHNICAL PROOFS

Throughout this appendix, $F$ denotes the generic cumulative distribution function, and $f$ denotes the generic probability density function for a random variable.

PROPOSITION 1. *Consider a pair of random variables* $(X, Y)$. *For any interval* $[a, b)$ *such that* $f_Y(y) > 0$ *for* $y \in [a, b)$, *we have*

$$\inf_{y\in[a,b)} F(x \mid Y = y) \le F\big(x \mid Y \in [a, b)\big) \le F\big(x \mid Y \in [a, b)\big) \le \sup_{y\in[a,b)} F(x \mid Y = y)$$

*for all* $x$.

PROOF OF PROPOSITION 1. By definition,

$$F\big(x \mid Y \in [a, b)\big) = \frac{\int_a^b \int_{-\infty}^x f(x, y)\,\mathrm{d}x\,\mathrm{d}y}{\int_a^b f_Y(y)\,\mathrm{d}y} = \frac{\int_a^b \int_{-\infty}^x f(x \mid y) f_Y(y)\,\mathrm{d}x\,\mathrm{d}y}{\int_a^b f_Y(y)\,\mathrm{d}y}.$$

Because for any $y \in [a, b)$,

$$\inf_{y\in[a,b)} F(x \mid Y = y) \le \int_{-\infty}^x f(x \mid y)\,\mathrm{d}x \le \sup_{y\in[a,b)} F(x \mid Y = y),$$

we have the desired conclusion. □

PROOF OF LEMMA 1. We start with the first conclusion. If $X_j$ is independent of $Y$, then $X_j$ will be independent of any $H$, which is a function of $Y$. Therefore, $K_j^{\mathbf{G}} = 0$ for all $\mathbf{G}$. Now suppose $K_j^{\mathbf{G}} = 0$ for all choices of $\mathbf{G}$. For any $y$, consider $H = 1$ if $Y \le y$ and $H = 2$ otherwise. Because $K_j^{\mathbf{G}} = 0$, $X_j$ is independent of $H$. Consequently, $\Pr(Y \le y \mid X_j) = \Pr(Y \le y)$ for all $y$, and $Y$ is independent of $X_j$.

For the second conclusion, suppose there exists $\mathbf{G}$ such that $K_j^{\mathbf{G}} = 0$. Then $X_j \perp H$ for the corresponding $H$. Therefore, $\Pr(Y \le a_1 \mid X_j) = \Pr(H = 1 \mid X_j) = \Pr(H = 1)$ is a constant, which contradicts our assumption. Therefore, we must have $K_j^{\mathbf{G}} \ne 0$.

Now we turn to the third conclusion. Because $X_j$ is not independent of $Y$, $K_j^* > 0$. Hence, it suffices to show that $K^{\mathbf{G}} \to K_j^*$ as $G \to \infty$. This is indeed true. By the definition of $K_j^*$, for any $\varepsilon > 0$, there exists $(y_1^*, y_2^*, x^*)$ such that

$$\left|K_j^* - \left|F_j(x^* \mid y_1^*) - F_j(x^* \mid y_2^*)\right|\right| < \varepsilon.$$

Because $F(x^* \mid y)$ is continuous in $y$, there exists $\delta > 0$ such that $|F_j(x^* \mid y) - F_j(x^* \mid y_1^*)| < \varepsilon$ for any $|y - y_1^*| < \delta$. Take $\phi = \Pr(|y - y_1^*| < \delta)$. Because $\max_{l=1,\ldots,G} \Pr(H = l) \to 0$, there exists $G^*$ such that $\Pr(H = l) < \frac{\phi}{2}$ for $G > G^*$. In such cases, there exists $[a_{l_1}, b_{l_1}) \subset (y_1^* - \delta, y_1^* + \delta)$. By Proposition 1, we have

$$\left|F_j(x^* \mid H = l_1) - F_j(x^* \mid y_1^*)\right| < \varepsilon.$$

Similarly, for sufficiently large $G$, there exists $l_2$ such that

$$\left|F_j(x^* \mid H = l_2) - F_j(x^* \mid y_2^*)\right| < \varepsilon.$$

Now note that

$$\left|F_j(x^* \mid H = l_1) - F_j(x^* \mid H = l_2)\right| \le K_j^{\mathbf{G}} \le K_j^*.$$

Hence

$$\begin{aligned}
&|K_j^* - K_j^{\mathbf{G}}| \\
&\quad \le \left|F_j(x^* \mid y_1^*) - F_j(x^* \mid y_2^*)\right| + \varepsilon - \left|F_j(x^* \mid H = l_1) - F_j(x^* \mid H = l_2)\right| \\
&\quad \le \sum_{i=1,2} \left|F_j(x^* \mid y_i^*) - F_j(x^* \mid H = l_i)\right| + \varepsilon \\
&\quad < 3\varepsilon.
\end{aligned}$$

Therefore, the conclusion follows. □

PROOF OF LEMMA 2. Because $K_j^*$ and $K_j^{\mathbf{G}}$ are invariant under monotone transformations, it suffices to consider the case $g_1(t) = t$, $g_2(t) = t$, and hence $X_j$ and $Y$ are jointly normal. Let $f_y(y)$ be the probability density function of $Y$, which is standard normal. For the first conclusion, note that if $\rho_j = 0$, then $X_j$ is independent of $Y$ and $K_j^* = 0$. On the other hand, if $\rho_j \ne 0$, $X_j \mid$

$Y = y \sim N(\rho_j y, (1-\rho_j^2))$. Therefore, $F_j(x \mid y) = \Phi(\frac{x-\rho_j y}{\sqrt{1-\rho_j^2}})$. It follows that $K_j^* \geq \lim_{y\to-\infty} F_j(0 \mid y) - \lim_{y\to\infty} F_j(0 \mid y) = 1$. Meanwhile, by definition, $K_j^* \leq 1$. Therefore, $K_j^* = 1$.

For the second conclusion, again by $X_j \mid Y \sim N(\rho_j Y, (1-\rho_j^2))$ and $F_j(x \mid y) = \Phi(\frac{x-\rho_j y}{\sqrt{1-\rho_j^2}})$, we have

$$\begin{aligned} F_j(x \mid H = l) &= \frac{\Pr(X_j \leq x, H = l)}{\Pr(H = l)} \\ &= G\int_{a_{l-1}}^{a_l} \Phi\biggl(\frac{x-\rho_j y}{\sqrt{1-\rho_j^2}}\biggr) f(y)\,\mathrm{d}y \\ &\in \biggl[\Phi\biggl(\frac{x-\rho_j a_{l-1}}{\sqrt{1-\rho_j^2}}\biggr), \Phi\biggl(\frac{x-\rho_j a_l}{\sqrt{1-\rho_j^2}}\biggr)\biggr]. \end{aligned}$$

Now, for $1 \leq l < m \leq G$,

$$\begin{aligned} &\sup_x \bigl|F_j(x \mid H = l) - F_j(x \mid H = m)\bigr| \\ &\quad\leq \sup_x \biggl(\Phi\biggl(\frac{x-\rho_j a_{l-1}}{\sqrt{1-\rho_j^2}}\biggr) - \Phi\biggl(\frac{x-\rho_j a_m}{\sqrt{1-\rho_j^2}}\biggr)\biggr) \\ &\quad= 2\Phi\biggl(\frac{\rho_j(a_m - a_{l-1})}{\sqrt{1-\rho_j^2}}\biggr) - 1. \end{aligned}$$

On the other hand,

$$\begin{aligned} &\sup_x \bigl|F_j(x \mid H = 1) - F_j(x \mid H = G)\bigr| \\ &\quad\geq \sup_x \biggl(\Phi\biggl(\frac{x-\rho_j a_1}{\sqrt{1-\rho_j^2}}\biggr) - \Phi\biggl(\frac{x-\rho_j a_{G-1}}{\sqrt{1-\rho_j^2}}\biggr)\biggr) \\ &\quad= 2\Phi\biggl(\frac{\rho_j(a_1 - a_{G-1})}{\sqrt{1-\rho_j^2}}\biggr) - 1 \geq 2\Phi\biggl(\frac{\rho_j(a_m - a_{l-1})}{\sqrt{1-\rho_j^2}}\biggr) - 1 \\ &\quad\geq \sup_x \bigl|F_j(x \mid H = l) - F_j(x \mid H = m)\bigr|. \end{aligned}$$

Therefore,

$$K_j^{\mathbf{G}} = \sup_x \bigl|F_j(x \mid H = 1) - F_j(x \mid H = G)\bigr|.$$

Moreover, note that $a_1 = -a_{G-1}$. By checking the derivatives, we have

$$K_j^{\mathbf{G}} = \bigl|F_j(0 \mid H = 1) - F_j(0 \mid H = G)\bigr|.$$

Hence

$$
\begin{aligned}
K_j^{\mathbf{G}} &= G\Bigg(\int_{-\infty}^{a_1} \Phi\Bigg(\frac{-\rho_j y}{\sqrt{1-\rho_j^2}}\Bigg) f(y)\,\mathrm{d}y - \int_{a_{G-1}}^{\infty} \Phi\Bigg(\frac{-\rho_j y}{\sqrt{1-\rho_j^2}}\Bigg) f(y)\,\mathrm{d}y\Bigg) \\
&= G\Bigg(\int_{-\infty}^{a_1} \Phi\Bigg(\frac{-\rho_j y}{\sqrt{1-\rho_j^2}}\Bigg) f(y)\,\mathrm{d}y - \int_{-\infty}^{a_1} \Bigg(1 - \Phi\Bigg(\frac{-\rho_j y}{\sqrt{1-\rho_j^2}}\Bigg)\Bigg) f(y)\,\mathrm{d}y\Bigg) \\
&= G\Bigg(\int_{-\infty}^{a_1} \Bigg(2\Phi\Bigg(\frac{-\rho_j y}{\sqrt{1-\rho_j^2}}\Bigg) - 1\Bigg) f(y)\,\mathrm{d}y\Bigg).
\end{aligned}
$$

Because $a_1 \le 0$, $\Phi(\frac{-\rho_j y}{\sqrt{1-\rho_j^2}})$ is strictly increasing in $\rho_j$ for each $y \in (-\infty, a_1)$. Hence $K_j^{\mathbf{G}}$ is strictly increasing in $\rho_j$. □

Now we prove Theorem 1. In order to prove this theorem, we need the following lemmas.

LEMMA 4. *If $\hat{a}_l$ is the sample $\frac{l}{G}$th quantile for $Y$, then with a probability greater than $1 - C\exp(-C\frac{n}{G^2})$, we have*

$$
\Pr(\hat{a}_l \le Y < \hat{a}_{l+1}) < \frac{2}{G}. \tag{9}
$$

LEMMA 5. *Under the conditions in Theorem 1, for any $\varepsilon > 0$, we have*:

(1)

$$
\begin{aligned}
&\Pr\big(|\hat{K}_j^{(o)} - K_j^{(o)}| \ge N\varepsilon\big) \\
&\qquad \le CN\big(\log^2 n\big)\exp\Bigg(-C\frac{n\varepsilon^2}{\log n}\Bigg) + CN\big(\log^2 n\big)\exp\Bigg(-C\frac{n}{\log^2 n}\Bigg);
\end{aligned} \tag{10}
$$

(2)

$$
\Pr\big(|\hat{K}_j - K_j| \ge N\varepsilon\big) \le CN\big(\log^2 n\big)\exp\Bigg(-C\frac{n\varepsilon^2}{\log n}\Bigg). \tag{11}
$$

LEMMA 6. *Under the conditions in Theorem 1, we have*

$$
\Pr\big(|K_j - K_j^{(o)}| \ge N\Delta_{\mathbf{S}}/4\big) \le CN\exp\Bigg(-C\frac{n}{\log^2 n}\Bigg).
$$

With Lemmas 4–6, we are ready to prove Theorem 1.

PROOF OF THEOREM 1. We first consider the first conclusion. Note that if $|\hat{K}_j^{(o)} - K_j^{(o)}| < N\Delta_{\mathbf{S}}/4$ for all $j$, we must have $\mathbf{D} \subset \hat{\mathbf{D}}$. This is indeed true because, combining it with condition (C1), we have

$$\hat{K}_j^{(o)} > K_j^{(o)} - N\Delta_{\mathbf{S}}/4 \geq \max_{j \notin \mathbf{S}} K_j^{(o)} + N\Delta/4 \qquad \text{for } j \in \mathbf{S},$$

$$\hat{K}_j^{(o)} < K_j^{(o)} + N\Delta_{\mathbf{S}}/4 \leq \max_{j \notin \mathbf{S}} K_j^{(o)} + N\Delta/4 \qquad \text{for } j \notin \mathbf{S}.$$

Hence, $\mathbf{S} \subset \hat{\mathbf{D}}$ and $\mathbf{D} \subset \hat{\mathbf{D}}$.

By (10), we have the desired conclusion.

For the second conclusion, we again have that, if $|\hat{K}_j - K_j^{(o)}| < N\Delta_{\mathbf{S}}/4$ for all $j$, we must have $\mathbf{D} \subset \hat{\mathbf{D}}$.

Combining (11) and Lemma 6, we have

$$\begin{aligned}&\Pr\bigl(|\hat{K}_j - K_j^{(o)}| > N\Delta_{\mathbf{S}}/4\bigr)\\ &\quad\leq CN\exp\biggl(-C\frac{n}{\log^2 n}\biggr) + CN\bigl(\log^2 n\bigr)\exp\biggl(-C\frac{n\Delta_{\mathbf{S}}^2}{\log n}\biggr).\end{aligned}$$

Then we have the desired conclusion. □

PROOF OF LEMMA 4. First, we show that, under the event $A = \sup_y |\hat{F}_y(y) - F_y(y)| \leq \frac{1}{8G}$, we must have (9). Indeed, under event $A$,

$$\begin{aligned}&\Pr(\hat{a}_l \leq Y < \hat{a}_{l+1})\\ &\quad= \Pr\biggl(\frac{l}{G} \leq \hat{F}_y(Y) < \frac{l+1}{G}\biggr)\\ &\quad\leq \Pr\biggl(\frac{l}{G} - \frac{1}{8G} \leq F_y(Y) < \frac{l+1}{G} + \frac{1}{8G}\biggr) = \frac{5}{4G} < \frac{2}{G}.\end{aligned}$$

Then note $\Pr(A) \geq 1 - C\exp(-C\frac{n}{G^2})$ by the Dvoretzky–Kiefer–Wolfowitz inequality, and the conclusion follows. □

PROOF OF LEMMA 5. We first show (10). Consider a single partition $\mathbf{G}_i$ with $G_i$ intervals bounded by the theoretical quantiles. Then $H_i^{(o)} = g$ if and only if $Y$ is between its $\frac{g}{G_i}$th and $\frac{g+1}{G_i}$th quantile. Set $K^{(o)}(G_i; g, g') = \sup_x |F_j(x \mid H_i^{(o)} = g) - F_j(x \mid H_i^{(o)} = g')|$. Then $\Pr(H_i^{(o)} = g) = \Pr(H_i^{(o)} = g') = \frac{1}{G_i}$. By Lemma A1 in Mai and Zou (2013), we have

$$\begin{aligned}&\Pr\bigl(|\hat{K}_j^{(o)}(G_i; g, g') - K_j^{(o)}(G_i; g, g')| \geq \varepsilon\bigr)\\ &\quad\leq C\exp\biggl(-Cn\frac{\varepsilon^2}{G_i}\biggr) + C\exp\biggl(-C\frac{n}{G_i^2}\biggr).\end{aligned}$$

Then if $|\hat{K}_j^{(o)}(G_i; g, g') - K_j^{(o)}(G_i; g, g')| \le \varepsilon$ for all $g, g'$, we must have

$$\begin{aligned}|\hat{K}_j^{(o)} - K_j^{(o)}| &= \Big|\max_{g,g'} \hat{K}_j^{(o)}(G_i; g, g') - \max_{g,g'} K_j^{(o)}(G_i; g, g')\Big| \\ &\le \max_{g,g'} |\hat{K}_j^{(o)}(G_i; g, g') - K_j^{(o)}(G_i; g, g')| \le \varepsilon.\end{aligned}$$

Therefore,

$$\begin{aligned}&\Pr\big(|\hat{K}_j^{(o)}(G_i) - K_j^{(o)}(G_i)| > \varepsilon\big) \\ &\quad\le C G_i^2 \exp\Big(-Cn\frac{\varepsilon^2}{G_i}\Big) + C G_i^2 \exp\Big(-C\frac{n}{G_i^2}\Big) \\ &\quad\le C(\log^2 n)\exp\Big(-Cn\frac{\varepsilon^2}{\log n}\Big) + C(\log^2 n)\exp\Big(-C\frac{n}{\log^2 n}\Big).\end{aligned}$$

Finally, note that

$$\Pr\big(|\hat{K}_j^{(o)}(G_i) - K_j^{(o)}(G_i)| > N\varepsilon\big) \le \sum_i \Pr\big(|\hat{K}_j^{(o)}(G_i) - K_j^{(o)}(G_i)| > \varepsilon\big),$$

and the conclusion follows. For (11), redefine $H_i = l$ if $Y$ is with in the $\frac{l}{G_i}$th and $\frac{l+1}{G_i}$th sample quantiles. Note that

$$\begin{aligned}&\Pr\big(|\hat{K}_j(G_i; g, g') - \hat{K}_j(G_i; g, g')| \ge \varepsilon\big) \\ &\quad\le \sum_{l=g,g'} \Pr\Big(\sup_x |\hat{F}_j(x \mid H_i = l) - F_j(x \mid H_i = l)| \ge \varepsilon/2\Big) \\ &\quad\le C\exp\Big(-Cn\frac{\varepsilon^2}{G_i}\Big),\end{aligned}$$

where the last inequality follows from the Dvoretzky–Kiefer–Wolfowitz inequality and the fact that there are $\frac{n}{G_i}$ observations in the $g$th and $g'$th slice, respectively. Then because $G_i \le \lceil \log n \rceil$, we have the desired conclusion. Finally, (11) can be proven in a similar way to (10). □

PROOF OF LEMMA 6. First, note that

$$\Pr\big(|K_j - K_j^{(o)}| \ge N\Delta_{\mathbf{S}}/4\big) \le \sum_i \Pr\big(|K_j(G_i) - K_j^{(o)}(G_i)| \ge \Delta_{\mathbf{S}}/4\big). \tag{12}$$

Therefore, we establish a bound for $\Pr(|K_j(G_i) - K_j^{(o)}(G_i)| \ge \Delta_{\mathbf{S}}/4)$.

Define

$$K_{0j} = \sup_x \Big(\sup_y F(x \mid y) - \inf_y F(x \mid y)\Big).$$

For any $x$ and $l$, we have

$$\inf_y F_j(x \mid y) \le F_j(x \mid H = l) \le \sup_y F_j(x \mid y).$$

It follows that $K_j(G_i) \le K_{0j}$ and $K_j^{(o)}(G_i) \le K_{0j}$. Moreover, for any $\varepsilon > 0$, there exists $(x^*, y_1^*, y_2^*)$ such that

$$K_{0j} \le F_j\big(x^* \mid y_1^*\big) - F_j\big(x^* \mid y_2^*\big) + \varepsilon.$$

Then there exists $[a_{l_i}, a_{l_i+1}) \in \mathbf{G}$ such that $y_i^* \in [a_{l_i}, a_{l_i+1})$. Hence,

$$K_{0j} - K_j^{(o)}(G_i) \le \varepsilon + \sum_{i=1,2} \big|F_j\big(x^* \mid y_1^*\big) - F_j\big(x^* \mid H = l_i\big)\big| \le \varepsilon + \Delta_{\mathbf{S}}/8,$$

where the last inequality follows from condition (C2) and Proposition 1. Because $\varepsilon$ is arbitrary, we have $K_{0j} - K_j^{(o)}(G_i) \le \Delta_{\mathbf{S}}/8$ and hence $K_j \le K_j^{(o)}(G_i) + \Delta_{\mathbf{S}}/8$. On the other hand, suppose

$$K_j^{(o)}(G_i) = F_j\big(x_0 \mid H_i^{(o)} = l_1\big) - F_j\big(x_0 \mid H_i^{(o)} = l_2\big).$$

Set $y_1^*$ such that $y_1^* \in \{y : H_i^{(o)} = l_1\}$ and $\inf_{y:\, H_i^{(o)}=l_1} F_j(x \mid y) = F_j(x \mid y_1^*)$. Note that $y_1^*$ can be $+\infty$ or $-\infty$. Then there exists $l_1'$ such that $y_1^* \in \{H_i^{(o)} = l_1\} \cap \{H_i = l_1'\}$. Also define $y_2^*$ as the number that $y_2^* \in \{y : H_i^{(o)} = l_2\}$ and $\sup_{y:\, H_i^{(o)}=l_1} F_j(x \mid y) = F_j(x \mid y_2^*)$. Note that $y_2^*$ can be $+\infty$ or $-\infty$ as well. Then there exists $l_2'$ such that $y_2^* \in \{H_i^{(o)} = l_2\} \cap \{H_i = l_2'\}$.

We claim that if $\Pr(H_i = l_k') \le 2/G$, we must have $K_j(G_i) \ge K_j^{(o)}(G_i) - \Delta_{\mathbf{S}}/4$. Indeed, by Proposition 1,

$$K_j \ge \inf_{y:\, H_i = l_1'} F_j(x_0 \mid y) - \sup_{y:\, H_i = l_2'} F_j(x_0 \mid y).$$

Then by condition (C2), if $\Pr(H_i = l_k') \le 2/G$, we must have

$$\begin{aligned} K_j &\ge \inf_{y:\, H_i = l_1'} F_j(x_0 \mid y) - \sup_{y:\, H_i = l_2'} F_j(x_0 \mid y) \\ &\ge F_j\big(x_0 \mid y_1^*\big) - F_j\big(x_0 \mid y_2^*\big) \\ &\ge \inf_{y:\, H_i^{(o)} = l_1} F_j\big(x^* \mid y\big) - \Delta_{\mathbf{S}}/8 - \sup_{y:\, H_i^{(o)} = l_2} F_j\big(x^* \mid y\big) - \Delta_{\mathbf{S}}/8 \\ &\ge K_j^{(o)} - \Delta_{\mathbf{S}}/4, \end{aligned}$$

where the last inequality again follows from condition (C2) and Proposition 1.

By Lemma 4, we have

$$\Pr\big(\Pr\big(H_i = l_k'\big) > 2/G\big) \le C \exp\left(-C \frac{n}{G_i^2}\right).$$

Therefore,

$$\Pr\bigl(\bigl|K_j(G_i) - K_j^{(o)}(G_i)\bigr| \geq \Delta_{\mathbf{S}}/4\bigr) \leq C\exp\biggl(-C\frac{n}{G_i^2}\biggr) \leq C\exp\biggl(-C\frac{n}{\log^2 n}\biggr). \tag{13}$$

Combining (12) and (13) we have the desired conclusion. □

PROOF OF LEMMA 3. For the first conclusion, note that

$$\begin{pmatrix} T_y(Y) \\ \mathbf{T}(\mathbf{X}) \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma^2 & \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma} \\ \boldsymbol{\Sigma}\boldsymbol{\beta} & \boldsymbol{\Sigma} \end{pmatrix}\right).$$

Straightforward calculation shows that

$$\bigl|\operatorname{cor}\bigl(T_y(Y), T_j(X_j)\bigr)\bigr| = \frac{|\alpha_j|}{\sqrt{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta} + \sigma^2}}$$

is monotone in $|\alpha_j|$. Now that, for any $G_i$, $K_j^{(o)}(G_i)$ is invariant under strictly monotone transformations. Therefore, by the second conclusion in Lemma 2, $K_j^{(o)}(G_i)$ is strictly increasing in $|\alpha_j|$, and the conclusion follows.

For the second conclusion, note that when $\boldsymbol{\Sigma}$ is blockwise independent, we must have $\boldsymbol{\alpha}_{\mathbf{D}^C} = 0$.

For the third conclusion, note that for $j > d$, we have $\alpha_j = \rho^{j-d}\alpha_d$. When $j > d + \frac{\log \min_{j\in\mathbf{D}} |\alpha_j|/|\alpha_d|}{\log|\rho|}$, we must have $|\alpha_j| < \min_{j\in\mathbf{D}} |\alpha_j|$, and the conclusion follows.

For the third conclusion, write $\boldsymbol{\Sigma} = (1-\rho)\mathbf{I} + \rho\mathbf{J}$, where $\mathbf{J}$ is a $p \times p$ matrix of 1. Then $\boldsymbol{\Sigma}^{-1} = (1-\rho)^{-1}I - \rho[\{1+(p-1)\rho\}(1-\rho)]^{-1}\mathbf{J}$. Write $c = 1^{\mathrm{T}}\boldsymbol{\beta} = \sum_{j\in\mathbf{S}}\beta_j$. For any $j \in \mathbf{S}$, we have $\beta_j = -\rho[\{1+(p-1)\rho\}(1-\rho)]^{-1}c$. Thus $\mathbf{D} \subseteq \mathbf{S} \Leftrightarrow 1^{\mathrm{T}}\boldsymbol{\beta} = 0$.

For the fourth conclusion, note that for any $j \in \mathbf{D}$, we have $\alpha_j = (1-\rho)\beta_j + \rho 1^{\mathrm{T}}\boldsymbol{\beta}$, while for $j \notin \mathbf{D}$, we have $\alpha_j = \rho 1^{\mathrm{T}}\boldsymbol{\beta}$. Hence, when $\rho > 0$ and $\beta_j$ has the same sign for all $j \in \mathbf{D}$, we have $\Delta_{\mathbf{D}} > 0$. □

**Acknowledgments.** We are grateful to the Editor, the Associate Editor and two referees for helpful suggestions. We thank Professor Lan Wang for providing the implementation of QA and Professor Yichao Wu for providing the implementation of ELS.

## REFERENCES

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–619. MR0803258

CHANG, J., TANG, C. Y. and WU, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *Ann. Statist.* **41** 2123–2148. MR3127860

COOK, R. D. and WEISBERG, S. (1991). Comment on "Sliced inverse regression for dimension reduction." *J. Amer. Statist. Assoc.* **86** 328–332.

COOK, R. D. and ZHANG, X. (2014). Fused estimators of the central subspace in sufficient dimension reduction. *J. Amer. Statist. Assoc.* **109** 815–827. MR3223752

FAN, J. and FAN, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.* **36** 2605–2637. MR2485009

FAN, J., FENG, Y. and SONG, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *J. Amer. Statist. Assoc.* **106** 544–557. MR2847969

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and LV, J. (2008). Sure independence screening for ultra-high dimensional feature space. *J. Roy. Statist. Soc. Ser. B* **20** 101–148.

FAN, J., SAMWORTH, R. and WU, Y. (2009). Ultra-dimensional variable selection via independenct learning: Beyond the linear model. *J. Mach. Learn. Res.* **10** 1829–1853.

FAN, J. and SONG, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38** 3567–3604. MR2766861

HASTIE, T. and TIBSHIRANI, R. (1998). Classification by pairwise coupling. *Ann. Statist.* **26** 451–471. MR1626055

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294

HE, X. and SHEN, L. (1997). Linear regression after spline transformation. *Biometrika* **84** 474–481. MR1467062

HE, X., WANG, L. and HONG, H. G. (2013). Quantile–adaptive model-free variable screening for high-dimensional heterogeneous data. *Ann. Statist.* **41** 342–369. MR3059421

HSING, T. and CARROLL, R. J. (1992). An asymptotic theory for sliced inverse regression. *Ann. Statist.* **20** 1040–1061. MR1165605

HUANG, J., HOROWITZ, J. L. and MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36** 587–613. MR2396808

LI, K.-C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–342. MR1137117

LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. MR3010900

LI, G., PENG, H., ZHANG, J. and ZHU, L. (2012). Robust rank correlation based screening. *Ann. Statist.* **40** 1846–1877. MR3015046

LV, J. and FAN, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37** 3498–3528. MR2549567

MAI, Q. and ZOU, H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika* **100** 229–234. MR3034336

SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VENABLES, W. N. and RIPLEY, B. D. (2002). *Modern Applied Statistics with S*, 4th ed. Springer, New York.

YANG, Y. and ZOU, H. (2015). A fast unified algorithm for solving group-lasso penalized learning problems. *Stat. Comput.* To appear. DOI:10.1007/s11222-014-9498-5.

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZHU, L. X. and NG, K. W. (1995). Asymptotics of sliced inverse regression. *Statist. Sinica* **5** 727–736. MR1347616

ZHU, L.-P., LI, L., LI, R. and ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **106** 1464–1475. MR2896849

DEPARTMENT OF STATISTICS
FLORIDA STATE UNIVERSITY
TALLAHASSEE, FLORIDA 32306
USA
E-MAIL: mai@stat.fsu.edu

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
MINNEAPOLIS, MINNESOTA 55455
USA
E-MAIL: zouxx019@umn.edu