Statistics and Its Interface Volume 10 (2017) 407–412

# Efficient feature screening for ultrahigh-dimensional varying coefficient models

Xin Chen, Xuejun Ma, Xueqin Wang[*], and Jingxiao Zhang

Feature screening in ultrahigh-dimensional varying coefficient models is a crucial statistical problem in economics, genomics, etc. Current methods not only suffer from circumstances when the models involve multiple index variables or group predictor variables, but also cannot handle nonlinear varying coefficient models. To address these real-life scenarios efficiently, we develop a screening procedure for ultrahigh-dimensional varying coefficient models utilizing conditional distance covariance (CDC). Extensive simulation studies and two real economic data examples show the effectiveness and the flexibility of our proposed method.

AMS 2000 subject classifications: 62G08, 62G20, 62H20.

Keywords and phrases: Ultrahigh-dimensionality, Varying coefficient models, Multiple index variables, Group variables, Conditional distance covariance.

## 1. INTRODUCTION

The varying coefficient models are natural extensions of classical parametric models with good flexibility and interpretability, which ameliorates the “curse of the dimensionality”, and are applied in economics, epidemiology, medical science, ecology and so on (Fan and Zhang 2008). While ultrahigh-dimensional data are becoming increasingly popular in data science, most variable selection methods using penalization do not perform well for ultrahigh-dimensional varying coefficient models owning to the challenges of computational expediency, statistical accuracy, and algorithmic stability (Wang and Xia 2009, Tang et al. 2012). Fan and Lv (2008) novelly proposed sure independence screening (SIS) for ultrahigh-dimensional data in linear model. Since then, various versions of SIS methods have been developed, with ranges from linear models to nonlinear models, from specific models to model-free models, and from parametric models to nonparametric models. Among them, Hall and Miller (2009) introduced a generalized Pearson correlation to screen variables. Zhu et al. (2011), Li et al. (2012), Shao and Zhang (2014) developed sure independent ranking and screening (SIRS), sure independent screening based on distance correlation (DC-SIS) and sure independent screening based on martingale difference correlation (MDC-SIS) respectively. Fan et al. (2009, 2010, 2011) considered ultrahigh-dimensional generalized linear models and additive models. For varying coefficient models, Fan et al. (2014) introduced nonparametric independence screening (NIS) which used marginal mean regression and spline approximation. Liu et al. (2014) developed a conditional correlation sure independence screening (CC-SIS) based on conditional Pearson correlation. Song et al. (2014) studied the longitudinal data analysis by ranking the magnitude of spline approximations of the nonparametric components.

In many regression problems, some predictors can be naturally grouped, such as groups of dummy variables in multifactor analysis of variance. NIS and CC-SIS suffer in the case of group predictor variables. On the other hand, existing methods do not perform well in the case of multiple index variables, which are very important models (Lee et al. 2012 and Park et al. 2015). Moreover, current methods cannot handle nonlinear varying coefficient models which are possible in reality and illustrated in later section. To overcome these drawbacks in applications, we accordingly develop a screening procedure for ultrahigh-dimensional varying coefficient models utilizing conditional distance covariance (CDC). Due to the nature of conditional distance covariance, our method can work efficiently in almost all scenarios.

The paper is organized as follows. In Section 2, we present our general model framework, and then introduce our procedure named as CDCS. Section 3 illustrates the finite sample performance with both Monte Carlo simulations studies and two real data examples. The article concludes with a brief discussion in Section 4.

## 2. METHODOLOGY

### 2.1 Conditional distance correlation

Conditional distance correlation (Chen et al., 2015 and Wang et al., 2015 ) can measure both linear and nonlinear conditional correlations. Let $\boldsymbol{Y}$,$\boldsymbol{W}$ and $\boldsymbol{Z}$ be $q$, $d$ and $r$ dimensional random vectors in $\boldsymbol{R}^q$, $\boldsymbol{R}^d$ and $\boldsymbol{R}^r$, respectively. $\phi_{\boldsymbol{Y},\boldsymbol{W}|\boldsymbol{Z}}(\boldsymbol{t},\boldsymbol{s})$ is the conditional joint characteristic function of $\boldsymbol{Y}$,$\boldsymbol{W}$ given $\boldsymbol{Z}$ ($\boldsymbol{t} \in \boldsymbol{R}^q, \boldsymbol{s} \in \boldsymbol{R}^q$). Conditional distance covariation(CDcov) between $\boldsymbol{Y}$ and $\boldsymbol{W}$ with finite moments given $\boldsymbol{Z}$ is defined as square root of

[*]Corresponding author.

$$
\begin{aligned}
&CDcov^2(\boldsymbol{Y},\boldsymbol{W}|\boldsymbol{Z})\\
&=\|\phi_{\boldsymbol{Y},\boldsymbol{W}|\boldsymbol{Z}}(\boldsymbol{t},\boldsymbol{s})-\phi_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{t})\phi_{\boldsymbol{W}|\boldsymbol{Z}}(\boldsymbol{s})\|^2\\
&=\frac{1}{c_qc_d}\int_{\boldsymbol{R}^{q+d}}\frac{|\phi_{\boldsymbol{Y},\boldsymbol{W}|\boldsymbol{Z}}(\boldsymbol{t},\boldsymbol{s})-\phi_{\boldsymbol{Y}|\boldsymbol{Z}}(\boldsymbol{t})\phi_{\boldsymbol{W}|\boldsymbol{Z}}(\boldsymbol{s})|^2}{|\boldsymbol{t}|_q^{q+1}|\boldsymbol{s}|_d^{d+1}}d\boldsymbol{r}d\boldsymbol{s},
\end{aligned}
$$

where $c_q=\pi^{(q+1)/2}/\Gamma((q+1)/2)$, $c_d=\pi^{(d+1)/2}/\Gamma((d+1)/2)$. Further, let $CDcov^2(\boldsymbol{Y}|\boldsymbol{Z})=CDcov^2(\boldsymbol{Y},\boldsymbol{Y}|\boldsymbol{Z})$. As the standardization of CDcov, CDcor between $\boldsymbol{Y}$ and $\boldsymbol{W}$ with finite moments given $\boldsymbol{Z}$ is defined as square root of

$$CDcor^2(\boldsymbol{Y},\boldsymbol{W}|\boldsymbol{Z})=\frac{CDcov^2(\boldsymbol{Y},\boldsymbol{W}|\boldsymbol{Z})}{\sqrt{CDcov^2(\boldsymbol{Y}|\boldsymbol{Z})CDcov^2(\boldsymbol{W}|\boldsymbol{Z})}},$$

if $CDcov^2(\boldsymbol{Y}|\boldsymbol{Z})CDcov^2(\boldsymbol{W}|\boldsymbol{Z})>0$ and 0 otherwise.

Suppose that $\boldsymbol{G}_i=(\boldsymbol{Y}_i,\boldsymbol{W}_i,\boldsymbol{Z}_i), i=1,\dots,n$ are random samples from $\boldsymbol{G}=(\boldsymbol{Y},\boldsymbol{W},\boldsymbol{Z})$. Let $\omega_i(\boldsymbol{Z})=K_{\boldsymbol{H}}(\boldsymbol{Z}-\boldsymbol{Z}_i)$, $\omega(\boldsymbol{Z})=\sum_{i=1}^n\omega_i(\boldsymbol{Z})$, where $K(\bullet)$ is a kernel function such as the Gaussian kernel, $\boldsymbol{H}$ is the bandwidth. Denote the Euclidean distance of $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ as $d_{ij}^{\boldsymbol{X}}=d(\boldsymbol{X}_i,\boldsymbol{X}_j)$, and similarly, $d_{ij}^{\boldsymbol{Y}}$ for $\boldsymbol{Y}$. Let

$$d_{ijkl}^a=d_{ijkl}+d_{ijlk}+d_{ilkj}$$

where $d_{ijkl}=(d_{ij}^{\boldsymbol{X}}+d_{kl}^{\boldsymbol{X}}-d_{ik}^{\boldsymbol{X}}-d_{jl}^{\boldsymbol{X}})(d_{ij}^{\boldsymbol{Y}}+d_{kl}^{\boldsymbol{Y}}-d_{ik}^{\boldsymbol{Y}}-d_{jl}^{\boldsymbol{Y}})$. $CDcov$ can be estimated by:

$$\widehat{CDcov}^2(\boldsymbol{Y},\boldsymbol{W}|\boldsymbol{Z})=\frac{1}{C_n^4}\sum_{i<j<k<l}\psi_n(\boldsymbol{G}_i,\boldsymbol{G}_j,\boldsymbol{G}_k,\boldsymbol{G}_l|\boldsymbol{Z}).$$

where $\psi_n(\bullet)$ is the symmetric random kernel of degree 4 defined in Schick(1997)

$$\psi_n(\boldsymbol{G}_i,\boldsymbol{G}_j,\boldsymbol{G}_k,\boldsymbol{G}_l|\boldsymbol{Z})=\frac{n^4\omega_i(\boldsymbol{Z})\omega_j(\boldsymbol{Z})\omega_k(\boldsymbol{Z})\omega_l(\boldsymbol{Z})}{12\omega^4(\boldsymbol{Z})}d_{ijkl}^a$$

Further, we can get accordingly $\widehat{CDcor}^2(\boldsymbol{Y},\boldsymbol{X}|\boldsymbol{Z})$.

### 2.2 A new screening procedure

Let $Y$ be the response, $\boldsymbol{X}=(X_1,\dots,X_p)^\top$ be the $p$ dimensional predictor variables. We consider the following varying coefficient models:

$$Y=X_1\sum_{l=1}^q\beta_{1l}(U_l)+\cdots+X_p\sum_{l=1}^q\beta_{pl}(U_l)+\varepsilon. \qquad (1)$$

where $(\beta_{k1}(U_1),\dots,\beta_{kq}(U_q))^\top$ are $q$ dimensional unknown smooth functions. $\boldsymbol{U}=(U_1,\dots,U_q)^\top$ are the $q(<p)$ dimensional multiple index variables. Fan et al. (2014) and Liu et al. (2014) supposed $\boldsymbol{U}$ to be one dimension. In this paper, we relax this assumption.

Define the true model index set $\mathcal{D}$ and its complement $\mathcal{D}^c$ by $\mathcal{D}=\{1\le j\le p:\sum_{l=1}^q\beta_{kl}(u_l)\ne 0$ for some $\boldsymbol{u}\in\boldsymbol{U}\}$ and $\mathcal{D}^c=\{1\le j\le p:\sum_{l=1}^q\beta_{kl}(u_l)=0$ for all $\boldsymbol{u}\in\boldsymbol{U}\}$.

The goal is to select a reduced model with a moderate scale which can almost fully contain $\mathcal{D}$ for ultrahigh-dimensional varying coefficient models. Because given $\boldsymbol{u}$, the varying coefficient models become linear regression models (Liu et al.,2014). It is natural to apply the CDcor to screen variable in ultrahigh-dimensional varying coefficient models. For $\boldsymbol{X}_k$ and $Y$ given the $\boldsymbol{U}$, $w_k$ is defined as

$$w_k=E\{CDcor^2(Y,\boldsymbol{X}_k|\boldsymbol{U})\}.$$

For a random sample $\{(Y_i,\boldsymbol{X}_i,\boldsymbol{U}_i)\}_{i=1}^n$ from $(Y,\boldsymbol{X},\boldsymbol{U})^\top$, we can have the estimator:

$$\hat{w}_k=\frac{1}{n}\sum_{i=1}^n\widehat{CDcor}^2(Y_i,X_{ik}|\boldsymbol{U}_i), \qquad k=1,\dots,p.$$

We then sort the magnitudes of all the components of $\hat{w}=(\hat{w}_1,\dots,\hat{w}_p)^\top$ in a decreasing order and select a submodel:

$$\widehat{\mathcal{D}}=\{k:\hat{w}_k\ge cn^{-\kappa},1\le k\le p\},$$

where $c$ and $0<\kappa<1/2$ are prespecified values. In practice, for given $d$, one can select a submodel:

$$\widehat{\mathcal{D}}=\{1\le k\le p:\hat{w}_k \text{ is among the first } d \text{ largest of all }\}$$

## 3. NUMERICAL STUDIES AND APPLICATION

### 3.1 Simulation studies

In this section, we assess the finite sample performance of the proposed method in following three criteria:

(1) $\mathcal{S}$: the minimum model size to include all active predictors.
(2) $\mathcal{P}_k$: the proportion of all active predictors are selected for given model size $d$.
(3) $\mathcal{P}$: the proportion of an individual predictor is selected for given model size $d$.

We set $p$ to be 1000, and $n$ to be 200. $d_1=[n/\log(n)]$, $d_2=2d_1$ and $d_3=n-1$, where $[a]$ denotes the integer part of $a$. All the simulation results are based on 1000 replications. We consider varying coefficient models with single index variable in the Examples 1 and 2, multiple index variables in Example 3, group predictor variables in Example 4, and a more complex model in Example 5.

**Example 1.** We consider the following varying coefficient model:

$$Y=\beta_1(U)X_1+\beta_2(U)X_2+\beta_3(U)X_3+\varepsilon$$

where $\beta_1(U)=\exp(2U-1),\beta_2(U)=4U(U-1),\beta_3(U)=2\cos(2\pi U)$, $\boldsymbol{X}=(X_1,\dots,X_p)^\top\sim N(0,\boldsymbol{I}_p)$, where $\boldsymbol{I}_p$ is a $p\times p$ identity matrix. $U\sim U(0,1)$, and $\varepsilon$ is generated from the standard normal distribution.

*Table 1. The mean, 25%, 50%,75%,85% and 95% quantiles of $\mathcal{S}$ in Example 1*

| Methods | Mean (SD) | 25% | 50% | 75% | 85% | 95% |
|---|---|---|---|---|---|---|
| CDCS | 5.7 (19.5) | 3 | 3 | 3 | 4.2 | 13 |
| NIS | 18.1 (38.2) | 3 | 4 | 14 | 29 | 83.2 |
| CC-SIS | 37.5 (76.9) | 8 | 18 | 41 | 63.2 | 117 |

*Table 2. The proportions of $\mathcal{P}$ and $\mathcal{P}_k$ in Example 1*

| Methods | | $d_1$ | $d_2$ | $d_3$ |
|---|---|---|---|---|
| CDCS | $\mathcal{P}_1$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}_2$ | 0.988 | 0.991 | 0.999 |
| | $\mathcal{P}_3$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}$ | 0.988 | 0.991 | 0.999 |
| NIS | $\mathcal{P}_1$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}_2$ | 0.881 | 0.942 | 0.991 |
| | $\mathcal{P}_3$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}$ | 0.881 | 0.941 | 0.991 |
| CC-SIS | $\mathcal{P}_1$ | 0.999 | 0.999 | 0.999 |
| | $\mathcal{P}_2$ | 0.733 | 0.886 | 0.984 |
| | $\mathcal{P}_3$ | 0.998 | 0.998 | 0.998 |
| | $\mathcal{P}$ | 0.732 | 0.884 | 0.981 |

Tables 1 and 2 summarize the simulation results of Example 1. CDCS is significantly better than NIS and CC-SIS because the 75%, 85% and 95% quantiles of their $\mathcal{S}$ are much larger than CDCS. Almost all of $\mathcal{P}$ and $\mathcal{P}_k$ of CDCS in Table 2 are larger than those of NIS and CC-SIS.

**Example 2.** We consider the nonzero coefficient functions:

$$\beta_2(U) = 2I(U > 0.4), \beta_{100} = 1 + U, \beta_{400} = 3 - 3U,$$
$$\beta_{600}(U) = 1 + 2\sin(2\pi U), \beta_{1000} = \exp(U/(U + 1))$$

where $U$ and $\boldsymbol{X}$ are generated as follows: first, take samples of $U^*$ and $\boldsymbol{X}$ from $(U^*, \boldsymbol{X}) \sim N(0, \Sigma)$, and $\Sigma = (\sigma_{ij})_{(p+1)\times(p+1)}$ and $\sigma_{ij} = 0.5^{|i-j|}$. Then take $U = \Phi(U^*)$, where $\Phi$ is the cumulative distribution function of the standard normal distribution. Thus, $U$ follows a uniform distribution in $(0, 1)$ and is correlated with $\boldsymbol{X}$, and all the predictor variables $X_1, \ldots, X_p$ are correlated with each other.

From Tables 3 and 4, we can see that CDCS again performs better than NIS and CC-SIS.

*Table 3. The mean, 25%, 50%,75%,85% and 95% quantiles of $\mathcal{S}$ in Example 2*

| Methods | Mean (SD) | 25% | 50% | 75% | 85% | 95% |
|---|---|---|---|---|---|---|
| CDCS | 6.1 (3.9) | 5 | 5 | 6 | 6 | 9 |
| NIS | 7.7 (11.5) | 5 | 5 | 7 | 9 | 14.1 |
| CC-SIS | 36.5 (86) | 12 | 20 | 37 | 50.2 | 86.1 |

*Table 4. The proportions of $\mathcal{P}$ and $\mathcal{P}_k$ in Example 2*

| Methods | | $d_1$ | $d_2$ | $d_3$ |
|---|---|---|---|---|
| CDCS | $\mathcal{P}_2$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}_{100}$ | 0.999 | 1.000 | 1.000 |
| | $\mathcal{P}_{400}$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}_{600}$ | 0.999 | 1.000 | 1.000 |
| | $\mathcal{P}_{1000}$ | 0.999 | 1.000 | 1.000 |
| | $\mathcal{P}$ | 0.997 | 1.000 | 1.000 |
| NIS | $\mathcal{P}_2$ | 0.999 | 1.000 | 1.000 |
| | $\mathcal{P}_{100}$ | 0.996 | 0.999 | 1.000 |
| | $\mathcal{P}_{400}$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}_{600}$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}_{1000}$ | 0.991 | 0.998 | 0.999 |
| | $\mathcal{P}$ | 0.986 | 0.997 | 0.999 |
| CC-SIS | $\mathcal{P}_2$ | 0.962 | 0.991 | 0.999 |
| | $\mathcal{P}_{100}$ | 0.952 | 0.989 | 0.997 |
| | $\mathcal{P}_{400}$ | 0.978 | 0.997 | 0.999 |
| | $\mathcal{P}_{600}$ | 0.970 | 0.994 | 0.999 |
| | $\mathcal{P}_{1000}$ | 0.855 | 0.960 | 0.992 |
| | $\mathcal{P}$ | 0.763 | 0.934 | 0.987 |

**Example 3.** In this example, we investigate the performance of models with multiple index variables. As Lee et al. (2012) and Park et al. (2015), we consider two index variables:

$$\begin{aligned} \beta_{11}(U_1) &= 1 + U_1^2 & \beta_{12}(U_2) &= 4(U_2 - 0.5)^2 \\ \beta_{21}(U_1) &= U_1 & \beta_{22}(U_2) &= 2\cos(2\pi U_2) \\ \beta_{31}(U_1) &= \exp(2U_1 - 1) & \beta_{32}(U_2) &= \sin(2\pi U_2) \end{aligned}$$

where $(U_1, U_2)$ are iid uniform distribution $U(0, 1)$. Since NIS and CC-SIS do not work in the case of multiple index variables, we compare the performance of the proposed method (CDCS) with three existing methods, SIRS (Zhu et al., 2011), DC-SIS (Li et al., 2012) and MDC-SIS (Shao and Zhang 2014), which are model-free feature screening procedures. The other settings are the same as Example 1.

Tables 5 and 6 show that CDCS is superior to SIRS, DC-SIS and MDC-SIS. It is because the latters only use the information of $Y$ and $\boldsymbol{X}$, not $\boldsymbol{U}$.

**Example 4.** This example is designed to illustrate the performance in the case of group independent predictor variables. We consider the following model:

$$\begin{aligned} Y = {} & \beta_1(U)X_1 + \beta_{12}^{(1)}(U)\mathbf{1}(X_{12} < q_1) \\ & + \beta_{12}^{(2)}(U)\mathbf{1}(q_1 \le X_{12} < q_2) + \beta_{22}(U)X_{22} + \varepsilon \end{aligned}$$

where $\beta_1(U) = 4U(U - 1)$, $\beta_{12}^{(1)}(U) = \exp(U/(U + 1))$, $\beta_{12}^{(2)}(U) = 3U$, $\beta_{22}(U) = 2\sin(2\pi U)$, $q_1$ and $q_2$ are the 30% and 60% quantiles of $X_{12}$. $(X_1, \ldots, X_p) \sim N(0, \mathbf{\Sigma})$, and $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$ and $\sigma_{ij} = 0.5^{|i-j|}$. The other settings are the same as Example 1. We write

$$\tilde{X}_{12} = \{\mathbf{1}(X_{12} < q_1), \mathbf{1}(q_1 \le X_{12} < q_2)\}^T.$$

Then these two indicator variables become a group. The predictor variable vector is $\mathbf{X} = (X_1, \ldots, X_{11}, \tilde{X}_{12}, X_{13}, \ldots, X_p) \in R^{p+1}$.

*Table 5. The mean, 25%, 50%, 75%, 85% and 95% quantiles of $\mathcal{S}$ in Example 3*

| Methods | Mean (SD) | 25% | 50% | 75% | 85% | 95% |
|---|---|---|---|---|---|---|
| CDCS | 3 (0.6) | 3 | 3 | 3 | 3 | 3 |
| SIRS | 192 (242) | 13 | 73 | 304 | 488 | 728 |
| DC-SIS | 93 (152) | 6 | 22.5 | 102 | 205 | 439 |
| MDC-SIS | 103 (185) | 4 | 18 | 100 | 233 | 547 |

*Table 6. The proportions of $\mathcal{P}$ and $\mathcal{P}_k$ in Example 3*

| Methods | | $d_1$ | $d_2$ | $d_3$ |
|---|---|---|---|---|
| CDCS | $\mathcal{P}_1$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}_2$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}_3$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}$ | 1.000 | 1.000 | 1.000 |
| SIRS | $\mathcal{P}_1$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}_2$ | 0.384 | 0.505 | 0.676 |
| | $\mathcal{P}_3$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}$ | 0.384 | 0.505 | 0.676 |
| DC-SIS | $\mathcal{P}_1$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}_2$ | 0.581 | 0.694 | 0.847 |
| | $\mathcal{P}_3$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}$ | 0.581 | 0.694 | 0.847 |
| MDC-SIS | $\mathcal{P}_1$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}_2$ | 0.624 | 0.715 | 0.831 |
| | $\mathcal{P}_3$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}$ | 0.624 | 0.715 | 0.831 |

*Table 7. The mean, 25%, 50%,75%,85% and 95% quantiles of $\mathcal{S}$ in Example 4*

| Methods | Mean (SD) | 25% | 50% | 75% | 85% | 95% |
|---|---|---|---|---|---|---|
| CDCS | 9.8 (34.2) | 3 | 4 | 5 | 7 | 22.1 |
| DC-SIS | 77.4 (98.4) | 19.8 | 46 | 93 | 138 | 249 |

*Table 8. The proportions of $\mathcal{P}$ and $\mathcal{P}_k$ in Example 4*

| Methods | | $d_1$ | $d_2$ | $d_3$ |
|---|---|---|---|---|
| CDCS | $\mathcal{P}_1$ | 0.990 | 0.994 | 0.999 |
| | $\mathcal{P}_{12}$ | 0.979 | 0.988 | 0.991 |
| | $\mathcal{P}_{22}$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}$ | 0.970 | 0.982 | 0.990 |
| DC-SIS | $\mathcal{P}_1$ | 0.994 | 0.997 | 1.000 |
| | $\mathcal{P}_{12}$ | 0.989 | 0.995 | 0.997 |
| | $\mathcal{P}_{22}$ | 0.455 | 0.684 | 0.920 |
| | $\mathcal{P}$ | 0.455 | 0.678 | 0.918 |

Since NIS and CC-SIS cannot be applied for screening group variables, we compare the performance of the proposed method (CDCS) with DC-SIS of Li et al. (2012). From Tables 7 and 8, we can see that CDCS is superior to DC-SIS. The 95% quantile of $\mathcal{S}$ of DC-SIS is 10 times more than that of CDCS. It implies CDCS is capable of handling group variable.

*Table 9. The mean, 25%, 50%,75%,85% and 95% quantiles of $\mathcal{S}$ in Example 5*

| Methods | Mean (SD) | 25% | 50% | 75% | 85% | 95% |
|---|---|---|---|---|---|---|
| CDCS | 4 (5.5) | 3 | 3 | 3 | 4 | 8 |
| NIS | 161 (217) | 16 | 62 | 215 | 356 | 677 |
| CC-SIS | 207 (222) | 51 | 122 | 275 | 409 | 730 |

*Table 10. The proportions of $\mathcal{P}$ and $\mathcal{P}_k$ in Example 5*

| Methods | | $d_1$ | $d_2$ | $d_3$ |
|---|---|---|---|---|
| CDCS | $\mathcal{P}_1$ | 1.000 | 1.000 | 1.000 |
| | $\mathcal{P}_2$ | 0.999 | 1.000 | 1.000 |
| | $\mathcal{P}_3$ | 0.998 | 1.000 | 1.000 |
| | $\mathcal{P}$ | 0.997 | 0.999 | 1.000 |
| NIS | $\mathcal{P}_1$ | 0.514 | 0.605 | 0.756 |
| | $\mathcal{P}_2$ | 0.906 | 0.960 | 0.995 |
| | $\mathcal{P}_3$ | 0.871 | 0.941 | 0.980 |
| | $\mathcal{P}$ | 0.397 | 0.541 | 0.734 |
| CC-SIS | $\mathcal{P}_1$ | 0.303 | 0.436 | 0.680 |
| | $\mathcal{P}_2$ | 0.797 | 0.915 | 0.987 |
| | $\mathcal{P}_3$ | 0.789 | 0.906 | 0.980 |
| | $\mathcal{P}$ | 0.178 | 0.358 | 0.657 |

Given index variables, the varying coefficient models (1) become linear regression models (Liu et al., 2014). However, in reality, we may have the following extended varying coefficient models (EVCM):

$$Y = f_1(X_1)\sum_{l=1}^{q}\beta_{1l}(U_l) + \cdots + f_p(X_p)\sum_{l=1}^{q}\beta_{pl}(U_l) + \varepsilon. \quad (2)$$

where $f_i(\cdot)$, $i = 1, \cdots, p$, are unknown functions. EVCM is a potential nonlinear varying coefficient models.

**Example 5.** We now consider a simple EVCM that contains a quadratic form of the first variable:

$$Y = \beta_1(U)X_1^2 + \beta_2(U)X_2 + \beta_3(U)X_3 + \varepsilon$$

where $\beta_1(U) = 2U$, $\beta_2(U) = 4U(1-U)$, $\beta_3(U) = \sin(2\pi U)$. The other settings are the same as Example 1. Note given $\boldsymbol{u}$, the relationship between $Y$ and $X_1$ is not linear, but nonlinear.

From Tables 9 and 10, we can see that NIS and CC-SIS do not work well. It is because NIS cannot estimate $\beta_1(U)$ well using B spline given $f_1(X_1) = X_1^2$, and CC-SIS only describes the conditional linear relationship.

## 3.2 Real data analysis

### 3.2.1 China economy data analysis

We apply CDCS to the China economy data which is collected from 1987 to 2012 and contains five variables: GDP (Gross Domestic Product), R&D (Research and Development), VTCI (Value of Technology Contracts Imported),

*Table 11. The mean, 50%,75%,85% and 95% quantiles of $\mathcal{S}$ in China economy data*

| Methods | Mean (SD) | 25% | 50% | 75% | 85% | 95% |
|---|---|---|---|---|---|---|
| CDCS | 6.2 (2.1) | 5 | 5 | 6.3 | 8 | 10 |
| NIS | 384 (186) | 242 | 354 | 503 | 596 | 740 |
| CC-SIS | 330 (203) | 168 | 308 | 484 | 570 | 662 |

CS (Capital Stock), EMP (Employment) and Year. GDP are commonly used to measure the economic performance of a whole country or region. R&D stands for independent innovation. The larger the R&D is, the stronger the innovation is. VTCI represents the usage of foreign technology. CS is obtained by perpetual inventory method (Goldsmith, 1951). EMP stands for the labour utilization. We take GDP as the response, Year as the index variable and R&D, VTCI, CS and EMP as predictor variables, denoted by $X_1, \ldots, X_4$. Variables $X_1, \ldots, X_4$ are known to be important in contributing to GDP. The response and predictor variables are transformed to have zero mean and unit stand deviation, and the index variable is transformed into the range of $[0, 1]$.

In order to evaluate the performances of CDCS, NIS and CC-SIS, we generate new variables $(X_1^*, \ldots, X_4^*)$ as follows: each time 20 samples are resampled from 26 original data points without replacement. $(X_5^*, \ldots, X_{1000}^*)$ are generated from the $N(0, \boldsymbol{I}_{996})$. We set $\mathcal{M} = \{1, 2, 3, 4\}$, and replicate for 1000 times.

From Table 11, we can see that the CDCS works very well in screening out redundant predictors since 95% quantile of $\mathcal{S}$ is very small, while NIS and CC-SIS do not perform well since the quantiles of $\mathcal{S}$ are very large, at least 50 times larger than CDCS.

### 3.2.2 Boston housing data analysis

We also apply our method to Boston housing data that concerns the median value of owner-occupied homes for 506 census tracts of Boston from the 1970 census, which can be found from BostonHousing in R pacakge inmlbench. Following Wang and Xia (2009), we take MEDV (median value of owner-occupied homes in USD 1000's) as the response $Y$, LSTAT (percentage of lower status of the population) as the index variable $U$, and the following predictors as the predictor variables: CRIM (per capita crime rate by town), RM (average number of rooms per dwelling), PTRATION (pupil-teacher ratio by town), NOX (nitric oxides concentration (parts per 10 million)), TAX (full-value property tax rate per USD 10,000), AGE (proportion of owner-occupied units built prior to 1940), denoted by $X_1, \ldots, X_6$. Before applying our method, both the response and the predictor variables (except for INT) are transformed to have zero mean and unit variance. The index variable LSTAT is transformed so that its marginal distribution is in [0, 1].

*Table 12. The mean, 50%,75%,85% and 95% quantiles of $\mathcal{S}$ in Boston housing data*

| Methods | Mean (SD) | 25% | 50% | 75% | 85% | 95% |
|---|---|---|---|---|---|---|
| CDCS | 3.9 (1.0) | 4 | 4 | 4 | 5 | 5 |
| NIS | 43.8 (63.8) | 8 | 19 | 51 | 80 | 171.3 |

Wang and Xia (2009) use LASSO to select variables and claim that $X_1, X_2, X_3$ are significant predictors, whose corresponding coefficients are nonzero. We only compare the performances with NIS because the bandwidth selection of CC-SIS using plug-in method seems not to be working for this dataset. We generate new variables $(X_1^*, \ldots, X_6^*)$ as follows: 200 samples are resampled from those 506 real data without replacement. $(X_7^*, \ldots, X_{1000}^*)$ are generated from the $N(0, \boldsymbol{I}_{994})$. We set $\mathcal{M} = \{1, 2, 3\}$, and replicate for 1000 times.

From Table 12, we can see that the CDCS performs much better than NIS because the sizes of CDCS are much smaller compared with NIS.

## 4. DISCUSSION

In the paper, we propose a new feature screening for ultrahigh-dimensional varying coefficient models based on conditional distance covariance. The numerical studies and real data analysis show CDCS performs much better than competitors NIS and CC-SIS. The underlying reason is that the nature of conditional distance covariance which can efficiently handle group variable, multiple index variable and non-linearity. The selection of bandwidth is challenging. The larger bandwidth is, the smoother our estimator becomes. We compare three methods in selecting the bandwidth: plug-in (Wand and Jones, 1994), smoothed cross-validation (Jones, Marron and Park, 1991) and least-squares cross-validation (Bowman, 1984), and find that CDCS is insensitive to these bandwidth selection methods. In our implementation, we use plug-in method as it is easy to calculate and performs very well in Examples 1–5.

## ACKNOWLEDGMENTS

The authors would like to thank the Editor, the Associate Editor and two Referees for their constructive comments which greatly improved the paper. Chen's research was supported by tier 1 funding of National University of Singapore (WBS no. R155000155112). Ma's research was supported by the program of China Statistical Research (No. 2015LZ42). Wang's research was partially supported by National Science Foundation of China(NSFC) for Excellent Young Scholar (11322108) and NSFC(11271383).

*Received 15 February 2016*

## REFERENCES

[1] Bowman A. (1984). An Alternative Method of Cross-validation for the Smoothing of Kernel Density Estimates. *Biometrika* **71** 353–360. MR0767163

[2] Chen X., Cook R., Zou C. (2015). Diagnostic Studies in Sufficient Dimension Reduction. *Biometrika* **102** 545–558. MR3394274

[3] Fan J., Feng Y., Song R. (2011) Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models. *Journal of the American Statistical Association* **106** 544–557. MR2847969

[4] Fan J., Lv J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space. *Journal of the Royal Statistical Society, Ser. B* **70** 849–911. MR2530322

[5] Fan J., Ma Y., Dai W. (2014) Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Varying Coefficient Models. *Journal of the American Statistical Association* **109** 1270–1284. MR3265696

[6] Fan J., Samworth R., Wu Y. (2009) Ultrahigh Dimensional Feature Selection: Beyond the Linear Model. *Journal of Machine Learning Research* **10** 2013–2038. MR2550099

[7] Fan J., Song R. (2010) Sure Independence Screening in Generalized Linear Models with NP-Dimensionality. *The Annals of Statistics* **38** 3567–3604. MR2766861

[8] Fan J., Zhang W. (2008) Statistical Methods with Varying Coefficient Models. *Statistics and Its Interface* **1** 179–195. MR2425354

[9] Hall P., Miller H. (2009) Using Generalized Correlation to Effect Variable Selection in Very High Dimensional Problems. *Journal of Computational and Graphical Statistics* 2009, **18** 533–550. MR2751640

[10] Goldsmith W. (1951). A Perpetual Inventory of National Wealth. *Studies in Income and Wealth* **14** New York: NBER, 1951.

[11] Jones M., Marron J., Park B. (1991). A Simple Root $n$ Bandwidth Selector. *The Annals of Statistics* **19** 1919–1932. MR1135156

[12] Lee Y., Mammen E., Park B. (2012). Flexible Generalized Varying Coefficient Regression Models. *The Annals of Statistics* **40** 1906–1933. MR3015048

[13] Li R., Wei Z., Zhu L. (2012). Feature Screening via Distance Correlation Learning. *Journal of the American Statistical Association* **107** 1129–1139. MR3010900

[14] Liu J., Li R., Wu S. (2014). Feature Selection for Varying Coefficient Models with Ultrahigh-Dimensional Covariates. *Journal of the American Statistical Association* **109** 266–274. MR3180562

[15] Park U., Mammen E., Lee K., Lee R. (2015). Varying Coefficient Regression Models: A Review and New Developments. *International Statistical Review* **83** 36–64. MR3341079

[16] Schick A. (1997). On U-Statistics with Random Kernels. *Statistics & Probability Letters* **34** 275–283 MR1458022

[17] Shao X., Zhang J. (2014). Martingale Difference Correlation and Its Use in High-Dimensional Variable Screening. *Journal of the American Statistical Association* **109** 1302–1318. MR3265698

[18] Song R., Yi F., Zou H. (2014) On Varying-Coefficient Independence Screening for High-Dimensional Varying-Coefficient Models. *Statistica Sinica* **24** 1735–1752. MR3308660

[19] Tang Y., Wang H., Zhu Z., Song X. (2012) A Unified Variable Selection Approach for Varying Coefficient Models. *Statistica Sinica* **22** 601–628. MR2954354

[20] Wand M., Jones M. (1994) Multivariate Plug–in Bandwidth Selection. *Computational Statistics* **9** 97–116. MR1280754

[21] Wang H., Xia Y. (2009) Shrinkage Estimation of the Varying Coefficient Model. *Journal of the American Statistical Association* **104** 747–757. MR2541592

[22] Wang X., Pan W., Hua W., Tian Y., Zhang H. (2015) Conditional Distance Correlation. *Journal of the American Statistical Association* **110** 1726–1734. MR3449068

[23] Zhu L., Li L., Li R., Zhu L. (2011). Model-Free Feature Screening for Ultrahigh Dimensional Data. *Journal of the American Statistical Association* **106** 1464–1475. MR2896849

Xin Chen
Department of Statistics and Applied Probability
National University of Singapore
117546
Singapore
E-mail address: stacx@nus.edu.sg

Xuejun Ma
College of Applied Sciences
Beijing University of Technology
Beijing, 100124
China
E-mail address: yinuoyumi@163.com

Xueqin Wang
Southern China Center for Statistical Science
School of Mathematics
Zhongshan School of Medicine
Xinhua College
Sun Yat-Sen University
Guangzhou, 510275
China
E-mail address: wangxq88@mail.sysu.edu.cn

Jingxiao Zhang
Center for Applied Statistics
School of Statistics
Renmin University of China
Beijing, 100872
China
E-mail address: zhjxiaoruc@163.com