IRTG 1792 Discussion Paper 2018-050

Variable selection and direction estimation for single-index models via DC-TGDR method

Wei Zhong *
Xi Liu *
Shuangge Ma *

* Xiamen University, People's Republic of China

This research was supported by the Deutsche Forschungsgemeinschaft through the International Research Training Group 1792 "High Dimensional Nonstationary Time Series".

http://irtg1792.hu-berlin.de
ISSN 2568-5619

Statistics and Its Interface Volume 11 (2018) 169–181

Variable selection and direction estimation for single-index models via DC-TGDR method

Wei Zhong, Xi Liu, and Shuangge Ma*

This paper is concerned with selecting important covariates and estimating the index direction simultaneously for high dimensional single-index models. We develop an efficient Threshold Gradient Directed Regularization method via maximizing Distance Covariance (DC-TGDR) between the single index and response variable. Due to the appealing property of distance covariance which can measure nonlinear dependence between random variables, the proposed method avoids estimating the unknown link function of the single index and dramatically reduces computational complexity compared to other methods that use smoothing techniques. It keeps the model-free advantage from the view of sufficient dimension reduction and requires neither predictors nor response variable to be continuous. In addition, the DC-TGDR method encourages a grouping effect. That is, it is capable of choosing highly correlated covariates in or out of the model together. We examine finite-sample performance of the proposed method by Monte Carlo simulations. In a real data analysis, we identify important copy number alterations (CNAs) for gene expression.

Keywords and phrases: Distance covariance, High-dimensional data, Threshold gradient directed regularization, Single-index models, Variable selection.

1. INTRODUCTION

With the development of modern technology for data collection, researchers are able to collect high dimensional data at relatively low cost in many fields. For example, RNA-seq technology is capable of profiling human tissues on a genome wide scale and measuring expression levels of thousands of genes along with certain clinical outcomes. With high dimensional data, nonlinear dependence between predictors and the response variable is often present. This makes traditional linear models not adequate. On the other hand, fully nonparametric models suffer from the "curse of dimensionality" problem. As a classic semiparametric method, single-index models which assume that the response only depends on the predictors through their single linear combination provide a balanced solution. They can not only maintain the flexibility of nonparametric models to deal with the nonlinear dependence but also retain the model interpretability of parametric models and avoid the "curse of dimensionality".

*Corresponding author.

Single-index models have been intensively studied in the literature, for instance, Powell, Stock and Stoker [16], Ichimura [7], Härdle, Hall and Ichimura [4], Horowitz and Härdle [5], Xia and Li [26] and among others. The traditional methods apply nonparametric smoothing techniques to estimate the unknown link function. For example, Ichimura [7] suggests replacing the unknown function with the leave-one-out Nadaraya-Watson estimator. To exclude irrelevant predictors and improve interpretability of high dimensional single-index models, penalized regression methods based on nonparametric smoothing techniques have also been developed, including Zhu and Zhu [31], Liang et al. [12], Radchenko [17], etc. To avoid estimating the unknown link function and reduce computational complexity, Zhu, Huang and Li [30] suggests directly using the simple linear quantile regression to estimate the index parameter vector for heteroscedastic single-index models and proves that the resulting index estimator is consistent under the linearity condition. Zhong et al. [29] further proposes penalized linear quantile regression to identify important covariates for high dimensional single-index models.

We consider a general class of single-index models

$$Y = g(\boldsymbol{\beta}^T X, \varepsilon), \tag{1}$$

where Y is a response variable, X is a p-dimensional predictor vector, $\boldsymbol{\beta}$ is an index parameter vector of interest and ε is a random error. Model (1), which is originally proposed in Li and Duan [10], implies that the response Y is independent of predictors X conditional on the index $\boldsymbol{\beta}^T X$. Many dimension reduction approaches are able to estimate the single-index direction in single-index models under the framework of central subspace. Examples include sliced inverse regression (SIR) by Li [9], minimum average variance estimation (MAVE) by Xia et al. [25], direction estimation by minimizing a Kullback-Leibler distance by Yin and Cook [27], etc. These methods either use inverse regression based on the linearity condition or involve nonparametric smoothing techniques. Recently, Sheng and Yin [22] suggests estimating single-index direction by maximizing distance covariance. Here, distance covariance proposed by Szekely, Rizzo and Bakirov [23] measures nonlinear dependence between random variables. Under regularity conditions, it is shown that the resulting estimator of single-index direction is root-n consistent and asymptotically normal.

In this paper, we develop an efficient Threshold Gradient Directed Regularization method via maximizing Dis-

tance Covariance (DC-TGDR) between the single index and response variable for high dimensional data. The TGDR method proposed by Friedman and Popescu [3] is an incremental stagewise parameter path searching method, which can keep coefficients of irrelevant variables as zero by imposing threshold on the updating direction. In the literature, the original TGDR method and its modified versions have shown good performance for variable selection problem in high-dimensional scenarios. For example, Ma and Huang [13] proposes a Clustering TGDR method for simultaneous cluster selection and within cluster gene selection.

Compared to the existing methods for single-index models, our proposed DC-TGDR method enjoys the following novel advantages. First, as a variant of the TGDR algorithm, it is capable of identifying important covariates efficiently and estimating the index direction simultaneously for high dimensional single-index models. Second, since distance covariance can measure nonlinear dependence between random variables, the DC-TGDR method avoids estimating the unknown link function of the single index, and so dramatically reduces computational complexity compared with other methods that use smoothing techniques. Third, the proposed method encourages a grouping effect. That is, it is capable of choosing highly correlated covariates in or out of the model together.

The rest of the article is organized as follows. In Section 2, we develop the DC-TGDR procedure to solve the coefficient paths in single-index models. Section 3 examines the finite sample performance of our proposed method along with alternative methods by intensive simulation studies. Section 4 implements the new method to analyze how copy number alternations (CNAs) regulate the expression level of certain gene in the Cell Development pathway. At last, concluding remarks are presented in Section 5.

2. METHODS

2.1 Preliminaries

Sheng and Yin [22] suggests estimating the index direction by maximizing distance covariance [23] in general single-index models. That is,

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \mathcal{V}^2(\boldsymbol{\beta}^T X, Y), \text{ subject to } \boldsymbol{\beta}^T \Sigma_X \boldsymbol{\beta} = 1, \tag{2}$$

where Σ_X stands for the nonsingular covariance matrix of X and $\mathcal{V}(\boldsymbol{\beta}^T X, Y)$ is the distance covariance between the single-index $\boldsymbol{\beta}^T X$ and response Y. Since the index parameter $\boldsymbol{\beta}$ is not identifiable, the direction of $\boldsymbol{\beta}$ rather than its true value is our primary interest to estimate. Note that $\boldsymbol{\beta}^T \Sigma_X \boldsymbol{\beta} = 1$ is the constraint to make the maximization procedure work. Some other constraints such as $\boldsymbol{\beta}^T \boldsymbol{\beta} = 1$ can be also used. Here, the distance covariance $\mathcal{V}(\boldsymbol{\beta}^T X, Y)$ is a new measure of the dependence between $\boldsymbol{\beta}^T X$ and Y, which is the non-negative square root of

$$\mathcal{V}^2(\boldsymbol{\beta}^T X, Y) \tag{3}$$

$$= \int_{\mathbb{R}^2} \left| \Phi_{\boldsymbol{\beta}^T X, Y}(t,s) - \Phi_{\boldsymbol{\beta}^T X}(t)\Phi_Y(s) \right|^2 w(t,s)\,dt\,ds,$$

where $\Phi(\cdot)$ denotes the characteristic function and $w(t,s)$ is a positive weight function. An appealing property of the distance covariance is that $\mathcal{V}(\boldsymbol{\beta}^T X, Y) = 0$ is equivalent to the independence between $\boldsymbol{\beta}^T X$ and Y. $(\mathbf{X}, \mathbf{Y}) = \{(X_i, Y_i) : i = 1, \ldots, n\}$ denotes an observed random sample from (X, Y). The empirical distance covariance between $\boldsymbol{\beta}^T X$ and Y is defined as the square root of

$$\mathcal{V}_n^2(\mathbf{X}\boldsymbol{\beta}, \mathbf{Y}) = T_1 + T_2 - 2T_3, \tag{4}$$

where

$$\begin{aligned} T_1 &= \frac{1}{n^2} \sum_{k,l=1}^{n} |\boldsymbol{\beta}^T X_k - \boldsymbol{\beta}^T X_l||Y_k - Y_l|, \\ T_2 &= \frac{1}{n^2} \sum_{k,l=1}^{n} |\boldsymbol{\beta}^T X_k - \boldsymbol{\beta}^T X_l| \frac{1}{n^2} \sum_{k,l=1}^{n} |Y_k - Y_l|, \\ T_3 &= \frac{1}{n^3} \sum_{k=1}^{n} \sum_{l,m=1}^{n} |\boldsymbol{\beta}^T X_k - \boldsymbol{\beta}^T X_l||Y_k - Y_m|. \end{aligned}$$

At the sample level, the index direction can be estimated by

$$\widehat{\boldsymbol{\beta}}_n = \arg\max_{\boldsymbol{\beta}} \mathcal{V}_n^2(\mathbf{X}\boldsymbol{\beta}, \mathbf{Y}), \text{ subject to } \boldsymbol{\beta}^T \Sigma_X \boldsymbol{\beta} = 1. \tag{5}$$

Sheng and Yin [22] applies the sequential quadratic programming procedure (SQP) to solve the above maximization problem and shows that the resulting direction estimator $\widehat{\boldsymbol{\beta}}_n$ is root-n consistent under some regularity conditions.

2.2 The DC-TGDR algorithm

In high dimensional applications, the SQP algorithm used in Sheng and Yin [22] can not perform variable selection and is computationally inefficient. Next, we develop an efficient Threshold Gradient Directed Regularization method via maximizing the empirical Distance Covariance between the single index $\mathbf{X}\boldsymbol{\beta}$ and response $\mathbf{Y}$ in (4). This algorithm is computationally fast and able to select important covariates and estimate the index direction simultaneously for high dimensional single-index models.

First, we compute the gradient of the objective function $\mathcal{V}_n^2(\mathbf{X}\boldsymbol{\beta}, \mathbf{Y})$ in (4) with respect to each coordinate β_j as follows.

$$g_j(\boldsymbol{\beta}) = \frac{\partial \mathcal{V}_n^2(\mathbf{X}\boldsymbol{\beta}, \mathbf{Y})}{\partial \beta_j} = \frac{\partial T_1}{\partial \beta_j} + \frac{\partial T_2}{\partial \beta_j} - 2\frac{\partial T_3}{\partial \beta_j}, \tag{6}$$

where

$$\begin{aligned} \frac{\partial T_1}{\partial \beta_j} &= \frac{1}{n^2} \sum_{k,l=1}^{n} \operatorname{sgn}(\boldsymbol{\beta}^T X_k - \boldsymbol{\beta}^T X_l)(X_{kj} - X_{kl})|Y_k - Y_l|, \\ \frac{\partial T_2}{\partial \beta_j} &= \frac{1}{n^2} \sum_{k,l=1}^{n} \operatorname{sgn}(\boldsymbol{\beta}^T X_k - \boldsymbol{\beta}^T X_l)(X_{kj} - X_{kl}) \end{aligned}$$

Algorithm 1 The DC-TGDR Method

Step 1. (Initialization) Let $m = 1$. Set $\widehat{\beta}_i^1 = 1$ and $\widehat{\beta}_j^1 = 0$ for $j \neq i$, where $i = \arg\max_{1\leq j\leq p} \mathcal{V}_n^2(\mathbf{x}_j, \mathbf{Y})$. Denote $\widehat{\boldsymbol{\beta}}^1 = (\widehat{\beta}_1^1, \widehat{\beta}_2^1, \ldots, \widehat{\beta}_p^1)^T$.

Step 2. (Update). Increase m by 1:

(1) Calculate the gradient of $\mathcal{V}_n^2(\mathbf{X}\boldsymbol{\beta}, \mathbf{Y})$ based on (6) at $\widehat{\boldsymbol{\beta}}^m$ and denote it as $\{g_j(\widehat{\boldsymbol{\beta}}^m)\}_1^p$;

(2) Compute the scaling factors $f_j(\widehat{\boldsymbol{\beta}}^m) = I[|g_j(\widehat{\boldsymbol{\beta}}^m)| \geq \tau \cdot \max_{0\leq k\leq p} |g_k(\widehat{\boldsymbol{\beta}}^m)|], j = 1, \ldots, p$;

(3) Update $\widehat{\boldsymbol{\beta}}$ by $\widehat{\boldsymbol{\beta}}^{m+1} = \widehat{\boldsymbol{\beta}}^m + \triangle\nu \cdot \mathbf{h}(\widehat{\boldsymbol{\beta}}^m)$, where $\mathbf{h}(\widehat{\boldsymbol{\beta}}^m) = \{f_j(\widehat{\boldsymbol{\beta}}^m) \cdot g_j(\widehat{\boldsymbol{\beta}}^m)\}_1^p$;

(4) Regulate the L_2 norm of $\widehat{\boldsymbol{\beta}}^{m+1}$, $\| \widehat{\boldsymbol{\beta}}^{m+1} \|_2$, to be 1.

Step 3. (Iteration). Repeat Step 2 until $|\mathcal{V}_n^2(\mathbf{X}\widehat{\boldsymbol{\beta}}^m, \mathbf{Y}) - \mathcal{V}_n^2(\mathbf{X}\widehat{\boldsymbol{\beta}}^{m+1}, \mathbf{Y})| < \delta$, where δ is a stopping rule. Or, Repeat Step 2 K times, where K is determined by cross-validation.

$$\times \frac{1}{n^2} \sum_{k,l=1}^{n} |Y_k - Y_l|,$$

$$\frac{\partial T_3}{\partial \beta_j} = \frac{1}{n^3} \sum_{k=1}^{n} \sum_{l,m=1}^{n} \operatorname{sgn}(\boldsymbol{\beta}^T X_k - \boldsymbol{\beta}^T X_l) \times (X_{kj} - X_{kl}) |Y_k - Y_m|.$$

Next, we present details of the DC-TGDR method in the following algorithm.

Note that in Step 1, we initiate the solution path by setting the coefficient corresponding to the predictor that marginally maximizes the distance covariance with Y to be 1, and others to be 0. This step essentially coincides with the sure independence screening using distance covariance in Li, Zhong and Zhu [11]. In Step 2(4), we regulate the L_2 norm of $\widehat{\boldsymbol{\beta}}^{m+1}$ due to the following reasons. For any constant c, $\mathcal{V}_n^2(c\boldsymbol{\beta}^T X, Y) = |c|\mathcal{V}_n^2(\boldsymbol{\beta}^T X, Y)$, thus we have to put a constraint on the norm of $\boldsymbol{\beta}$ to make the maximization procedure work. On the other hand, the index parameter $\boldsymbol{\beta}$ is not identifiable in the single-index model. The direction of $\boldsymbol{\beta}$, instead of its true value, is our primary interest. Thus, we take the L_2 constraint $\boldsymbol{\beta}^T\boldsymbol{\beta} = 1$ to obtain the direction of $\boldsymbol{\beta}$. The computational complexity to compute the empirical distance covariance is $O(n^2)$ [6]. Thus, the computational complexity of the DC-TGDR algorithm is $O(n^2pK)$, where K is the total number of iterations. In the DC-TGDR algorithm, $\triangle\nu$, τ and δ (or K) are tuning parameters which control diversity and sparsity of the resulting estimated coefficients. We will discuss how to select tuning parameters in the following subsection.

2.3 Selection of tuning parameters

The tuning parameter $\triangle\nu$ controls the increment size at each iteration. Since we regulate the L_2 norm of $\widehat{\boldsymbol{\beta}}$ to be one at each iteration, the algorithm is not very sensitive to step size $\triangle\nu$ if we fix it at a moderate value. On the other hand, although smaller step sizes can prevent the proposed algorithm (and also other stagewise algorithms) from being too aggressive, step sizes that are too small result in slow convergence. Therefore, we set $\triangle\nu$ as 0.1 in this algorithm, and all simulation results in Section 3 illustrate that 0.1 is a proper value.

We update the parameters along with the regularized direction $\mathbf{h}(\widehat{\boldsymbol{\beta}}^m)$ in the mth iteration until the difference of the objective functions between two iterations is small enough. That is, $|\mathcal{V}_n^2(\mathbf{X}\widehat{\boldsymbol{\beta}}^m, \mathbf{Y}) - \mathcal{V}_n^2(\mathbf{X}\widehat{\boldsymbol{\beta}}^{m+1}, \mathbf{Y})| < \delta$. Thus, δ is the stopping rule of the algorithm. The smaller δ is, the more iterations are needed to ensure convergence, and the more coefficients will be estimated as nonzero. Thus, it controls both computation time and sparsity of the estimator. In our simulations, we use five-fold cross-validation to select the optimal value of δ from $[1, 5, 10, 15, \ldots, 100] \times 10^{-6}$.

The tuning parameter τ sets a threshold for each coordinate of the gradient vector such that some coefficients with relatively small gradients are estimated to be zero. Thus, it is important to investigate the effect of τ in the DC-TGDR algorithm. How τ controls the diversity and sparsity of coefficients has been studied in the TGDR method for linear models by Friedman and Popescu [3]. Next, we illustrate how τ affects the single-index direction estimator in a simple simulation. We consider a single-index model $Y = \sin\left(\boldsymbol{\beta}^T X / \| \boldsymbol{\beta} \|\right) + 0.1\varepsilon$, which has been studied by Zhu and Zhu [31], Peng and Huang [14] and Radchenko [17]. Set $\boldsymbol{\beta} = (3, 1.5, 0, 0, 0, 0, 2, 0, \cdots, 0)_{20}^T$ and $\varepsilon \sim N(0, 1)$. X is generated from $N(0, \Sigma)$, where $\Sigma = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = 0.5^{|i-j|}$. The sample size n is 100. We apply the DC-TDGR method with three different threshold values $\tau = 0.2$, 0.5 and 0.9. The coefficient paths in each case are plotted in Figure 1.

When $\tau = 0$, the scaling factors $f_j(\widehat{\boldsymbol{\beta}}^m) = 1$ for all $j = 1, \ldots, p$ which makes each coordinate of the estimator updated. The coefficient paths solved with $\tau = 0$ are similar to those derived by imposing an L_2 penalty to the objective function. On the other hand, when $\tau = 1$, only one scaling factor $f_j(\widehat{\boldsymbol{\beta}}^m) = 1$, where gradient of $\widehat{\beta}_j^m$ has the largest absolute value. This makes most of coordinates of the estimator remain the current values or zeros. In this case, the coefficient paths are like those in the L_1 penalized regression, and the resulting estimator is generally sparse.

We also plot the solution paths of the single-index direction estimates when τ is varying from 0 to 1 in Figure 2.

According to Figures 1 and 2, we can see that all important predictors can be selected in three cases but a larger value of τ produces the sparser model. Thus, it is natural to choose the value of τ close to 1 when one has some prior information that the model is sparse or the dimension of single-index direction is very high. Conversely, if the original dimension of single-index direction is not high, and hence there is no need to select important variables, setting the value τ close to 0 may produce a more desirable result. With-

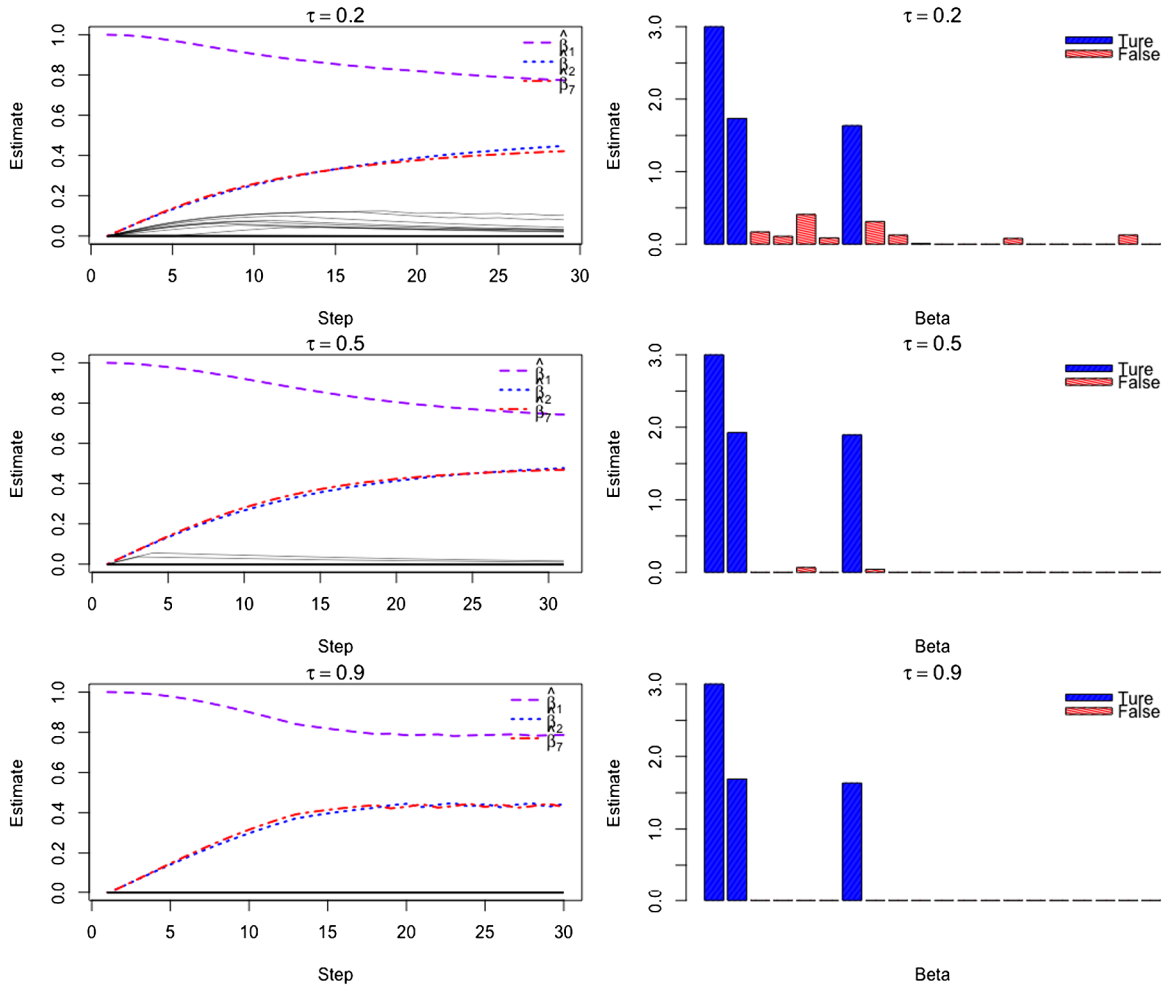

Figure 1. Solution paths with $\tau = 0.2,\ 0.5$ *and* 0.9.

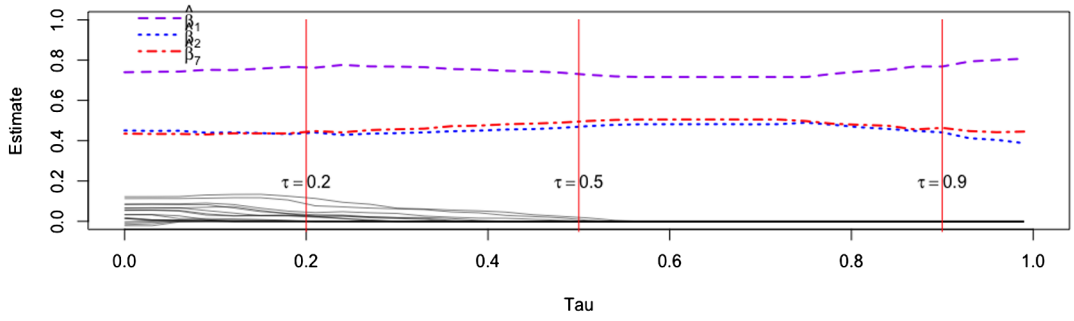

Figure 2. Solution paths as a function of τ.

out any prior information about the single-index direction, one can apply the cross-validation technique to choose the optimal value of τ. In our simulations, we use five-fold cross-validation to determine τ from the set $[0.5, 0.55, \ldots, 0.95, 1]$.

2.4 Grouping effect

In high dimensional data problems, it is of importance to study the "grouped variables" situation. For example, highly correlated CNAs, which can be regarded as a cluster structure, are likely to have similar impact on the expression level of a gene. It is meaningful to select grouped variables in or out of the model together. Many penalization methods in the literature, such as Group Lasso [28] and Elastic Net [32], encourage the grouping effect. Segal, Dahlquist and Conklin [21] suggests using a regularized regression procedure to identify grouped genes. The proposed DC-TGDR method

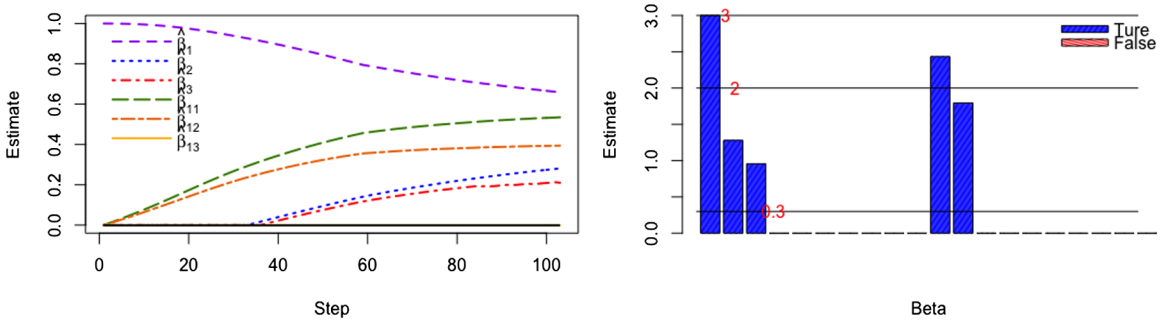

Figure 3. Grouping effect of the DC-TGDR algorithm.

also exhibits the grouping effect and encourages the coefficients of highly correlated predictors to have similar sizes.

When predictors are highly correlated, they are likely to have similar gradient values. Thus, they have the similar chances of being updated in Step 2(2)-(3) of the DC-TGDR algorithm when $\tau < 1$. In other words, they as a group can be selected together with a high probability. In the most extreme case, when $\mathbf{x}_i = \mathbf{x}_j, i, j \in \{1, \ldots, p\}$, $\widehat{\beta}_i = \widehat{\beta}_j$ and when $\mathbf{x}_i = -\mathbf{x}_j$, $\widehat{\beta}_i = -\widehat{\beta}_j$ are guaranteed by the DC-TGDR method, which is the same with the elastic net.

To demonstrate the grouping effect of the DC-TGDR method, the coefficient paths of another simulated example are displayed in Figure 3. The simulated model is $Y = (\boldsymbol{\beta}^T X)^2 + \varepsilon$ with $\boldsymbol{\beta} = (\underbrace{3, 2, 0.3, 0, \cdots, 0}_{10}, \underbrace{3, 2, 0.3, 0, \cdots, 0}_{10})^T$ and $\varepsilon \sim N(0, 1)$. The first three predictors are generated from a multivariate normal distribution with correlation $\rho^{ij} = 0.9^{|i-j|}$ for $i, j \in \{1, 2, 3\}$ but all other predictors are generated independently from $N(0, 1)$. The sample size n is 200.

In Figure 3, when $\tau = 0.5$, the proposed method shows shrinkage effect as well as grouping effect. For the shrinkage effect, since $\beta_{13} = 0.3$ is relatively small compared with other nonzero coefficients and X_{13} is independent of other predictors, the estimated coefficient of X_{13} shrinks to zero and X_{13} is deleted from the model. On the other hand, the grouping effect can be illustrated by $\widehat{\beta}_3$. Although the value of β_3 is also 0.3, its estimate $\widehat{\beta}_3$ is close to $\widehat{\beta}_2$ because of the high correlation among X_1, X_2 and X_3. The difference between $\widehat{\beta}_3$ and $\widehat{\beta}_{13}$ stems from the grouping effect of the DC-TGDR method.

3. SIMULATION

In this section, we compare the finite-sample performance of the DC-TGDR method with recently developed methods including HD-SIM in Radchenko [17] and RoSIS-PeQuan in Zhong et al. [29]. The HD-SIM method is an L_1 regularization method which recovers the index direction by minimizing the restricted least squares criterion. It updates the nonparametric link function by a B-spline method and parametric single-index direction iteratively. The RoSIS-PeQuan is a two-step procedure which first utilizes a model-free robust screening to reduce the dimension and further estimates the parameter index using a penalized linear quantile regression.

To comprehensively assess their empirical performances, we consider the following four single-index models,

(A) $Y = \sin(\boldsymbol{\beta}^T X / \parallel \boldsymbol{\beta} \parallel) + 0.1\varepsilon$

(B) $Y = (\boldsymbol{\beta}^T X)^2 + \varepsilon$

(C) $Y = \exp(2 - \boldsymbol{\beta}^T X/2) + (2 - \boldsymbol{\beta}^T X/2)^2 + \exp(\boldsymbol{\beta}^T X/2)\varepsilon$

(D) $Y = \begin{cases} 1 & \boldsymbol{\beta}^T X > 2 \\ 0 & \text{otherwise.} \end{cases}$

Models (A) and (B) are classic single-index models where the conditional mean of Y given X only depends on a single linear combination of predictors. Model (C) from Zhong et al. [29] is a heteroscedastic single-index model where both the conditional mean and conditional variance of Y are based on a single index. In Model (D) which is examined in Sheng and Yin [22], the response Y is binary.

We generate covariates using three different kinds of correlation structures as below.

(i) X is from $N(0, \Sigma)$, where $\Sigma = (\sigma_{ij})_{p\times p}$ with $\sigma_{ij} = 0.5^{|i-j|}$ and the true coefficient vector $\boldsymbol{\beta} = (3, 1.5, 0, 0, 0, 0, 2, 0, \cdots, 0)_p^T$, where only three predictors are important.

(ii) X is generated in the following way:

$$
\begin{aligned}
&X_i = Z_1 + \varepsilon_i^x, Z_1 \sim N(0,1), i = 1, \ldots, 5,\\
&X_i = Z_2 + \varepsilon_i^x, Z_2 \sim N(0,1), i = 6, \ldots, 10,\\
&X_i = Z_3 + \varepsilon_i^x, Z_3 \sim N(0,1), i = 11, \ldots, 15,\\
&X_i \overset{i.i.d}{\sim} N(0,1), i = 16, \ldots, p.\\
&\varepsilon_i^x \overset{i.i.d}{\sim} N(0, 0.01), i = 1, \ldots, 15.
\end{aligned}
$$

Set $\boldsymbol{\beta} = (\underbrace{1, \ldots, 1, 0.6, \ldots, 0.6, 0.4, \ldots, 0.4}_{15}, 0, \cdots, 0)_p^T$. There are three groups of strongly correlated important predictors.

(iii) X is from $N(0, \Sigma)$, where $\sigma_{ij} = 0.6$ for $i, j = 1, \ldots, 5$ but $i \neq j$, all other $\sigma_{ij} = 0$ for $i \neq j$ and all diago-

nal elements $\sigma_{jj} = 1$ for $j = 1, \ldots, p$. This structure indicates that the first five predictors are correlated as a group and equally important for the response. Set $\boldsymbol{\beta} = (\underbrace{1, 1, 1, 1, 1}_{5}, 2, 1, 0, 0, 0, 0, 1.5, 0, \cdots, 0)^T_p$. There are 8 important predictors.

Two types of random error term are considered, $\varepsilon \sim N(0, 1)$ or $\varepsilon \sim t(2)$. We set the sample size $n = 100$ or 200 and the dimension of predictors $p = 400$. We consider the mean and standard variation of the following four criteria based on 100 replicates to compare empirical performances.

Size: The number of non-zero estimated regression coefficients $\widehat{\beta}_j \neq 0$ for $1 \leq j \leq p$;

C: The number of truly non-zero coefficients correctly estimated to be non-zero;

IC: The number of truly zero coefficients incorrectly estimated to be non-zero;

AE: The absolute estimation error of $\widehat{\boldsymbol{\beta}}$, $\sum_{j=1}^{p} \left| \widehat{\beta}_j \text{sign}(\widehat{\beta}_{j,1}) / \parallel \widehat{\boldsymbol{\beta}} \parallel - \beta_j \text{sign}(\beta_{j,1}) / \parallel \boldsymbol{\beta} \parallel \right|$.

Tables 1-3 summarize the simulation results measured by the aforementioned four criteria. Note that, according to assumptions in Radchenko [17], the HD-SIM method can be only applied to the simulated Models (A) and (B) which are homoscedastic models. Thus, it is unfair to evaluate the performance of HD-SIM in Models (C) and (D). The PeQuan method is based on quantile regression and cannot be used for binary data in Model (D). Overall, the DC-TGDR method performs better than the two alternative methods in most simulation scenarios, especially when the sample size is large, the error is heavy-tailed or the grouping effect exists in the model. Specially, the PeQuan method has difficulty in identifying important variables in Model (B) because the link function of the single index is quadratic. Thus, the empirical performance of the PeQuan method is generally worse for Model (B) under all three different correlation structures. For correlation structures (ii) and (iii) where the grouping effects are present, neither HD-SIM nor PeQuan can satisfactorily detect the grouped variables, but our method works well in identifying the groups. Moreover, the DC-TGDR method has an outstanding performance in variable/group selection and direction estimation when the response is binary in Model (D).

In addition, we evaluate finite-sample performance of the DC-TGDR algorithm in higher dimensional situations by setting $p = 1000$ when $n = 200$. We consider the same four models with three different correlation structures as in the previous simulations. The results are summarized in Table 4. We can see that our DC-TGDR method can still work satisfactorily and outperform the existing methods for the higher dimensional cases.

In summary, the DC-TGDR method is capable of selecting important predictors accurately and estimating the index direction simultaneously for general single-index models. Meanwhile, it also encourages the grouping effects to automatically identify groups of highly correlated variables. All these simulations results support the applicability of the proposed DC-TGDR for various single-index models.

4. DATA ANALYSIS

With the development of profiling technology, researchers are now able to collect measurements on multiple layers of cellular molecules, such as DNA copy number alternations (CNAs), gene expressions (GEs), protein expressions and so on. Various diseases are caused by abnormality in gene expressions which are partly regulated by CNAs. Therefore, it is of significant importance to investigate how CNAs influence the expression level of genes. Various studies have been performed to examine the relationship between CNAs and gene expression. For example, Schäfer et al. [19] proposes an approach based on a modified correlation coefficient and an explorative Wilcoxon test to search genetic regions where CNAs and GEs display strong equally directed deviations from the reference levels. Peng et al. [15] developes a remMap method for multivariate linear regression analysis between CNAs and GEs to identify master predictors.

In this section, we analyze the TCGA data[1] and examine the Cell Development pathway, in which genes control cellular differentiation, growth as well as apoptosis. The abnormal and unregulated cell growth is an underlying reason of forming neoplasms. With the lump invading or spreading to parts of the body diffusely, cancer develops into a fatal disease. Therefore, the Cell Development pathway is closely related with the occurrence of cancer. In our data set, GE and CNA measurements are available on 275 patients. There are 563 gene expressions and 341 CNAs in the pathway. Nonlinear relationships between GEs and CNAs often exhibit as shown in later figures. Thus, the commonly used linear regression is not adequate to model the data and may suffer from model misspecification. As an illustrative example, we study how CNAs regulate the expression level of gene FAS which controls the production of FAS receptor. As one of the most important members of the death receptor family, FAS receptor can trigger apoptosis. Since apoptosis plays an instrumental role in regulation of the immune system, the aberrant expression of FAS gene may result in oncogenesis and drug resistance of malignant tumours [18, 1]. Thus, it is critical to model the relationship between the expression of FAS gene and CNAs.

First, we conduct exploratory analysis. We plot the marginal relationships between the expression level of gene FAS and four CNAs, UBB, RRAGA, BCL10 and CDK5R1, which are selected as the most influential predictors by the DC-TGDR method. Nonlinear relationships clearly exist in Figure 4. In addition, the correlation heatmap of all CNAs after reordering via the hierarchical clustering is plotted in the left panel of Figure 5. We can see that there are groups of highly correlated CNAs. Therefore, it is necessary to ac-

[1] http://cancergenome.nih.gov

Table 1. Simulation Results for Correlation Structure (i)

Model	Method	Size	C	IC	AE
			n=100, p=400, $\varepsilon \sim N(0,1)$		
	HD-SIM	16.62(6.75)	3(0)	13.62(6.75)	0.29(0.07)
(A)	PeQuan	3.84(1.87)	2.94(0.24)	0.9(1.83)	0.19(0.18)
	DC-TGDR	4.74(1.79)	3(0)	1.74(1.79)	0.22(0.14)
	HD-SIM	11.18(6.43)	2.84(0.47)	8.34(6.3)	0.32(0.42)
(B)	PeQuan	27.02(10.45)	1.08(0.83)	25.94(9.98)	3.57(2.04)
	DC-TGDR	10.46(14.32)	2.72(0.78)	7.74(14.78)	0.61(1.36)
(C)	PeQuan	4.44(3.98)	2.54(0.65)	1.9(3.79)	0.82(0.81)
	DC-TGDR	4.32(1.63)	3(0)	1.32(1.63)	0.26(0.17)
(D)	DC-TGDR	8.12(4.34)	2.96(0.2)	5.16(4.29)	0.51(0.24)
			n=100, p=400, $\varepsilon \sim t(2)$		
	HD-SIM	10.36(5.62)	2.84(0.55)	7.52(5.47)	0.67(0.44)
(A)	PeQuan	6.12(5.19)	2.92(0.34)	3.2(5.19)	0.5(0.67)
	DC-TGDR	4.82(1.98)	3(0)	1.82(1.98)	0.24(0.16)
	HD-SIM	12.82(8.73)	2.68(0.77)	10.14(8.77)	0.54(0.52)
(B)	PeQuan	26.2(9.58)	1.02(0.77)	25.18(9.23)	3.23(1.96)
	DC-TGDR	12.76(18.47)	2.7(0.84)	10.06(19.12)	0.36(0.49)
(C)	PeQuan	4.88(4.2)	2.68(0.62)	2.2(4.05)	0.66(0.79)
	DC-TGDR	4.58(2.78)	2.96(0.2)	1.62(2.76)	0.26(0.18)
(D)	DC-TGDR	8.26(3.74)	3(0)	5.26(3.74)	0.57(0.29)
			n=200, p=400, $\varepsilon \sim N(0,1)$		
	HD-SIM	19(5.54)	3(0)	16(5.54)	0.2(0.04)
(A)	PeQuan	4.62(4.69)	3(0)	1.62(4.69)	0.12(0.16)
	DC-TGDR	3.6(0.78)	3(0)	0.6(0.78)	0.13(0.06)
	HD-SIM	10.38(4.13)	3(0)	7.38(4.13)	0.06(0.11)
(B)	PeQuan	44.1(25.19)	1.16(1.08)	42.94(24.4)	4.1(2.88)
	DC-TGDR	3.6(1.01)	3(0)	0.6(1.01)	0.09(0.05)
(C)	PeQuan	7.78(10.46)	2.86(0.45)	4.92(10.39)	0.63(0.76)
	DC-TGDR	3.48(0.68)	3(0)	0.48(0.68)	0.19(0.13
(D)	DC-TGDR	4.62(1.56)	3(0)	1.62(1.56)	0.19(0.09)
			n=200, p=400, $\varepsilon \sim t(2)$		
	HD-SIM	15.46(7.64)	2.94(0.42)	12.52(7.58)	0.47(0.25)
(A)	PeQuan	3.88(2.67)	2.96(0.2)	0.92(2.65)	0.15(0.18)
	DC-TGDR	3.76(0.92)	3(0)	0.76(0.92)	0.16(0.08)
	HD-SIM	16.88(5.45)	2.96(0.28)	13.92(5.48)	0.15(0.23)
(B)	PeQuan	42.88(22.08)	1.08(0.9)	41.8(21.45)	3.61(2.8)
	DC-TGDR	4.18(1.65)	3(0)	1.18(1.65)	0.09(0.05)
(C)	PeQuan	5.48(7.17)	2.82(0.56)	2.66(7.08)	0.41(0.56)
	DC-TGDR	3.76(0.85)	3(0)	0.76(0.85)	0.16(0.06)
(D)	DC-TGDR	4.54(1.73)	3(0)	1.54(1.73)	0.23(0.13)

[1] The true coefficients $\boldsymbol{\beta}_0 = (3, 1.5, 0, 0, 0, 0, 2, 0, \cdots, 0)_p^T$.

[2] When the estimate of β_i is less than 0.001, we regard the corresponding predictor as unimportant.

[3] The means of four criteria based on 100 data replicates with the standard deviations in parentheses.

[4] Tuning parameters τ and δ in DC-TGDR are selected by five-fold cross validation.

commodate nonlinear relationships and grouping structures in this data analysis.

Next, we consider the framework of robust single-index models and use the DC-TGDR method to identify important CNAs and estimate the index direction for gene FAS. For a comparison purpose, we also apply the RoSIS-PeQuan method and the HD-SIM method. The estimates of single-index direction by the three methods are listed in Table 5. The DC-TGDR method selects 25 important CNAs for gene FAS. The PeQuan method chooses 12 CNAs as important ones but it misses gene UBB which is selected as the most important CNA by the other two methods. On the other hand, the HD-SIM method detects only one important CNA. Studies on the functions and interactions of genes in the Cell Development pathway have partially supported the validity of the DC-TGDR method. Among the selected CNAs, UBB, RRAGA and BCL10 are the top three which make the largest impact on the expression level of

Table 2. Simulation Results for Correlation Structure (ii)

Model	Method	Size	C	IC	AE
			n=100, p=400, $\varepsilon \sim N(0,1)$		
	HD-SIM	9.06(9.3)	3.86(3.1)	5.2(8.38)	3.59(0.13)
(A)	PeQuan	31.38(9.79)	2.72(1.91)	28.66(8.8)	4.68(1.65)
	DC-TGDR	70.26(64.62)	9.8(5.24)	60.46(64.62)	4.4(3.44)
	HD-SIM	16.98(8.58)	5.88(2.24)	11.1(8)	3.67(0.3)
(B)	PeQuan	29.32(8.88)	1.64(1.4)	27.68(8.18)	4.12(1.11)
	DC-TGDR	40.5(46.59)	12.18(4.9)	28.32(48.96)	1.99(2.36)
(C)	PeQuan	13.26(9.01)	4.12(1.83)	9.14(7.8)	3.78(0.43)
	DC-TGDR	27.8(21.96)	12.7(3.54)	15.1(20.99)	2.26(1.77)
(D)	DC-TGDR	23.9(11.46)	12.5(3.05)	11.4(9.77)	2.16(0.8)
			n=100, p=400, $\varepsilon \sim t(2)$		
	HD-SIM	6.48(6.15)	2.46(2.48)	4.02(5.34)	3.61(0.08)
(A)	PeQuan	31.18(9.26)	2.14(1.85)	29.04(8.54)	5.18(1.96)
	DC-TGDR	86.8(69.4)	8.96(5.86)	77.84(68.34)	5.55(4.13)
	HD-SIM	13.48(9.45)	5.42(2.46)	8.06(8.38)	3.62(0.07)
(B)	PeQuan	27.62(11.18)	1.72(1.63)	25.9(10.21)	4.44(1.51)
	DC-TGDR	27.9(24.17)	13.2(3.93)	14.7(26.18)	1.41(0.96)
(C)	PeQuan	12.48(9.66)	4.32(2.22)	8.16(8.16)	3.81(0.41)
	DC-TGDR	27.54(23.7)	12.48(3.6)	15.06(22.43)	2.35(1.58)
(D)	DC-TGDR	24.82(11.94)	12.92(2.59)	11.9(10.69)	2.15(0.85)
			n=200, p=400, $\varepsilon \sim N(0,1)$		
	HD-SIM	19.02(7.21)	7.24(2)	11.78(6.98)	3.57(0.16)
(A)	PeQuan	53.56(22.79)	4.4(3.04)	49.16(21.15)	5.24(1.93)
	DC-TGDR	34.88(42.15)	13.56(3.23)	21.32(43.99)	1.75(2.37)
	HD-SIM	23.54(4.99)	8.94(0.98)	14.6(4.86)	3.51(0.22)
(B)	PeQuan	44.92(27)	2.8(3.03)	42.12(24.96)	5.15(2.04)
	DC-TGDR	18.52(3.13)	14.92(0.27)	3.6(3.06)	0.62(0.24)
(C)	PeQuan	22.04(17.79)	5.34(2.08)	16.7(16.32)	3.68(0.24)
	DC-TGDR	18.74(6)	14.32(1.94)	4.42(6.11)	1.19(0.92)
(D)	DC-TGDR	18.38(4.14)	14.24(1.88)	4.14(3.31)	0.83(0.59)
			n=200, p=400, $\varepsilon \sim t(2)$		
	HD-SIM	16.2(8.96)	5.64(2.67)	10.56(7.74)	3.6(0.1)
(A)	PeQuan	49.2(24.56)	3.82(3.52)	45.38(22.67)	4.88(1.91)
	DC-TGDR	31.24(32.87)	13.56(2.79)	17.68(33.82)	1.96(2.63)
	HD-SIM	23.2(5.45)	8.76(1.13)	14.44(5.08)	3.58(0.14)
(B)	PeQuan	47.36(23.37)	2.24(2.26)	45.12(22.06)	4.38(1.59)
	DC-TGDR	17.76(3.04)	14.84(0.47)	2.92(2.93)	0.59(0.24)
(C)	PeQuan	21.08(18.42)	5.4(1.86)	15.68(17.25)	3.65(0.34)
	DC-TGDR	19.88(6.14)	14.7(1.04)	5.18(6.33)	1.21(0.87)
(D)	DC-TGDR	19.46(4.51)	14.54(1.63)	4.92(4.03)	0.83(0.46)

[1] The true coefficients

$$\boldsymbol{\beta} = (\underbrace{1,1,1,1,1,0.6,0.6,0.6,0.6,0.6,0.4,0.4,0.4,0.4,0.4}_{15},0,\cdots,0)_p^T$$

.

[2] When the estimate of β_i is less than 0.001, we regard the corresponding predictor as unimportant.

[3] The means of four criteria based on 100 data replicates with the standard deviations in parentheses.

[4] Tuning parameters τ and δ in DC-TGDR are selected by five-fold cross validation.

gene FAS. In the literature, studies have revealed that the ubiquitin encoded by UBB is closely related to the degradation of abnormal proteins, which is consequently involved in the regulation of gene expression [2]. Also, the proteins encoded by RRAGA belong to the ubiquitously expressed Ras family, which controls processes such as cell adhesion, apoptosis, and cell migration [20]. BCL10 has clinical significance in lymphoma, a type of cancer developed in the immune system [24]. In summary, previous studies have illustrated that the expression level of gene FAS is potentially regulated by UBB; and FAS, RRAGA and BCL10 are functionally and jointly associated with the abnormality of cell

Table 3. Simulation Results for Correlation Structure (iii)

Model	Method	Size	C	IC	AE
			n=100, p=400, $\varepsilon \sim N(0,1)$		
	HD-SIM	5.46(4.06)	2.88(1.3)	2.58(3.65)	2.77(0.13)
(A)	PeQuan	20.94(9.62)	5.34(1.64)	15.6(9.13)	3.2(1.38)
	DC-TGDR	13.5(4.22)	7.54(0.58)	5.96(4.03)	1.32(0.37)
	HD-SIM	4.92(4.85)	1.28(0.78)	3.64(4.76)	2.83(0.29)
(B)	PeQuan	27.16(9.85)	1.78(1.43)	25.38(9.07)	4.67(1.96)
	DC-TGDR	44.24(42.78)	5.24(3.17)	39(44.26)	2.6(2.53)
(C)	PeQuan	12.58(10.53)	4.86(2.06)	7.72(9.34)	3.14(1.26)
	DC-TGDR	12.06(5.28)	6.82(1.06)	5.24(4.83)	1.72(0.64)
(D)	DC-TGDR	17.92(7.63)	7.08(0.88)	10.84(7.12)	2.03(0.46)
			n=100, p=400, $\varepsilon \sim t(2)$		
	HD-SIM	5.22(3.18)	2.5(1.36)	2.72(2.63)	2.79(0.21)
(A)	PeQuan	23.12(9.91)	5.38(1.43)	17.74(9.14)	3.9(1.61)
	DC-TGDR	12.06(3.56)	6.02(0.47)	6.04(3.53)	2.31(0.26)
	HD-SIM	5.72(6.75)	1.26(0.8)	4.46(6.72)	2.84(0.21)
(B)	PeQuan	27.24(11.01)	2.14(1.47)	25.1(10.01)	4.36(1.92)
	DC-TGDR	42.82(41.76)	5.3(3.15)	37.52(43.12)	2.86(2.92)
(C)	PeQuan	12.5(10.2)	4.14(2.09)	8.36(8.71)	3.08(1.35)
	DC-TGDR	12.82(6.15)	6.88(1.14)	5.94(5.81)	1.77(0.78)
(D)	DC-TGDR	18.02(7.16)	7.16(0.82)	10.86(6.67)	1.97(0.43)
			n=200, p=400, $\varepsilon \sim N(0,1)$		
	HD-SIM	13.04(4.45)	7.56(0.67)	5.48(4.09)	1.98(0.34)
(A)	PeQuan	16.86(8.78)	6.8(1.4)	10.06(8.16)	1.76(0.84)
	DC-TGDR	8.78(1.3)	7.86(0.35)	0.92(1.14)	0.79(0.19)
	HD-SIM	9.38(5.24)	4.52(1.82)	4.86(4.51)	2.62(0.2)
(B)	PeQuan	49.44(26.26)	3.12(2.29)	46.32(24.33)	5.8(2.92)
	DC-TGDR	16.82(31.6)	7.52(1.63)	9.3(32.8)	0.84(1.41)
(C)	PeQuan	14.16(12.35)	6.76(1.8)	7.4(11.64)	1.91(1.28)
	DC-TGDR	8.02(1.13)	7.54(0.65)	0.48(0.74)	1.06(0.26)
(D)	DC-TGDR	11.06(3.13)	7.7(0.46)	3.36(2.92)	1.1(0.29)
			n=200, p=400, $\varepsilon \sim t(2)$		
	HD-SIM	12.06(4.85)	6.56(1.86)	5.5(4.11)	2.31(0.3)
(A)	PeQuan	16.78(9.78)	6.24(1.12)	10.54(9.41)	1.99(0.67)
	DC-TGDR	9.98(2.14)	7.94(0.24)	2.04(2.07)	0.8(0.21)
	HD-SIM	9.12(6.54)	3.92(1.99)	5.2(5.86)	2.71(0.22)
(B)	PeQuan	43.08(27.31)	2.68(2.22)	40.4(25.37)	5.95(2.9)
	DC-TGDR	10.3(3.87)	7.58(1.68)	2.72(5.41)	0.63(0.55)
(C)	PeQuan	14.18(14.56)	6.6(1.81)	7.58(14.19)	1.82(1.43)
	DC-TGDR	8.24(1.32)	7.52(0.58)	0.72(1.09)	1.04(0.32)
(D)	DC-TGDR	11.7(3.41)	7.64(0.56)	4.06(3.18)	1.15(0.31)

[1] The true coefficients

$$\boldsymbol{\beta} = (\underbrace{1,1,1,1,1}_{5},2,1,0,0,0,0,1.5,0,\cdots,0)_p^T$$

.

[2] When the estimate of β_i is less than 0.001, we regard the corresponding predictor as unimportant.

[3] The means of four criteria based on 100 data replicates with the standard deviations in parentheses.

[4] Tuning parameters τ and δ in DC-TGDR are selected by five-fold cross validation.

development. In other words, the important CNAs identified by the proposed DC-TGDR method are consistent with the existing discoveries.

Besides, we plot the correlation heatmap of the 25 selected CNAs by the DC-TGDR method in the right panel of Figure 5. It illustrates that the DC-TGDR method selects a group of strongly correlated CNAs, which includes PRL, DUSP22, TXNDC5, IL17A, VEGFA, CDKN1A and MDGA1. The average pairwise Pearson correlation values among these seven CNAs is 0.79. Hence, they form a bight yellow square in the lower right corner. However, with PeQuan, only PRL and CDKN1A are selected out among these

Table 4. Simulation Results When $p = 1000$, $n = 200$

Model	Method	Size	C	IC	AE
			Correlation(i), $\varepsilon \sim N(0,1)$		
	HD-SIM	17.54(6.46)	3(0)	14.54(6.46)	0.22(0.05)
Model A	PeQuan	4.34(1.81)	2.84(0.42)	1.5(2.01)	0.42(0.44)
	DC-TGDR	3.52(0.79)	3(0)	0.52(0.79)	0.14(0.07
	HD-SIM	11.16(5.21)	2.94(0.42)	8.22(5.33)	0.14(0.38)
Model B	PeQuan	19.5(17.65)	0.54(0.76)	18.96(17.19)	2.52(1.87)
	DC-TGDR	4.12(2)	3(0)	1.12(2)	0.09(0.04)
Model C	PeQuan	7.58(6.89)	2.4(0.81)	5.18(6.62)	0.92(0.75)
	DC-TGDR	3.44(0.88)	3(0)	0.44(0.88)	0.14(0.09)
Model D	DC-TGDR	5.12(1.83)	3(0)	2.12(1.83)	0.22(0.14)
			Correlation(i), $\varepsilon \sim t(2)$		
	HD-SIM	14.96(5.7)	3(0)	11.96(5.7)	0.49(0.19)
Model A	PeQuan	5(5.54)	3(0)	2(5.54)	0.19(0.25)
	DC-TGDR	3.76(1.06)	3(0)	0.76(1.06)	0.17(0.08)
	HD-SIM	16.64(7.38)	3(0)	13.64(7.38)	0.12(0.1)
Model B	PeQuan	44.72(24.97)	1.02(0.94)	43.7(24.26)	4.36(3)
	DC-TGDR	4.6(2.03)	3(0)	1.6(2.03)	0.1(0.05)
Model C	PeQuan	6.5(11.17)	2.66(0.66)	3.84(11.03)	0.64(0.8)
	DC-TGDR	3.5(0.71)	2.98(0.14)	0.52(0.68)	0.19(0.16)
Model D	DC-TGDR	5.28(1.75)	3(0)	2.28(1.75)	0.25(0.11)
			Correlation(ii), $\varepsilon \sim N(0,1)$		
	HD-SIM	18.3(7.62)	7.16(2.23)	11.14(6.51)	3.56(0.18)
Model A	PeQuan	65.18(9.12)	4.88(2.38)	60.3(8.15)	5.98(2.1)
	DC-TGDR	47.08(64.96)	11.96(4.62)	35.12(67.67)	2.29(2.95)
	HD-SIM	22.34(4.87)	8.68(1.19)	13.66(4.34)	3.57(0.17)
Model B	PeQuan	46.44(23.29)	1.5(1.59)	44.94(22.28)	4.64(1.86)
	DC-TGDR	20.76(6.85)	14.92(0.27)	5.84(6.8)	0.61(0.23)
Model C	PeQuan	19.42(16.91)	4.82(1.97)	14.6(15.58)	4(0.84)
	DC-TGDR	24.06(16.88)	14.48(1.78)	9.58(16.55)	1.46(1.03)
Model D	DC-TGDR	21.88(7.73)	14.46(0.95)	7.42(7.46)	0.98(0.37)
			Correlation(ii), $\varepsilon \sim t(2)$		
	HD-SIM	14.22(8.55)	5.16(2.85)	9.06(7.71)	3.61(0.07)
Model A	PeQuan	59.52(16.39)	3.32(2.27)	56.2(15.5)	6.07(2.47)
	DC-TGDR	49.74(82.67)	13.08(3.3)	36.66(84.34)	2.28(3.09)
	HD-SIM	23.04(6.33)	8.68(1.22)	14.36(5.9)	3.55(0.18)
Model B	PeQuan	51.94(23.17)	2.26(2.53)	49.68(21.88)	4.96(2.18)
	DC-TGDR	19.6(5.23)	14.94(0.24)	4.66(5.2)	0.65(0.26)
Model C	PeQuan	15.04(14.28)	4.72(1.77)	10.32(13.28)	3.82(0.56)
	DC-TGDR	23.56(10.81)	14.78(0.86)	8.78(10.66)	1.35(0.92)
Model D	DC-TGDR	23(10.11)	14.44(1.53)	8.56(9.76)	1.06(0.62)
			Correlation(iii), $\varepsilon \sim N(0,1)$		
	HD-SIM	9.28(2.83)	6.46(1.05)	2.82(2.68)	2.4(0.26)
Model A	PeQuan	17.78(12.77)	6.02(1.53)	11.76(12.37)	2.25(1.16)
	DC-TGDR	12.06(3.96)	7.92(0.27)	4.14(3.9)	0.86(0.26)
	HD-SIM	7.22(6.04)	2.88(1.45)	4.34(5.75)	2.77(0.23)
Model B	PeQuan	48.88(24.15)	2.42(2.07)	46.46(22.5)	5.2(2.87)
	DC-TGDR	34.24(78.57)	7.14(2.33)	27.1(79.72)	1.1(2.23)
Model C	PeQuan	9.4(7.03)	6.18(2.21)	3.22(6.25)	2.05(1.57)
	DC-TGDR	11.74(4.8)	7.72(0.5)	4.02(4.64)	1.12(0.38)
Model D	DC-TGDR	14.58(6.02)	7.56(0.58)	7.02(5.67)	1.24(0.2
			Correlation(iii), $\varepsilon \sim t(2)$		
	HD-SIM	8.68(4.73)	4.92(1.94)	3.76(4.25)	2.61(0.19)
Model A	PeQuan	23.24(14.41)	6.48(1.25)	16.76(13.83)	2.66(1.3)
	DC-TGDR	13.52(4.52)	7.86(0.35)	5.66(4.38)	0.99(0.24)
	HD-SIM	6.28(5.12)	2.54(1.58)	3.74(4.66)	2.81(0.18)
Model B	PeQuan	41.38(25.94)	1.96(1.74)	39.42(24.51)	5.24(2.79)
	DC-TGDR	25.44(49.67)	7.5(1.85)	17.94(51.15)	1.08(2.19)
Model C	PeQuan	17.48(16.48)	6.24(2.26)	11.24(15.77)	1.94(1.36)
	DC-TGDR	10.76(4.05)	7.44(0.7)	3.32(3.75)	1.23(0.44)
Model D	DC-TGDR	16.96(6.98)	7.68(0.47)	9.28(6.76)	1.34(0.31)

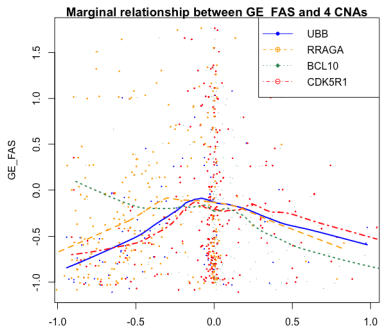

Figure 4. Marginal regression of FAS gene expression level versus four CNAs.

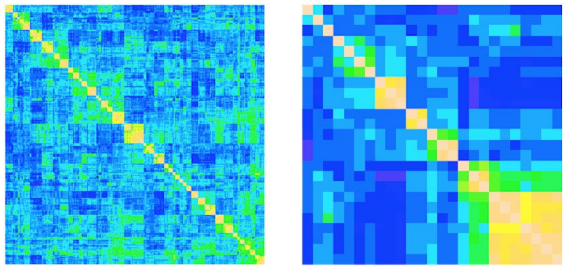

Figure 5. The left panel is the correlation heatmap of all CNAs and the right one is the correlation heatmap of CNAs selected by DC-TGDR method. The bright yellow squares indicate that there exist grouped CNAs.

Table 5. Estimation of Single-index Direction

CNAs	β^1	β^2	β^3	CNAs	β^1	β^2	β^3
FAS	1.69			EP300	-0.62		
IL17A	-0.41			CDKN1A	-0.45	-1.51	
DUSP22	-0.31			DEDD	-0.22		
BNIPL	-1.21	-1.09		UBB	7.38		10
TP53	1.91	4.52		MDGA1	-0.64		
GRIK2	0.12			FARP2	0.69		
CIB1	-2.06			TXNDC5	-0.40		
BCL10	-2.54			VEGFA	-0.84		
CDK5R1	2.08			PRL	-0.71	-1.44	
NOTCH2	-1.56	-0.75		DDAH2		1.43	
BAG4	0.36			ALOX12		-0.71	
TNFRSF10D	1.84			CCL2		0.33	
SEMA4D	0.15			VCP		-0.08	
FOXO3	0.75			TPD52L1		1.09	
CASP8AP2	0.63	1.33		RRAGA	3.52	3.09	

[1] The response variable is the expression level of Gene FAS.
[2] β^1 is solved by DC-TGDR, β^2 is solved by PeQuan and β^3 is solved by HD-SIM.

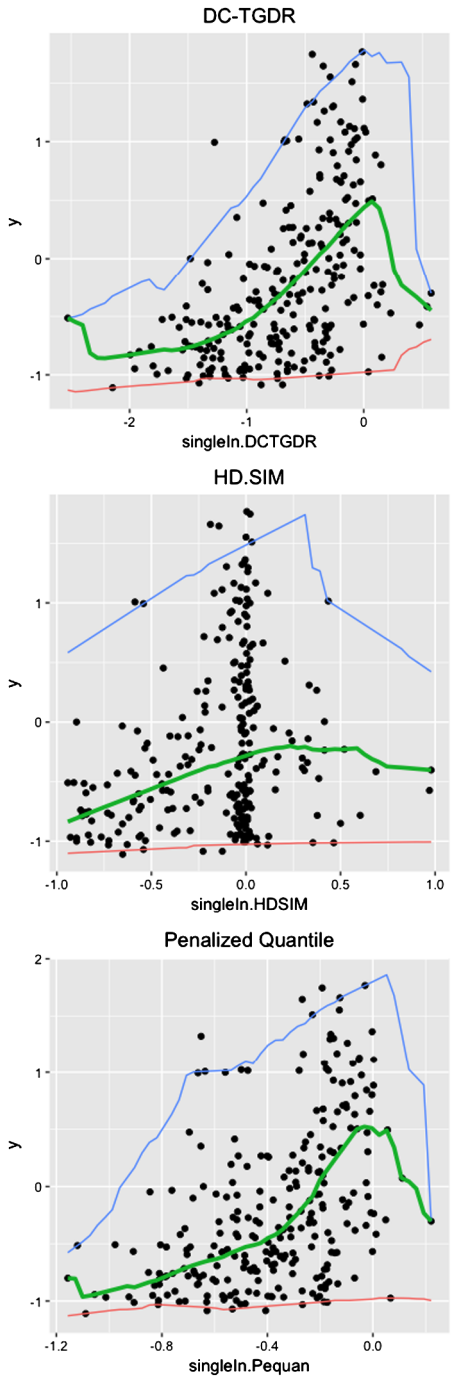

Figure 6. Scatter plots of the expression level of gene FAS against the estimated single indices by the three methods with local linear quantile estimated curves at the 2.5%, 50% and 97.5% quantiles.

seven CNAs. This result also demonstrates that the proposed approach encourages the grouping effect.

In addition, we show the scatter plots of the expression level of gene FAS against the estimated single indices by the three methods in Figure 6. It is easy to see that the relationship between the response and single index is not linear and the variance of gene FAS increases as the single index increases, i.e. the heteroscedasticity is also present. Then, we use the local linear quantile regression to estimate the conditional median, 2.5% quantile and 97.5% quantile of

the response variable conditional on the single index for each method, where the confidence intervals are shown between the red(lower) line and blue(upper) line in Figure 6.

5. CONCLUSION

In this article, we have developed a new DC-TGDR method which can perform variable selection and index direction estimation simultaneously for general single-index models. This method inherits the advantages of both the distance covariance and Threshold Gradient Directed Regularization algorithm. Since the distance covariance is able to measure nonlinear dependence between random variables, the DC-TGDR method avoids estimating unknown link function of the single index so it can reduce computational complexity. As a variant of the original TGDR method, the DC-TGDR method also encourages a grouping effect which is important in many high dimensional problems. Both Monte Carlo simulations and real data analysis demonstrate the favorable empirical performances compared with the existing methods for single-index models. For future study, we may consider the multiple indices models where more than two linear combinations of predictors are considered because the distance covariance can measure the dependence between two random vectors. Besides, we may also investigate the relationship between CNAs and multivariate gene expressions simultaneously using the distance covariance in the pathway studies.

ACKNOWLEDGMENTS

We thank the editor and reviewers for insightful comments, which have led to a significant improvement of this article.

Funding: Wei Zhong's research was supported by NNSFC grants 11301435 and 11671334.

Received 11 November 2016

REFERENCES

[1] Cao, Y., Miao, X., Huang, M., Deng, L., Lin, D., Zeng, Y. and Shao, J. (2010), Polymorphisms of death pathway genes FAS and FASL and risk of nasopharyngeal carcinoma, *Molecular Carcinogenesis*, **49**, 944–950.

[2] Conaway, R. C., Brower, C. S. and Conaway, J. W. (2002), Emerging roles of ubiquitin in transcription regulation, *Science*, **296**, 1254–1258.

[3] Friedman, J. and Popescu, B. E. (2004), Gradient directed regularization for linear regression and classification. Technical report, *Department of Statistics, Stanford University*.

[4] Härdle, W., Hall, P., Ichimura, H. (1993), Optimal smoothing in single-index models, *Annals of Statistics*, **21**, 157–178.

[5] Horowitz, J. L. and Härdle, W. (1996), Direct semiparametric estimation of single-index models with discrete covariates, *Journal of the American Statistical Association*, **91**, 1632–1639.

[6] Huo, X. and Székely, G. (2012). Fast computing for distance covariance, *Technometrics,* **58**, 435–447. MR3556612

[7] Ichimura, H. (1993), Semiparametric least squares (sls) and weighted sls estimation of single-index models, *Journal of Econometrics*, **28**, 71–120. MR1230981

[8] Kong, E. and Xia, Y. (2007), Variable Selection for the single-index model, *Biometrika*, **94**, 217–229. MR2367831

[9] Li, K.-C. (1991), Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**, 316–327. MR1137117

[10] Li, K.-C. and Duan, N. (1989), Regression analysis under link violation, *The Annals of Statistics*, **17**, 1009–1052.

[11] Li, R., Zhong, W. and Zhu, L. (2012), Feature screening via distance correlation learning, *Journal of the American Statistical Association,* **107**, 1129–1139.

[12] Liang, H., Liu, X., Li, R. and Tsai, C. (2010), Estimation and testing for partially linear single-index models, *Annals of Statistics*, **38**, 3811–3836.

[13] Ma, S. and Huang, J. (2007), Clustering threshold gradient descent regularization: with applications to microarray studies, *Bioinformatics*, **23**, 466–472.

[14] Peng, H. and Huang, T. (2011), Penalized least squares for single-index models, *Journal of Statistical Planning and Inference,* **141**, 1362–1379. MR2747907

[15] Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D. Y., Pollack, J. R. and Wang, P. (2010), Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer, *The Annals of Applied Statistics,* **3**, 1236–1265. MR2758084

[16] Powell, J. L., Stock, J. H. and Stoker, T. M. (1989), Semiparametric estimation of index coefficient, *Econometrica,* **51**, 1403–1430.

[17] Radchenko, P. (2015), High dimensional single index models, *Journal of Multivariate Analysis*, **139**, 266–282.

[18] Randhawa, S. R., Chahine, B. G., Lowery-Nordberg, M., Cotelingam, J. D. and Casillas, A. M. (2010), Underexpression and overexpression of Fas and Fas ligand: a double-edged sword, *Annals of Allergy, Asthma & Immunology,* **104**, 286–292.

[19] Schäfer, M., Schwender, H., Merk, S., Haferlach, C., Ickstadt, K. and Dugas, M. (2009), Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities, *Bioinformatics,* **25**, 3228–3235.

[20] Schürmann, A., Brauers, A., Massmann, S. Becker, W. and Joost, H. (1995), Cloning of a novel family of mammalian GTP-binding proteins (RagA, RagBs, RagB1) with remote similarity to the Ras-related GTPases, *Journal of Biological Chemistry*, **270**, 28982–28988.

[21] Segal, M. R., Dahlquist, K. D. and Conklin, B. R. (2003), Regression approach for microarray data analysis, *Journal of Computational Biology*, **10**, 961–980.

[22] Sheng, W. and Yin, X. (2013), Direction estimation in single-index models via distance covariance, *Journal of Multivariate Analysis*, **122**, 148–161. MR3189314

[23] Székely, G. J., Rizzo, M. L. and Bakirov, N. K. (2007), Measuring and testing dependence by correlation of distances, *Annals of Statistics*, **35**, 2769–2794.

[24] Willis, T. G., Jadayel, D. M., Du, M. Q., Peng, H., Perry, A. R., Abdul-Rauf, M., Price, H., Karran, L., Majekodunmi, O., Wlodarska, I. and others, Bcl10 is involved in t (1; 14)(p22; q32) of MALT B cell lymphoma and mutated in multiple tumor types, *Cell*, **96**, 35–45.

[25] Xia, Y. C., Tong, H., Li, W. K., Zhu, L. X. (2002), An adaptive estimation of optimal regression subspace, *Journal of the Royal Statistical Society, Series B*, **64**, 363–410.

[26] Xia, Y. C. and Li, W. K. (1999), On single-index coefficient regression models, *Journal of the American Statistical Association*, **94**, 1275–1285.

[27] Yin, X. and Cook, R. D. (2005), Direction estimation in single-index regressions, *Biometrika*, **92**, 371–384.

[28] Yuan, M. and Lin, Y. (2006), Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B*, **68**, 49–67.

[29] Zhong, W., Zhu, L, P., Li, R. and Cui, H. (2016), Regularized quantile regression and robust feature screening for single index models, *Statistica Sinica*, **26**, 69–95.
[30] Zhu, L. P., Huang, M. and Li, R. (2012), Semiparametric quantile regression with high-dimensional covariates, *Statistica Sinica*, **22**, 1379–1401. MR3027092
[31] Zhu, L. P. and Zhu, L. X. (2009), Nonconcave penalized inverse regression in single-index models with high dimensional predictors, *Journal of Multivariate Analysis*, **100**, 862–875. MR2498719
[32] Zou, H. and Hastie, T. (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, Series B,* **67**, 301–320. MR2137327

Wei Zhong
Wang Yanan Institute for Studies in Economics (WISE)
Department of Statistics
School of Economics
and Fujian Key Laboratory of Statistical Science
Xiamen University
China
E-mail address: wzhong@xmu.edu.cn

Xi Liu
Department of Statistics and Applied Probability
University of California
Santa Barbara, CA
USA
E-mail address: xiliu@umail.ucsb.edu

Shuangge Ma
Department of Biostatistics
Yale University
New Haven, CT
USA
Wang Yanan Institute for Studies in Economics (WISE)
Department of Statistics
School of Economics
Xiamen University
China
E-mail address: shuangge.ma@yale.edu

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

001 "Data Driven Value-at-Risk Forecasting using a SVR-GARCH-KDE Hybrid" by Marius Lux, Wolfgang Karl Härdle and Stefan Lessmann, January 2018.

002 "Nonparametric Variable Selection and Its Application to Additive Models" by Zheng-Hui Feng, Lu Lin, Ruo-Qing Zhu asnd Li-Xing Zhu, January 2018.

003 "Systemic Risk in Global Volatility Spillover Networks: Evidence from Option-implied Volatility Indices " by Zihui Yang and Yinggang Zhou, January 2018.

004 "Pricing Cryptocurrency options: the case of CRIX and Bitcoin" by Cathy YH Chen, Wolfgang Karl Härdle, Ai Jun Hou and Weining Wang, January 2018.

005 "Testing for bubbles in cryptocurrencies with time-varying volatility" by Christian M. Hafner, January 2018.

006 "A Note on Cryptocurrencies and Currency Competition" by Anna Almosova, January 2018.

007 "Knowing me, knowing you: inventor mobility and the formation of technology-oriented alliances" by Stefan Wagner and Martin C. Goossen, February 2018.

008 "A Monetary Model of Blockchain" by Anna Almosova, February 2018.

009 "Deregulated day-ahead electricity markets in Southeast Europe: Price forecasting and comparative structural analysis" by Antanina Hryshchuk, Stefan Lessmann, February 2018.

010 "How Sensitive are Tail-related Risk Measures in a Contamination Neighbourhood?" by Wolfgang Karl Härdle, Chengxiu Ling, February 2018.

011 "How to Measure a Performance of a Collaborative Research Centre" by Alona Zharova, Janine Tellinger-Rice, Wolfgang Karl Härdle, February 2018.

012 "Targeting customers for profit: An ensemble learning framework to support marketing decision making" by Stefan Lessmann, Kristof Coussement, Koen W. De Bock, Johannes Haupt, February 2018.

013 "Improving Crime Count Forecasts Using Twitter and Taxi Data" by Lara Vomfell, Wolfgang Karl Härdle, Stefan Lessmann, February 2018.

014 "Price Discovery on Bitcoin Markets" by Paolo Pagnottoni, Dirk G. Baur, Thomas Dimpfl, March 2018.

015 "Bitcoin is not the New Gold - A Comparison of Volatility, Correlation, and Portfolio Performance" by Tony Klein, Hien Pham Thu, Thomas Walther, March 2018.

016 "Time-varying Limit Order Book Networks" by Wolfgang Karl Härdle, Shi Chen, Chong Liang, Melanie Schienle, April 2018.

017 "Regularization Approach for NetworkModeling of German EnergyMarket" by Shi Chen, Wolfgang Karl Härdle, Brenda López Cabrera, May 2018.

018 "Adaptive Nonparametric Clustering" by Kirill Efimov, Larisa Adamyan, Vladimir Spokoiny, May 2018.

019 "Lasso, knockoff and Gaussian covariates: a comparison" by Laurie Davies, May 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

020 "A Regime Shift Model with Nonparametric Switching Mechanism" by Haiqiang Chen, Yingxing Li, Ming Lin and Yanli Zhu, May 2018.

021 "LASSO-Driven Inference in Time and Space" by Victor Chernozhukov, Wolfgang K. Härdle, Chen Huang, Weining Wang, June 2018.

022 " Learning from Errors: The case of monetary and fiscal policy regimes" by Andreas Tryphonides, June 2018.

023 "Textual Sentiment, Option Characteristics, and Stock Return Predictability" by Cathy Yi-Hsuan Chen, Matthias R. Fengler, Wolfgang Karl Härdle, Yanchu Liu, June 2018.

024 "Bootstrap Confidence Sets For Spectral Projectors Of Sample Covariance" by A. Naumov, V. Spokoiny, V. Ulyanov, June 2018.

025 "Construction of Non-asymptotic Confidence Sets in 2 -Wasserstein Space" by Johannes Ebert, Vladimir Spokoiny, Alexandra Suvorikova, June 2018.

026 "Large ball probabilities, Gaussian comparison and anti-concentration" by Friedrich Götze, Alexey Naumov, Vladimir Spokoiny, Vladimir Ulyanov, June 2018.

027 "Bayesian inference for spectral projectors of covariance matrix" by Igor Silin, Vladimir Spokoiny, June 2018.

028 "Toolbox: Gaussian comparison on Eucledian balls" by Andzhey Koziuk, Vladimir Spokoiny, June 2018.

029 "Pointwise adaptation via stagewise aggregation of local estimates for multiclass classification" by Nikita Puchkin, Vladimir Spokoiny, June 2018.

030 "Gaussian Process Forecast with multidimensional distributional entries" by Francois Bachoc, Alexandra Suvorikova, Jean-Michel Loubes, Vladimir Spokoiny, June 2018.

031 "Instrumental variables regression" by Andzhey Koziuk, Vladimir Spokoiny, June 2018.

032 "Understanding Latent Group Structure of Cryptocurrencies Market: A Dynamic Network Perspective" by Li Guo, Yubo Tao and Wolfgang Karl Härdle, July 2018.

033 "Optimal contracts under competition when uncertainty from adverse selection and moral hazard are present" by Natalie Packham, August 2018.

034 "A factor-model approach for correlation scenarios and correlation stress-testing" by Natalie Packham and Fabian Woebbeking, August 2018.

035 "Correlation Under Stress In Normal Variance Mixture Models" by Michael Kalkbrener and Natalie Packham, August 2018.

036 "Model risk of contingent claims" by Nils Detering and Natalie Packham, August 2018.

037 "Default probabilities and default correlations under stress" by Natalie Packham, Michael Kalkbrener and Ludger Overbeck, August 2018.

038 "Tail-Risk Protection Trading Strategies" by Natalie Packham, Jochen Papenbrock, Peter Schwendner and Fabian Woebbeking, August 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.

IRTG 1792 Discussion Paper Series 2018

For a complete list of Discussion Papers published, please visit irtg1792.hu-berlin.de.

039 "Penalized Adaptive Forecasting with Large Information Sets and Structural Changes" by Lenka Zbonakova, Xinjue Li and Wolfgang Karl Härdle, August 2018.

040 "Complete Convergence and Complete Moment Convergence for Maximal Weighted Sums of Extended Negatively Dependent Random Variables" by Ji Gao YAN, August 2018.

041 "On complete convergence in Marcinkiewicz-Zygmund type SLLN for random variables" by Anna Kuczmaszewska and Ji Gao YAN, August 2018.

042 "On Complete Convergence in Marcinkiewicz-Zygmund Type SLLN for END Random Variables and its Applications" by Ji Gao YAN, August 2018.

043 "Textual Sentiment and Sector specific reaction" by Elisabeth Bommes, Cathy Yi-Hsuan Chen and Wolfgang Karl Härdle, September 2018.

044 "Understanding Cryptocurrencies" by Wolfgang Karl Härdle, Campbell R. Harvey, Raphael C. G. Reule, September 2018.

045 "Predicative Ability of Similarity-based Futures Trading Strategies" by Hsin-Yu Chiu, Mi-Hsiu Chiang, Wei-Yu Kuo, September 2018.

046 "Forecasting the Term Structure of Option Implied Volatility: The Power of an Adaptive Method" by Ying Chen, Qian Han, Linlin Niu, September 2018.

047 "Inferences for a Partially Varying Coefficient Model With Endogenous Regressors" by Zongwu Cai, Ying Fang, Ming Lin, Jia Su, October 2018.

048 "A Regime Shift Model with Nonparametric Switching Mechanism" by Haiqiang Chen, Yingxing Li, Ming Lin, Yanli Zhu, October 2018.

049 "Strict Stationarity Testing and GLAD Estimation of Double Autoregressive Models" by Shaojun Guo, Dong Li, Muyi Li, October 2018.

050 "Variable selection and direction estimation for single-index models via DC-TGDR method" by Wei Zhong, Xi Liu, Shuangge Ma, October 2018.

IRTG 1792, Spandauer Straße 1, D-10178 Berlin
http://irtg1792.hu-berlin.de

This research was supported by the Deutsche Forschungsgemeinschaft through the IRTG 1792.