Statistica Sinica Preprint No: SS-2018-0176	
Title	A Lack-Of-Fit Test with Screening in Sufficient Dimension Reduction
Manuscript ID	SS-2018-0176
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202018.0176
Complete List of Authors	Yaowu Zhang Wei Zhong and Liping Zhu
Corresponding Author	Liping Zhu
E-mail	zhulp1@hotmail.com
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Statistica Sinica

A Lack-Of-Fit Test with Screening in Sufficient Dimension Reduction

Yaowu Zhang[1], Wei Zhong[2] and Liping Zhu[3]

Shanghai University of Finance and Economics[1],

Xiamen University[2] *and Renmin University of China*[3]

Abstract: It is of fundamental importance to infer how the conditional mean of the response varies with the predictors. Sufficient dimension reduction techniques reduce the dimension by identifying a minimal set of linear combinations of the original predictors without loss of information. This paper is concerned with testing whether a given small number of linear combinations of the original ultrahigh dimensional covariates is sufficient to characterize the conditional mean of the response. We first introduce a novel consistent lack-of-fit test statistic when the dimensionality of covariates is moderate. The proposed test is shown to be n-consistent under the null hypothesis and root-n-consistent under the alternative hypothesis. A bootstrap procedure is also developed to approximate p-values and its consistency has been theoretically studied. To deal with ultrahigh dimensionality, we introduce a two-stage lack-of-fit test with screening (LOFTS) procedure based on data splitting strategy. The data are randomly partitioned into two equal halves. In the first stage, we apply the martingale difference correlation based screening to one half of the data and select a moderate set of covariates.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

In the second stage, we perform the proposed test based on the selected covariates using the second half of the data. The data splitting strategy is crucial to eliminate the effect of spurious correlations and avoid the inflation of Type-I error rates. We also demonstrate the effectiveness of our two-stage test procedure through comprehensive simulations and two real-data applications.

Key words and phrases: Bootstrap; Central mean subspace, Data splitting, Lack-of-fit test, Sufficient dimension reduction, Variable Selection.

1. Introduction

Let $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ be a covariate vector and $\mathbf{y} = (Y_1, \ldots, Y_q)^{\mathrm{T}} \in \mathbb{R}^q$ be a response vector. It is of fundamental importance to infer how the conditional mean of $\mathbf{y}$ varies with the predictors. Sufficient dimension reduction techniques have become important and useful in high dimensional data analysis. It is aimed to identify a few linear combinations of the original high dimensional covariates while retaining all the information about $E(\mathbf{y} \mid \mathbf{x})$. Cook and Li (2002) assumed that there exists a $p \times d_0$ matrix $\boldsymbol{\beta}$ such that

$$E(\mathbf{y} \mid \mathbf{x}) = E(\mathbf{y} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}), \tag{1.1}$$

which implies that the conditional mean function $E(\mathbf{y} \mid \mathbf{x})$ depends on $\mathbf{x}$ only through d_0 linear combinations $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$. This model not only retains the

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

flexibility of nonparametric modeling but also enjoys the interpretability of parametric modeling. Since $\boldsymbol{\beta}$ is not identifiable, Cook and Li (2002) defined the Central Mean Subspace (CMS), denoted by $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$, as the smallest column space of $\boldsymbol{\beta}$, where the corresponding smallest column numbers, denoted by d_0, is defined as the structural dimension. To recover $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$, Li and Duan (1989) suggested to use the ordinary least squares estimator when $\mathbf{x}$ follows an elliptical distribution particularly when $d_0 = 1$. Cook and Li (2002) proved that the column space of $\{\text{var}(\mathbf{x})\}^{-1}\text{cov}(\mathbf{x}, \mathbf{y})$ belongs to the CMS $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$ when $\mathbf{x}$ satisfies the linearity condition. Xia, et. al (2002) proposed a minimum average variance estimation (MAVE) when the covariates are continuous. Ma and Zhu (2012) developed a semiparametric approach to dimension reduction. Ma and Zhu (2014) further investigated the inference and estimation efficiency of the central mean subspace for sufficient dimension reduction. Zhu and Zhong (2015) considered estimation of the CMS for multivariate response data. One can refer to Ma and Zhu (2013a) for a comprehensive review on dimension reduction.

Most work in the dimension reduction literature focused on estimation of the central mean subspace. However, model diagnostic studies have not received much attention within the context of dimension reduction. It is fundamental to study whether a given small number of linear combinations

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

of the original high dimensional covariates is sufficient to characterize the conditional mean of $\mathbf{y}$. That is, we test the following null hypothesis, for a given $d_0 \geq 1$,

$$H_0: \quad E(\mathbf{y} \mid \mathbf{x}) = E(\mathbf{y} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}), \text{ for some } p \times d_0 \text{ matrix } \boldsymbol{\beta}. \tag{1.2}$$

Some efforts have been devoted to model checking. For example, Stute and Zhu (1998) studied nonparametric tests for the validity of generalized linear models with a given parametric link structure based on certain empirical processes marked by the residuals. Xia, et. al (2004) considered model-checking for single-index models. Verzelen and Villers (2010) proposed a new goodness-of-fit test for high dimensional Gaussian Linear models based on the Fisher statistic. Guo, Wang and Zhu (2016) introduced a model-adaptation concept in lack-of-fit testing and proposed a dimension-reduction model-adaptive (DRMA) test for checking parametric single-index models. Shah and Buhlmann (2018) developed Residual Prediction goodness-of-fit tests to assess the validity of high dimensional linear models. For the choice of structural dimension d_0, Cook and Li (2004) provided a sequential test procedure. Zhu, Yu and Zhu (2010) proposed a sparse eigendecomposition strategy by introducing an ℓ_1 penalty to shrink small sample eigenvalues to zero. Ma and Zhang (2015) considered an information criterion-based method to determine the structural dimension

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

of the reduction model. However, the challenges associated with designing a general test for (1.2) especially for ultrahigh dimensional covariates are not to be addressed.

For ultrahigh dimensional data where the number of covariates is much higher than the sample size, the aforementioned dimension reduction methods do not work because their asymptotic normality results may require the dimensionality divergence rate satisfy $p = o(n^{1/3})$ (Zhu, Zhu and Feng , 2010). In addition, as pointed by Zhang, Yao and Shao (2018), the testing problem such as $H_0 : E(\varepsilon \mid \mathbf{x}) = 0$ almost surely without assuming a parametric model is very challenging since we are targeting a general class of alternative, and the power may decrease quickly owing to the growing dimension and nonlinear dependence. It is natural and crucial to assume the sparsity principle that only a small set of covariates, denoted by $\mathcal{A}$, truly contributes to the response. Let $\mathbf{x}_{\mathcal{A}} = \{X_k, k \in \mathcal{A}\}$ stand for the covariates indexed by $\mathcal{A}$. Under the sparsity assumption, the null hypothesis (1.2) can be written as

$$H_0 : \quad E(\mathbf{y} \mid \mathbf{x}) = E(\mathbf{y} \mid \boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}} \mathbf{x}_{\mathcal{A}}), \text{ for some } |\mathcal{A}| \times d_0 \text{ matrix } \boldsymbol{\beta}_{\mathcal{A}}, \quad (1.3)$$

where $|\mathcal{A}|$ represents the cardinality of $\mathcal{A}$. Without loss of generality, we assume $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}}^{\mathrm{T}}, \mathbf{0}_{d_0 \times (p-|\mathcal{A}|)})^{\mathrm{T}}$ where $\mathbf{0}_{d_0 \times (p-|\mathcal{A}|)}$ denotes a $d_0 \times (p - |\mathcal{A}|)$ matrix of zeros. However, $\mathcal{A}$ is generally unknown. Sure independence

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

screening approaches (Fan and Lv , 2008; Zhu, et. al , 2011; Li, Zhong and Zhu , 2012) have been developed to screen out irrelevant covariates and estimate $\mathcal{A}$ for ultrahigh dimensional data. Refer to Liu, Zhong and Li (2015) for a review on variable screening. In particular, Shao and Zhang (2014) proposed a martingale difference correlation (MDC) which imposes few parametric assumptions on the mean regression form $E(\mathbf{y} \mid \mathbf{x})$ and retains the model-free flavor of sufficient dimension reduction.

In this paper, we first assume that there exists a surrogate index set $\mathcal{S}$ with a moderate size such that $\mathcal{A} \subseteq \mathcal{S}$ and develop a novel consistent lack-of-fit test statistic for (1.3) based on the moderate covariates set $\mathcal{S}$. We demonstrate that the hypothesis based on $\mathcal{S}$ is equivalent to (1.3) as long as $\mathcal{A} \subseteq \mathcal{S}$ in Theorem 1. The proposed test is shown to be n-consistent under the null hypothesis and root-n-consistent under the alternative hypothesis. We suggest a bootstrap procedure to approximate the p-values and theoretically show that the bootstrap procedure is consistent. The second goal is to introduce a new two-stage approach based on data random splitting strategy for testing (1.2) when the dimensionality of covariates is ultrahigh. To be specific, we first randomly partition data into two equal halves. In the first stage, we apply the MDC-screening to one half of the data and select a moderate set of covariates to estimate the index set $\mathcal{S}$. In the second

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

stage, we perform the proposed test for (1.2) based on the selected set in the second half of the data. Note that the data split strategy is crucial to eliminate the effect of spurious correlations and avoid the inflation of Type-I error rates of the test. Furthermore, to avoid potential type-I error rate inflation when some important covariates are missed with a nonignorable probability, we also provide a multi-split strategy in the extension.

The rest of this paper is organized as follows. Section 2 introduces the details of the two-stage test procedure. In Section 3, we study the theoretical justification for the test procedure. Section 4 demonstrates the finite-sample performance through comprehensive simulations and two real data applications. We add an extension on multi-splitting strategy in Section 5. All technical proofs are relegated to the supplemental material.

A word on notation. Let $\mathbf{x}_{\mathcal{S}}$ be the covariate vector indexed by $\mathcal{S}$, $|c|$ be the absolute value of a generic constant c. For a complex-valued function ψ, $\|\psi\|^2 = \psi^{\mathrm{T}}\overline{\psi}$ and $\overline{\psi}$ is the conjugate of ψ, and for a matrix $\boldsymbol{\beta} \in \mathbb{R}^{p\times d_0}$, $\|\boldsymbol{\beta}\| = \{\mathrm{tr}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta})\}^{1/2}$. In addition, $\mathrm{span}(\boldsymbol{\beta})$ denotes the column space of $\boldsymbol{\beta}$, $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$ and $\mathcal{S}_{E(\mathbf{y}|\mathbf{x}_{\mathcal{S}})}$ denote the central mean subspace of $\mathbf{y}$ given $\mathbf{x}$ and the central mean subspace of $\mathbf{y}$ given $\mathbf{x}_{\mathcal{S}}$, respectively. The sign $\xrightarrow{D}$ denotes convergence in distribution.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

2. A New Testing Procedure

In this section, we first propose a lack-of-fit test statistic in the population level based on a surrogate index set $\mathcal{S}$ with a moderate size such that $\mathcal{A} \subseteq \mathcal{S}$. Then, we estimate the test statistic and further develop a two-stage lack-of-fit test with screening procedure.

2.1 A Lack-of-Fit Test Statistic

Under the sparsity assumption, this hypothesis can be formulated as (1.3) where $\mathcal{A}$ represents the index set of covariates which truly contributes to the response. However, $\mathcal{A}$ is generally unknown which makes it practically infeasible to directly propose a test for (1.3). To deal with this issue, we first suppose that these exists a surrogate index set $\mathcal{S}$ with a moderate size which satisfies that $\mathcal{A} \subseteq \mathcal{S}$. Then, we consider the following null hypothesis

$$H_0 : \quad E(\mathbf{y} \mid \mathbf{x}) = E(\mathbf{y} \mid \boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}} \mathbf{x}_{\mathcal{S}}), \text{ for some } |\mathcal{S}| \times d_0 \text{ matrix } \boldsymbol{\beta}_{\mathcal{S}}. \tag{2.1}$$

The natural question then arises: whether testing (2.1) is equivalent to testing (1.3)? The following theorem answers this question.

Theorem 1. *In addition to the sparsity assumption, we assume that both* $\mathcal{S}_{E(\mathbf{y}|\mathbf{x})}$ *and* $\mathcal{S}_{E(\mathbf{y}|\mathbf{x}_{\mathcal{S}})}$ *exist and are uniquely defined, then testing (2.1) is equivalent to testing (1.3) for an arbitrary index set* $\mathcal{S}$ *as long as* $\mathcal{A} \subseteq \mathcal{S}$.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

We emphasize the importance of Theorem 1 because it guarantees that testing (2.1) is equivalent to testing (1.3) as long as $\mathcal{A} \subseteq \mathcal{S}$. This allows the two-stage procedure feasible for ultrahigh dimensional testing problems which will be discussed in the next subsection.

Next, we propose a new consistent lack-of-fit test for testing (2.1) in the population level based on the index set $\mathcal{S}$. In the sufficient dimension reduction context without any further regression model assumption, we define the error term $\boldsymbol{\varepsilon} \stackrel{\text{def}}{=} \mathbf{y} - E(\mathbf{y} \mid \boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}})$. The null hypothesis H_0 in (2.1) is equivalent to $E(\boldsymbol{\varepsilon} \mid \mathbf{x}_{\mathcal{S}}) = 0$. It is further equivalent to $\left\|E\{\boldsymbol{\varepsilon}\exp(i\mathbf{s}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}})\}\right\|^2 = 0$ for all $\mathbf{s} \in \mathbb{R}^{|\mathcal{S}|\times 1}$ using Fourier transformation, where i stands for an imaginary unit, i.e., $i^2 = -1$. We further note that

$$\left\|E\{\boldsymbol{\varepsilon}\exp(i\mathbf{s}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}})\}\right\|^2 = E\Big[\boldsymbol{\varepsilon}_1^{\mathrm{T}}\boldsymbol{\varepsilon}_2\exp\{i\mathbf{s}^{\mathrm{T}}(\mathbf{x}_{1,\mathcal{S}} - \mathbf{x}_{2,\mathcal{S}})\}\Big],$$

where $(\mathbf{x}_{1,\mathcal{S}}, \mathbf{y}_1)$ and $(\mathbf{x}_{2,\mathcal{S}}, \mathbf{y}_2)$ are two independent copies of $(\mathbf{x}_{\mathcal{S}}, \mathbf{y})$. Then, for an arbitrary weight function $\omega(\mathbf{s}) > 0$, testing H_0 in (2.1) equals to checking whether

$$E\left\{\int_{\mathbb{R}^{|\mathcal{S}|}} \boldsymbol{\varepsilon}_1^{\mathrm{T}}\boldsymbol{\varepsilon}_2\exp\{i\mathbf{s}^{\mathrm{T}}(\mathbf{x}_{1,\mathcal{S}} - \mathbf{x}_{2,\mathcal{S}})\}\omega(\mathbf{s})d\mathbf{s}\right\} = 0, \tag{2.2}$$

where the expectation E is taken with respect to $(\mathbf{x}_{1,\mathcal{S}}, \mathbf{y}_1)$ and $(\mathbf{x}_{2,\mathcal{S}}, \mathbf{y}_2)$. Then, the left-hand side of (2.2) can be considered as a test statistic. Borrowing the ideas of Székely, Rizzo and Bakirov (2007) and Shao and

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Zhang (2014), we specifically choose $\omega(\mathbf{s}) = \left(c_0\|\mathbf{s}\|^{1+|\mathcal{S}|}\right)^{-1}$ where $c_0 = \pi^{(1+|\mathcal{S}|)/2}/\Gamma\{(1+|\mathcal{S}|)/2\}$. Then, by $E(\boldsymbol{\varepsilon}_1) = E(\boldsymbol{\varepsilon}_2) = 0$ and Lemma 1 of Székely, Rizzo and Bakirov (2007), this test statistic has a closed form,

$$\begin{aligned} T &\overset{\text{def}}{=} E\left[\int_{\mathbb{R}^{|\mathcal{S}|}} \left(c_0\|\mathbf{s}\|^{1+|\mathcal{S}|}\right)^{-1} \boldsymbol{\varepsilon}_1^{\mathrm{T}}\boldsymbol{\varepsilon}_2 \exp\{i\mathbf{s}^{\mathrm{T}}(\mathbf{x}_{1,\mathcal{S}} - \mathbf{x}_{2,\mathcal{S}})\} d\mathbf{s}\right] \\ &= E\left\{\int_{\mathbb{R}^{|\mathcal{S}|}} \left(c_0\|\mathbf{s}\|^{1+|\mathcal{S}|}\right)^{-1} \boldsymbol{\varepsilon}_1^{\mathrm{T}}\boldsymbol{\varepsilon}_2 d\mathbf{s}\right\} \\ &- E\left[\int_{\mathbb{R}^{|\mathcal{S}|}} \left(c_0\|\mathbf{s}\|^{1+|\mathcal{S}|}\right)^{-1} \boldsymbol{\varepsilon}_1^{\mathrm{T}}\boldsymbol{\varepsilon}_2 \left[1 - \cos\{\mathbf{s}^{\mathrm{T}}(\mathbf{x}_{1,\mathcal{S}} - \mathbf{x}_{2,\mathcal{S}})\}\right] d\mathbf{s}\right] \\ &= -E\left(\boldsymbol{\varepsilon}_1^{\mathrm{T}}\boldsymbol{\varepsilon}_2 \|\mathbf{x}_{1,\mathcal{S}} - \mathbf{x}_{2,\mathcal{S}}\|\right). \end{aligned} \tag{2.3}$$

In general, $T \geq 0$. And $T = 0$ if and only if H_0 in (2.1) is true. This motivates us to utilize a consistent estimator of T as our test statistic for testing (2.1). The large values of T provide stronger evidence against the null hypothesis (2.1).

2.2 Two-Stage Lack-of-Fit Test with Screening

In order to estimate the test statistic T, we first study how to find an index set $\mathcal{S}$ to contain the true covariates set $\mathcal{A}$ and estimate the error term $\boldsymbol{\varepsilon} = \mathbf{y} - E(\mathbf{y} \mid \boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}})$. To this end, we propose a two-stage testing procedure based on data splitting strategy. We randomly partition the random sample $\mathcal{D} \overset{\text{def}}{=} \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, n\}$ into two halves. In the first stage, we screen out as many irrelevant covariates as possible based on the

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

first half $\mathcal{D}_1 \overset{\text{def}}{=} \{(\mathbf{x}_i, \mathbf{y}_i), i = 1, \ldots, n_1\}$ to obtain an index set $\mathcal{S}$ with a moderate size, where n_1 is the integer part of $n/2$. In the second stage, we develop a novel consistent lack-of-fit test for testing (1.3) using the second half of the data $\mathcal{D}_2 \overset{\text{def}}{=} \{(\mathbf{x}_i, \mathbf{y}_i), i = n_1 + 1, \ldots, n_1 + n_2\}$.

Stage 1: Feature Screening

Feature screening approaches are developed to screen out irrelevant covariates and retain the truly relevant ones in a moderate set for ultrahigh dimensional data. In the first stage, we apply the martingale difference correlation (MDC) based screening proposed by Shao and Zhang (2014) to the first half of the data and select a moderate set of covariates.

The martingale difference divergence (MDD) of $\mathbf{y}$ given each covariate X_j is defined by

$$\text{MDD}(\mathbf{y} \mid X_j)^2 = \frac{1}{c_q} \int_{\mathbb{R}^q} \frac{\|g_{\mathbf{y},X_j}(\mathbf{s}) - E(\mathbf{y}) g_{X_j}(\mathbf{s})\|^2}{\|\mathbf{s}\|^{1+q}} d\mathbf{s}, \tag{2.4}$$

where $g_{\mathbf{y},X_j}(\mathbf{s}) = E(\mathbf{y}e^{i\mathbf{s}^{\mathrm{T}}X_j})$, $g_{X_j}(\mathbf{s}) = E(e^{i\mathbf{s}^{\mathrm{T}}X_j})$, $c_q = \pi^{(1+q)/2}/\Gamma(1+q)/2$ and $\Gamma(\cdot)$ is the gamma function. The martingale difference correlation $\text{MDC}(\mathbf{y} \mid X_j)$ is the normalized version of $\text{MDD}(\mathbf{y} \mid X_j)$. $\text{MDC}(\mathbf{y} \mid X_j) = 0$ if and only if $E(\mathbf{y} \mid X_j) = E(\mathbf{y})$ almost surely when $E(\|\mathbf{y}\|^2 + X_j^2) < \infty$. That is, when $\text{MDC}(\mathbf{y} \mid X_j) = 0$, the conditional mean of $\mathbf{y}$ given X_j is independent of X_j. Shao and Zhang (2014) proposed to utilize the estimated MDC of the response given a covariate as the marginal utility to rank the

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

importance of all covariates and select a moderate set of covariates with top ranks. One may refer to Shao and Zhang (2014) for the calculation of the sample martingale difference correlation.

As mentioned by Cook and Li (2002), regression analysis is mostly concerned with inferring about the conditional mean of the response given the covariates. The MDC-based screening shares this spirit and inherits the model-free flavor of sufficient dimension reduction. We apply the MDC-based screening to $\mathcal{D}_1$, the first half of the data, and select the set of covariates defined by

$$\mathcal{S} = \{j : \widehat{\mathrm{MDC}}(\mathbf{y} \mid X_j) \text{ is among the top } s \text{ largest of all } p \text{ sample } \widehat{\mathrm{MDC}} \text{ values}\}.$$

By slight abuse of notation, we still use $\mathcal{S}$ to represent the selected set by screening. Under some regularity assumptions, the sure screening property holds for the MDC-based screening. That is, $P(\mathcal{A} \subseteq \mathcal{S}) \to 1$ as the sample size approaches the infinity. Then, Theorem 1 and the sure screening property together justifies that testing (2.1) is asymptotically equivalent to testing (1.3).

Stage 2: A Lack-of-Fit Test

Next, we elaborate how to estimate the test statistic T. First, we suggest to use the profile least squares approach to recover $\mathcal{S}_{E(\mathbf{y}|\mathbf{x}_\mathcal{S})}$. It amounts

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

to minimize the profile least squares and obtain the following estimator

$$\widehat{\boldsymbol{\beta}}_{\mathcal{S},-d_0} \stackrel{\text{def}}{=} \underset{\mathbf{b}\in\mathbb{R}^{(|\mathcal{S}|-d_0)\times d_0}}{\arg\min} \sum_{i=n_1+1}^{n} \|\mathbf{y}_i - \widehat{\mathbf{m}}(\mathbf{x}_{\mathcal{S},d_0,i} + \mathbf{b}^{\mathrm{T}}\mathbf{x}_{\mathcal{S},-d_0,i})\|^2,$$

where $\mathbf{x}_{\mathcal{S},d_0}$ is the vector of the first d_0 elements of $\mathbf{x}_{\mathcal{S}}$ and $\mathbf{x}_{\mathcal{S},-d_0}$ is the vector of the rest. Here we restrict the upper $d_0 \times d_0$ submatrix of $\boldsymbol{\beta}_{\mathcal{S}}$ to be an identity matrix to ensure that $\boldsymbol{\beta}_{\mathcal{S}}$ itself is identifiable (Ma and Zhu , 2013b) for a given d_0. $\widehat{\boldsymbol{\beta}}_{\mathcal{S},-d_0}$ is a $(|\mathcal{S}| - d_0) \times d_0$ matrix composed of the lower $(|\mathcal{S}| - d_0)$ rows of $\boldsymbol{\beta}_{\mathcal{S}}$. For an arbitrary $\mathbf{b} \in \mathbb{R}^{(|\mathcal{S}|-d_0)\times d_0}$, we estimate $\mathbf{m}(\mathbf{x}_{\mathcal{S},d_0,i} + \mathbf{b}^{\mathrm{T}}\mathbf{x}_{\mathcal{S},-d_0,i})$ with the leave-one-out kernel estimator, defined as

$$\widehat{\mathbf{m}}(\mathbf{x}_{\mathcal{S},d_0,i}+\mathbf{b}^{\mathrm{T}}\mathbf{x}_{\mathcal{S},-d_0,i}) \stackrel{\text{def}}{=} \sum_{j=n_1+1,j\neq i}^{n} \frac{K_h(\mathbf{x}_{\mathcal{S},d_0,j} + \mathbf{b}^{\mathrm{T}}\mathbf{x}_{\mathcal{S},-d_0,j} - \mathbf{x}_{\mathcal{S},d_0,i} - \mathbf{b}^{\mathrm{T}}\mathbf{x}_{\mathcal{S},-d_0,i})\mathbf{y}_j}{K_h(\mathbf{x}_{\mathcal{S},d_0,j} + \mathbf{b}^{\mathrm{T}}\mathbf{x}_{\mathcal{S},-d_0,j} - \mathbf{x}_{\mathcal{S},d_0,i} - \mathbf{b}^{\mathrm{T}}\mathbf{x}_{\mathcal{S},-d_0,i})},$$

where $K_h(\cdot) = K(\cdot/h)/h^{d_0}$, $K(\cdot)$ is a product of d_0 univariate kernel functions and h is the bandwidth. Then we estimate T by

$$T_{n_2} \stackrel{\text{def}}{=} \mathrm{tr}\bigg(-\frac{1}{n_2^2} \sum_{i=n_1+1}^{n_1+n_2} \sum_{j=n_1+1}^{n_1+n_2} \widehat{\boldsymbol{\varepsilon}}_i \widehat{\boldsymbol{\varepsilon}}_j^{\mathrm{T}} \|\mathbf{x}_{i,\mathcal{S}} - \mathbf{x}_{j,\mathcal{S}}\| \bigg), \tag{2.5}$$

where $\widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \widehat{\mathbf{m}}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}})$. In practice, larger values of T_{n_2} provide stronger evidence against H_0 in (2.1).

Since the null hypothesis (2.1) is concerned with studying whether a given small number of linear combinations of covariates is sufficient to characterize the conditional mean of $\mathbf{y}$, the test based on T_{n_2} is essentially a lack-of-fit test. Thus, we name this two-stage test procedure as a Lack-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Of-Fit Test with Screening (LOFTS) procedure and summarizes it in the following algorithm.

Algorithm 1 The LOFTS Procedure

Step 1. Randomly split the random sample into two even halves, $\mathcal{D}_1$ and $\mathcal{D}_2$.

Step 2. Apply the MDC-based screening to $\mathcal{D}_1$ and select the moderate set $\mathcal{S}$.

Step 3. Test (2.1) based on the test statistic T_{n_2} using $\mathcal{D}_2$. The associated p-value can be obtained using the bootstrap procedure (Algorithm 2 in Section 3).

Step 4. Reject (2.1) and (1.3) if p-value$< \alpha$, the significance level.

REMARK 1: It is worth noting that the data splitting technique is crucial in the proposed two-stage LOFTS procedure for ultrahigh dimensional data. If we do not split the data, a naive two-stage procedure is as follows. In the first stage, the MDC-screening is applied to the full sample data. In the second stage, the proposed test is conducted based on the selected covariates using the same data. In theory, this method works well and is even more efficient if $\mathcal{S}$ happens to be $\mathcal{A}$ exactly in the first stage. However, it is usually difficult to achieve in ultrahigh dimensional problems. Some inactive covariates which may contribute to the response in the finite sample level are often recruited in the first screening stage of the naive two-stage procedure, so the type-I error rates will be inflated for testing (2.1). One can refer to the simulations results in Section 4. This phenomenon owns

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

to spurious correlations that are inherent in ultrahigh dimension problems (Fan, Guo and Hao , 2012). The data splitting technique can eliminate spurious correlations and further avoid the size inflation. Because the two halves of the data set are independent, a covariate which has a large spurious sample correlation with the response over the first half has a small chance to be highly correlated with the response in the second half. Hence, its influence on the size of the test in the second stage is negligible.

REMARK 2: Feature screening can efficiently reduce the dimensionality of covariates in the first stage and retain the truly important covariates in the asymptotical sense. However, some important covariates may be missed in the finite sample level. In this sense, the choice of the reduced model size may be crucial for the screening procedure to work. Fan and Lv (2008) suggested a hard thresholding where the reduced model size is proportional to $[n/\log n]$. Wu, Boos and Stefanski (2007), Zhu, et. al (2011) and Li, Zhong, Li and Wu (2014) proposed a soft-thresholding rule by introducing auxiliary variables. To reduce this risk in practice, one may choose a relatively large set $\mathcal{S}$ if he believes that the size of important covariates is relatively large; or apply the iterative version of the MDC-based screening to avoid missing some important covariates which are marginally uncorrelated with the response. Another strategy to enhance the performance of

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

data splitting technique is multiple data splitting which will discussed in Section 5.

REMARK 3: The number of linear combinations of covariates, d_0, in (2.1) is prespecified before the lack-of-fit test procedure. The null hypothesis H_0 (2.1) holds trivially if we specify $d_0 = |\mathcal{S}|$ by letting $\boldsymbol{\beta}_{\mathcal{A}}$ is an $|\mathcal{S}|\times|\mathcal{S}|$ identity matrix. It is of great interest to determine the smallest number of linear combinations of covariates to sufficiently capture regression information of $E(\mathbf{y} \mid \mathbf{x}_{\mathcal{S}})$. For instance, the optimal value of d_0 is 1 for a general single-index model. For a given dimension d_0, if H_0 is rejected at some level of significance, then we can conclude that $\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}}$ is not sufficient to characterize the conditional mean $E(\mathbf{y} \mid \mathbf{x}_{\mathcal{S}})$ and some additional linear combinations of $\mathbf{x}_{\mathcal{S}}$ are further needed. In practice, we can sequentially perform the two-stage LOFTS procedure for d_0 starting from 1 to $|\mathcal{S}|$ until we fail to reject H_0 to determine the optimal value of d_0.

3. Theoretical Properties

In this section, we study theoretical properties of the proposed test, including the asymptotic distribution under the null hypothesis and the asymptotic distributions under both global and local alternative hypothesises. We also propose a bootstrap procedure to calculate the associated p-value. The

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

regularity conditions are provided in the Appendix.

Theorem 2 in the following states the asymptotic null distribution of the test statistic under the null hypothesis (2.1).

Theorem 2. *Assume Conditions (C1)-(C5) hold, under* H_0 *in (2.1),*

$$n_2 T_{n_2} \xrightarrow{D} \|\zeta(\mathbf{s})\|_\omega^2 \stackrel{def}{=} \int_{\mathbf{s}\in\mathbb{R}^{|\mathcal{S}|}} \|\zeta(\mathbf{s})\|^2 (c_0\|\mathbf{s}\|^{1+|\mathcal{S}|})^{-1} d\mathbf{s}, \quad \text{as } n_2 \to \infty,$$

where $\zeta(\mathbf{s})$ *denotes a complex-valued Gaussian random process with mean zero and covariance function* $cov\{\zeta(\mathbf{s}), \zeta^T(\mathbf{s}_0)\}$ *defined in (??) of the Supplement.*

However, the asymptotic distribution of T_{n_2} under H_0 is unfortunately not tractable because $\|\zeta(\mathbf{s})\|_\omega^2$ hinges upon the unknown joint distribution of $(\mathbf{x}_\mathcal{S}, \mathbf{y})$. In practice, we propose the bootstrap procedure in Algorithm 2 to calculate the associated p-value.

Theorem 3 states the consistency of the bootstrap procedure.

Theorem 3. *Assume Conditions (C1)-(C5) hold, we have that* $n_2\widetilde{T}_{n_2} \xrightarrow{D} \|\zeta(\mathbf{s})\|_\omega^2$, *as* $n_2 \to \infty$.

We remark that although it is not tractable in Theorem 2, the asymptotic distribution of T_{n_2} under H_0 is necessary to derive. Because Theorem 3 shows that the asymptotic null distribution of the bootstrapped test statistic is same as that of the original test statistic. It implies that the bootstrap

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Algorithm 2 The Bootstrap Procedure

Step 1. Obtain $\widehat{\boldsymbol{\beta}}_{\mathcal{S}}$ and $\widehat{\mathbf{m}}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}})$ using the second half $\mathcal{D}_2$, and calculate the residuals $\widehat{\boldsymbol{\varepsilon}}_i = \mathbf{y}_i - \widehat{\mathbf{m}}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S},i})$, for $i = n_1+1, n_1+2, \ldots, n$. Then, compute the test statistic T_{n_2} in (2.5).

Step 2. Draw the weights δ_i independently from $\{1, -1\}$ at random with probability 0.5. Let $\widetilde{\boldsymbol{\varepsilon}}_i = \widehat{\boldsymbol{\varepsilon}}_i\delta_i$ and generate $\widetilde{\mathbf{y}}_i = \widehat{\mathbf{m}}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S},i}) + \widetilde{\boldsymbol{\varepsilon}}_i$, for $i = n_1+1, n_1+2, \ldots, n$.

Step 3. Repeat Step 1 and calculate the test statistic $\widetilde{T}_{n_2}$ based on (2.5) using the new bootstrapped data set $(\mathbf{x}_i, \widetilde{\mathbf{y}}_i), i = n_1+1, n_1+2, \ldots, n$.

Step 4. Repeat Step 2 and 3 1,000 times to obtain $\widetilde{T}_{n_2}^{(1)}, \widetilde{T}_{n_2}^{(2)}, \ldots, \widetilde{T}_{n_2}^{(1,000)}$. The associated p-value is obtained by $1000^{-1}\sum_{b=1}^{1000} I(\widetilde{T}_{n_2}^{(b)} \geq T_{n_2})$, where $I(\cdot)$ is an indicator function. Reject H_0 if the p-value $< \alpha$, a given significance level.

procedure is able to provide asymptotically valid inference for the proposed lack-of-fit test.

Next, we consider two kinds of alternative hypothesises. The global alternative hypothesis can be specified in the following

$$H_{1g} : E(\mathbf{y} \mid \mathbf{x}) = E(\mathbf{y} \mid \mathbf{B}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}}), \text{ for some } |\mathcal{S}| \times d_1 \text{ matrix } \mathbf{B}_{\mathcal{S}}, d_0 < d_1 \leq |\mathcal{S}| \quad (3.1)$$

Under H_{1g}, d_0 linear combinations of covariates are not sufficient to recover the CMS $\mathcal{S}_{E(\mathbf{y}|\mathbf{x}_{\mathcal{S}})}$. We also consider a sequence of local alternatives:

$$H_{1l} : \mathbf{y} = \mathbf{m}(\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}}) + C_{n_2}\mathbf{g}(\mathbf{B}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}}) + \boldsymbol{\varepsilon}, \quad (3.2)$$

$$\text{for some } |\mathcal{S}| \times d_1 \text{ matrix } \mathbf{B}_{\mathcal{S}}, d_0 < d_1 \leq |\mathcal{S}|,$$

where $\boldsymbol{\beta}_{\mathcal{S}}$ is a subspace of $\mathbf{B}_{\mathcal{S}}$ and $C_{n_2} \to 0$ makes H_{1l} become local alterna-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

tives. Under H_{1l}, we have that $E(\boldsymbol{\varepsilon} \mid \mathbf{x}_{\mathcal{S}}) = \mathbf{0}$ and $\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}}$ is not sufficient to characterize the conditional mean function $E(\mathbf{y} \mid \mathbf{x}_{\mathcal{S}})$. However, as $n_2 \to \infty$, H_{1l} approaches H_0. Then the asymptotic distributions under both global and local alternative hypothesises are presented in Theorem 4.

Theorem 4. *Assume conditions (C1)-(C5) in the Appendix hold.*

(i) Under global alternative in (3.1), as $n_2 \to \infty$,

$$n_2{}^{1/2}(T_{n_2} - T) \xrightarrow{D} N(0, \sigma_0^2),$$

where the variance $\sigma_0^2 \overset{def}{=} 4var(Z_1 + Z_2 + Z_3)$, *and* Z_1, Z_2 *and* Z_3 *are defined in (??)-(??) of the Supplement, respectively.*

(ii) Under the local alternative in (3.2) with $C_{n_2} = n_2^{-1/2}$, *as* $n_2 \to \infty$,

$$n_2 T_{n_2} \xrightarrow{D} \|\zeta_0(\mathbf{s})\|_{\omega}^2 \overset{def}{=} \int_{\mathbf{s}\in\mathbb{R}^{|\mathcal{S}|}} \|\zeta_0(\mathbf{s})\|^2 (c_0\|\mathbf{s}\|^{1+|\mathcal{S}|})^{-1} d\mathbf{s},$$

where $\zeta_0(\mathbf{s})$ *denotes a complex-valued Gaussian random process with the mean function defined in (??) and the covariance function* $cov\{\zeta_0(\mathbf{s}), \zeta_0(\mathbf{s}_0)\}$ *defined in (??) of the Supplement.*

4. Numerical Studies

Example 1. We examine the finite-sample performance of the proposed two-stage test procedure by simulations. Consider the following two regres-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

sion models

$$\begin{aligned} \text{Model (I)}: \ Y &= (3+\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})^2 + c(\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x})^2 + \varepsilon, \\ \text{Model (II)}: \ Y &= \boldsymbol{\beta}_3^{\mathrm{T}}\mathbf{x} + (3+\boldsymbol{\beta}_4^{\mathrm{T}}\mathbf{x})^2 + c(\boldsymbol{\beta}_5^{\mathrm{T}}\mathbf{x})^2 + \varepsilon, \end{aligned}$$

where $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ is generated from a multivariate normal distribution with mean zero and covariance matrix $\Sigma = (\sigma_{kl})_{p\times p}$ with $\sigma_{kl} = 0.5^{|k-l|}$ for $k, l = 1, \ldots, p$, and $\varepsilon \sim N(0,1)$. Here, we set $\boldsymbol{\beta}_1 = (0.25, 0.25, 0.25, 0.25, 0, \ldots, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (0, 1, 0, 0, 0, \ldots, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_3 = (3, 0, 3, 0, 0, \ldots, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_4 = (0, 0.5, 0, 0.5, 0, \ldots, 0)^{\mathrm{T}}$ and $\boldsymbol{\beta}_5 = (0, 0, 2, 0, 0, \ldots, 0)^{\mathrm{T}}$. In both models, the value $c = 0$ corresponds to the null hypothesises while $c \neq 0$ represents the alternatives. Thus, the CMS $\mathcal{S}_{E(Y|\mathbf{x})}$ only depends on $\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}$ under H_0 but two linear combinations $(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}, \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x})$ under H_1 in Model (I). For Model (II), $\mathcal{S}_{E(Y|\mathbf{x})}$ is two-dimensional under the null but three-dimensional under the alternative.

We consider the sample size $n = 200$ and the covariate dimension $p = 2000$. Each sample is randomly divided into two equal halves. We perform the LOFTS procedure in Algorithm 1: Utilize MDC-based screening based on the first half $\mathcal{D}_1 = \{(\mathbf{x}_i, Y_i), i = 1, \ldots, 100\}$ to obtain a selected model $\mathcal{S}$ and test the hypothesis (2.1) based on the second half $\mathcal{D}_2 = \{(\mathbf{x}_{i,\mathcal{S}}, Y_i), i = 101, \ldots, 200\}$. To compare its empirical performance, we further consider the following two procedures: (1) A naive two-stage test procedure, denoted

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

by "NAIVE". We perform both the MDC-based screening and the lack-of-fit test on the same full sample. (2) An oracle test procedure, denoted by "Oracle". In the second stage, we directly conduct the test based on $\{(\mathbf{x}_{i,\mathcal{A}}, Y_i), i = 101, \ldots, 200\}$ as the true model $\mathcal{A}$ is known. We repeat simulations 1,000 times and summarize their finite-sample performances.

REMARK: For simplicity, we test the null hypothesis (2.1) with $d_0 = 1$ for Model (I) and $d_0 = 2$ for Model (II) to compare the performances of the test procedures in simulations. As a practical byproduct of the two-stage LOFTS procedure, one can sequentially perform the procedure for d_0 starting from 1 to $|\mathcal{S}|$ and the optimal value of d_0 is determined when the corresponding null hypothesis fails to be rejected at some significance level.

Screening Performances: In the two-stage LOFTS procedure, the first-stage screening performance is crucial for the follow-up test according to Theorem 1. The MDC-based screening method is effective to include almost all truly important covariates into the selected models in this example. Since it is not our main contribution in this paper, we only report the screening performance in the supplementary material. One may refer to Shao and Zhang (2014) for more numerical justifications of the MDC-based screening.

Size Performances: Next we evaluate the size performances of four test

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

procedures for Models (I) and (II) when $c = 0$, including our proposed LOFTS, the naive two-stage method, the oracle procedure and the DRMA procedure proposed by Guo, Wang and Zhu (2016). Since the DRMA procedure is proposed for parametric single index model, we only report their results for Model (I). The critical values of the lack-of-fit test procedure are decided through the proposed bootstrap procedure in Algorithm 2. We take four significance levels into consideration: $\alpha = 0.01, 0.02, 0.05$ and 0.10, and two different sizes of the selected models: $|\mathcal{S}| = 8$ and 16. The empirical Type-I error rates based on 1,000 repetitions for Models (I) and (II) are charted in Table 1. In addition, the QQ plots of the empirical p-values and the uniform distribution are charted in Panels (A) and (B) of Figure 1. It can be clearly seen that the empirical Type-I error rates of the LOFTS procedure, the DRMA procedure and the oracle method are pretty close to the user-specified significance levels. However, the empirical Type-I error rates of the naive two-stage method are obviously larger than the significance levels, especially when the selected model size becomes large. The inflation of Type-I errors in the naive two-stage procedure is due to spurious correlations between the response and some unimportant covariates in the ultrahigh dimensional data. The results further support the importance of the data splitting strategy which can efficiently eliminate

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

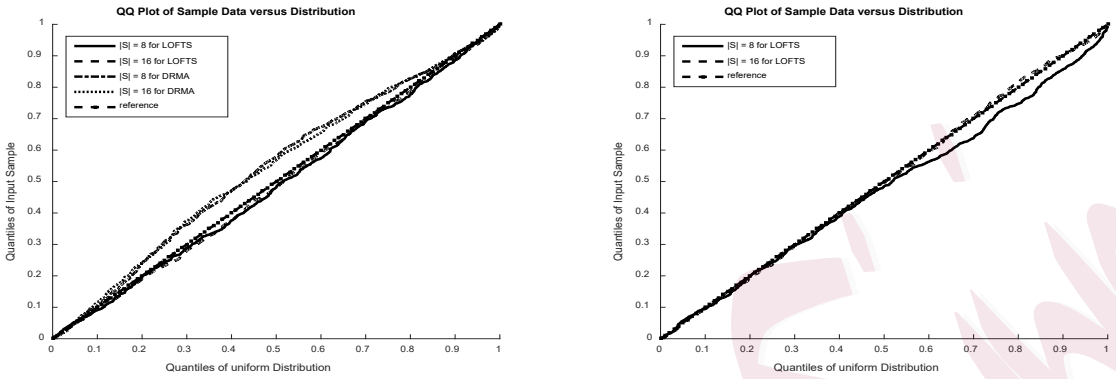
the effect of spurious correlations.

Table 1: The empirical Type-I error rates when $c = 0$.

Model	α	LOFTS		NAIVE		Oracle	DRMA	
		$\|\mathcal{S}\| = 8$	$\|\mathcal{S}\| = 16$	$\|\mathcal{S}\| = 8$	$\|\mathcal{S}\| = 16$		$\|\mathcal{S}\| = 8$	$\|\mathcal{S}\| = 16$
(I)	0.01	0.010	0.011	0.021	0.041	0.015	0.011	0.013
	0.02	0.017	0.020	0.044	0.064	0.020	0.020	0.025
	0.05	0.046	0.054	0.096	0.117	0.052	0.051	0.049
	0.10	0.114	0.105	0.152	0.209	0.095	0.098	0.088
(II)	0.01	0.013	0.017	0.021	0.026	0.009	-	-
	0.02	0.026	0.027	0.031	0.039	0.016	-	-
	0.05	0.045	0.049	0.076	0.087	0.054	-	-
	0.10	0.105	0.106	0.118	0.151	0.112	-	-

Power Performances: When $c \neq 0$, the previous null hypothesises are no longer true. For instance, the response depends on three different linear combinations in Model (II) when $c \neq 0$. We consider a sequence of values of $c = 0.2, 0.4, 0.6, 0.8$ and 1 to compare the empirical powers of the LOFTS and oracle procedures. Note that the "oracle" means the set of truly important covariates is assumed known in the second test stage. We choose the two significance levels: $\alpha = 0.05$ and 0.10, and two reduced model sizes, $|\mathcal{S}| = 8$ and 16. Table 2 summarizes the simulation results which show that the proposed two-stage LOFTS procedure is powerful to detect the significance of the tests. As the signal intensity parameter c increases, the empirical powers also increases. Once the true set $\mathcal{A}$ can be contained, the smaller the selected model size is, the larger the empirical

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)



(A) (B)

Figure 1: QQ plots of the empirical p-values and the uniform distribution for Model (I) in Panel (A) and Model (II) in Panel (B) in Example 1.

powers are. This phenomenon further confirms the importance of screening out irrelevant covariates in the ultrahigh dimensional test problems. Note that the outstanding size and power performances of the oracle procedure also demonstrate the advantage of the proposed lack-of-fit test.

Sequential Test Performances: By performing our proposed LOFTS sequentially, we can determine the smallest number of linear combinations of covariates to sufficiently capture regression information of $E(\mathbf{y} \mid \mathbf{x}_{\mathcal{S}})$. The procedure is conducted as follows. Start with $d_0 = 1$, test the null hypothesis (1.2) using the LOFTS. If the hypothesis is rejected, increase d_0 by one and perform the test again. Stop when the first null hypothesis is not rejected in the test series. The corresponding value of d_0, denoted

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 2: The empirical powers when $c \neq 0$ at $\alpha = 0.05$ or 0.10.

Model	c	LOFTS				Oracle		DRMA			
		$\|\mathcal{S}\| = 8$		$\|\mathcal{S}\| = 16$				$\|\mathcal{S}\| = 8$		$\|\mathcal{S}\| = 16$	
		0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10	0.05	0.10
(I)	0.2	0.243	0.369	0.162	0.247	0.572	0.692	0.099	0.160	0.068	0.127
	0.4	0.652	0.756	0.469	0.580	0.965	0.988	0.253	0.343	0.173	0.284
	0.6	0.853	0.908	0.702	0.779	0.995	0.997	0.496	0.620	0.382	0.495
	0.8	0.944	0.976	0.831	0.896	1.000	1.000	0.722	0.806	0.561	0.674
	1.0	0.963	0.980	0.893	0.937	1.000	1.000	0.836	0.906	0.702	0.796
(II)	0.2	0.965	0.977	0.646	0.768	1.000	1.000	-	-	-	-
	0.4	0.999	0.999	0.921	0.971	1.000	1.000	-	-	-	-
	0.6	0.998	0.999	0.963	0.982	1.000	1.000	-	-	-	-
	0.8	1.000	1.000	0.943	0.976	1.000	1.000	-	-	-	-
	1.0	0.997	0.997	0.923	0.965	1.000	1.000	-	-	-	-

by $\widehat{d}$, is the estimate of d^* that represents the smallest number of linear combinations of covariates to sufficiently recover the CMS. We report the empirical distributions of $\widehat{d}$ at the significance level $\alpha = 0.05$ based on 1000 simulations for Model (I) and (II) in Table 3. It can be seen that the LOFTS sequential tests are able to estimate the true structural dimension correctly with large probabilities, especially when $|\mathcal{S}| = 8$ and $c = 0$ or c is large.

We also compare our results with the method of iterative Hessian transformation (IHT) proposed by Cook and Li (2004) and the validated information criterion based method (VIC) by Ma and Zhang (2015), we only report the corresponding results when the reduced model size is 8, for simplicity. For the VIC method, since we are targeting on the structural dimen-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 3: The empirical distributions of $\widehat{d}$ at the significance level $\alpha = 0.05$.

Model	c	d^*	$\lvert\mathcal{S}\rvert = 8$				$\lvert\mathcal{S}\rvert = 16$			
			1	2	3	4	1	2	3	4
(I)	0	1	**0.954**	0.032	0.002	0.012	**0.946**	0.038	0.004	0.012
	0.2	2	**0.757**	0.218	0.009	0.016	**0.838**	0.128	0.015	0.019
	0.4	2	0.348	**0.607**	0.017	0.028	**0.531**	0.432	0.017	0.020
	0.6	2	0.147	**0.803**	0.022	0.028	0.298	**0.652**	0.018	0.032
	0.8	2	0.056	**0.897**	0.018	0.029	0.169	**0.786**	0.019	0.026
	1.0	2	0.037	**0.907**	0.023	0.033	0.107	**0.844**	0.010	0.039
(II)	0	2	0.000	**0.955**	0.019	0.026	0.003	**0.948**	0.030	0.019
	0.2	3	0.001	0.035	**0.923**	0.041	0.012	0.346	**0.607**	0.035
	0.4	3	0.011	0.001	**0.940**	0.048	0.078	0.073	**0.807**	0.042
	0.6	3	0.050	0.002	**0.907**	0.041	0.234	0.024	**0.705**	0.037
	0.8	3	0.000	0.000	**0.969**	0.031	0.006	0.052	**0.901**	0.041
	1.0	3	0.001	0.003	**0.965**	0.031	0.034	0.059	**0.870**	0.037

sion of the central mean space, we only examine the validated information criterion in semiparametric principal Hessian direction estimators. From Table 4, we can see that our LOFTS procedure outperforms both methods in our limited experiments. The IHT method often under estimates the structural dimension, this is mainly because the largest estimated eigenvalues often dominates the others. Although the estimated dimension using the validated information criterion converges to the true structural dimension in probability, there is no guarantee for the finite sample performance, especially when the sample size is small and the reduced model size is large. Our LOFTS procedure, however, avoids this problem through the proposed bootstrap procedure.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 4: The empirical distributions of $\widehat{d}$ of IHT and VIC when $|\mathcal{S}| = 8$.

Model	c	d^*	IHT				VIC			
			1	2	3	4	1	2	3	4
(I)	0	1	**0.960**	0.039	0.001	0.000	**0.836**	0.164	0.000	0.000
	0.2	2	**0.926**	0.073	0.001	0.000	**0.830**	0.170	0.000	0.000
	0.4	2	**0.821**	0.168	0.011	0.000	**0.787**	0.213	0.000	0.000
	0.6	2	**0.730**	0.258	0.012	0.000	**0.714**	0.284	0.002	0.000
	0.8	2	**0.583**	0.398	0.019	0.000	**0.601**	0.397	0.002	0.000
	1	2	0.466	**0.514**	0.020	0.000	0.437	**0.563**	0.000	0.000
(II)	0	2	0.395	**0.558**	0.047	0.000	0.105	**0.538**	0.357	0.000
	0.2	3	0.026	**0.941**	0.033	0.000	0.004	**0.858**	0.138	0.000
	0.4	3	0.002	**0.928**	0.070	0.000	0.000	**0.787**	0.213	0.000
	0.6	3	0.000	**0.897**	0.103	0.000	0.000	**0.701**	0.299	0.000
	0.8	3	0.000	**0.857**	0.142	0.001	0.000	**0.613**	0.381	0.006
	1	3	0.000	**0.808**	0.190	0.002	0.000	**0.554**	0.430	0.016

Example 2. We apply our proposed two-stage LOFTS procedure to a rat eye microarray expression data set which is available from Gene Expression Omnibus with accession number GSE5680. In this study, 120 twelve-week-old male rats were selected for tissue harvesting from the eyes and 31,042 different probe sets were measured for microarray analysis. In Scheetz, et. al (2006) and Huang, Ma and Zhang (2008), 18,976 probes that were considered adequately expressed and exhibited at least two-fold variation were retained in order to take deep insight into genetic variation involved in human's eye disease. The response variable TRIM32 at probe 1389163_at, one of the selected 18,976 probes, was recently found to cause

http://www.ncbi.nlm.nih.gov/geo

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Bardet-Biedl syndrome (Chiang, et. al , 2006). In our study, we aim to check if there exists a single linear combination of the gene expression levels that is sufficient to predict the expression level of the gene TRIM32.

We randomly partition this random sample into two halves, each with 60 observations, and marginally standardize all variables. We perform the MDC-based screening method to reduce the covariate dimension from 18,975 to 8 and 16, respectively. Denote $\mathbf{x}_{\mathcal{S}_1} = (X_1, \ldots, X_8)^{\mathrm{T}}$ and $\mathbf{x}_{\mathcal{S}_2} = (X_1, \ldots, X_{16})^{\mathrm{T}}$ the covariates retained in the screening stage. We apply the profile least squares approach to estimate $\boldsymbol{\beta}_{\mathcal{S}_1}$ based on $\{(\mathbf{x}_{j,\mathcal{S}_1}, Y_j), j = 61, \ldots, 120\}$ and $\boldsymbol{\beta}_{\mathcal{S}_2}$ based on $\{(\mathbf{x}_{j,\mathcal{S}_2}, Y_j), j = 61, \ldots, 120\}$. To ensure identifiability of $\boldsymbol{\beta}_{\mathcal{S}_1}$ and $\boldsymbol{\beta}_{\mathcal{S}_2}$, we fix the coefficient of X_1 to be 1. Table 5 exhibits the estimate coefficients (denoted by "coef"), along with their respective standard deviations (denoted by "std") and p-values. With two different model sizes, both estimates agree very well: X_4, X_6, X_7 and X_8, in addition to X_1, are important at the significance level $\alpha = 0.05$, X_3 and X_5 are important if only eight covariates are retained, X_9 becomes important if sixteen covariates are selected.

Next, we check whether a single linear combination of the retained covariates suffices to predict the expression level of TRIM32, based on $\{(\mathbf{x}_{j,\mathcal{S}_1}, Y_j), j = 61, \ldots, 120\}$ and $\{(\mathbf{x}_{j,\mathcal{S}_2}, Y_j), j = 61, \ldots, 120\}$, respectively.

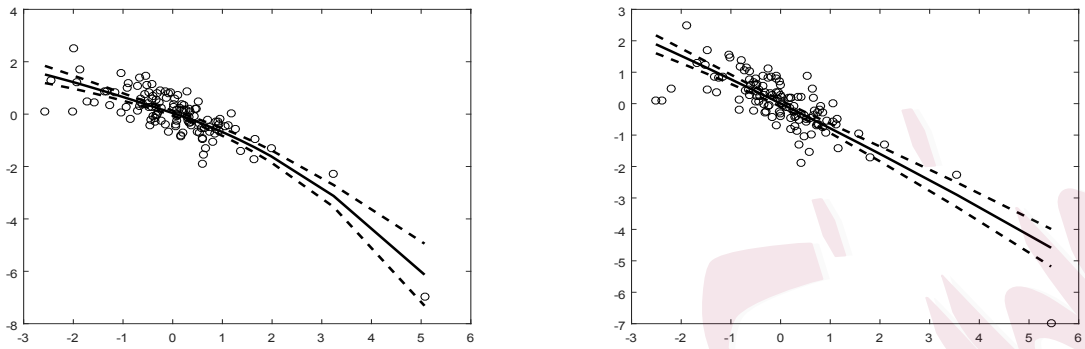
Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 5: The estimated coefficients, the standard errors and the p-values when $|\mathcal{S}| = 8$ and $|\mathcal{S}| = 16$, respectively, in Example 2.

$\lvert\mathcal{S}\rvert$		$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\widehat{\beta}_6$	$\widehat{\beta}_7$	$\widehat{\beta}_8$	
8	coef	0.043	1.772	-6.750	-5.762	-6.783	4.364	-3.251	
	std	0.824	0.828	2.794	2.596	2.643	1.848	1.346	
	p-value	0.959	0.037	0.019	0.030	0.013	0.022	0.019	
		$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{\beta}_4$	$\widehat{\beta}_5$	$\widehat{\beta}_6$	$\widehat{\beta}_7$	$\widehat{\beta}_8$	
16	coef	0.711	0.270	-2.087	-1.191	-1.778	1.990	-1.355	
	std	0.741	0.460	0.914	0.730	0.714	0.893	0.630	
	p-value	0.341	0.559	0.026	0.108	0.016	0.030	0.036	
		$\widehat{\beta}_9$	$\widehat{\beta}_{10}$	$\widehat{\beta}_{11}$	$\widehat{\beta}_{12}$	$\widehat{\beta}_{13}$	$\widehat{\beta}_{14}$	$\widehat{\beta}_{15}$	$\widehat{\beta}_{16}$
	coef	-1.766	-0.941	-0.679	0.964	-0.393	0.764	-0.607	0.115
	std	0.757	0.606	0.919	0.732	0.866	0.635	0.656	0.708
	p-value	0.023	0.126	0.463	0.193	0.652	0.234	0.358	0.872

The p-values obtained by our test procedures are 0.765 and 0.479 respectively, indicating that we have no evidence to reject the null hypothesis and a single linear combination indeed suffices to describe how the expression level of the gene TRIM32 varies with other genes. To further justify this test result, we chart the scatterplots of the response versus the standardized $(\mathbf{x}_{j,\mathcal{S}_1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{S}_1})$ and $(\mathbf{x}_{j,\mathcal{S}_1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{S}_2})$ in Panels (A) and (B) of Figure 2, respectively. The solid lines are fitted by local linear approximation where the bandwidths are decided through leave-one-out cross validation and the dashed lines are the 95% pointwise confidence intervals. It is clearly observed that the response can be described very well using only one single linear combination of the selected covariates.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)



(A) (B)

Figure 2: The scatterplots of the response versus standardized $(\mathbf{x}_{j,\mathcal{S}_1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{S}_1})$ in panel (A) and versus standardized $(\mathbf{x}_{j,\mathcal{S}_2}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{S}_2})$ in panel (B) in Example 2.

To further examine the prediction performance of single-index models based on the selected covariates, we calculate the mean squared prediction errors based on leave-one-out cross validation. The mean squared prediction error are 0.3801 based on $\{(\mathbf{x}_{j,\mathcal{S}_1}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{S}_1}, Y_j), j = 1, \ldots, 120\}$, 0.4297 based on $\{(\mathbf{x}_{j,\mathcal{S}_2}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\mathcal{S}_2}, Y_j), j = 1, \ldots, 120\}$. This indicates that the selected covariates are probably truly predictive for the expression level of the gene TRIM32 and a single linear combination of these covariates is probably sufficient to characterize the conditional mean of the response.

5. An Extension: Multiple Splitting

In the proposed two-stage testing procedure, the sure screening property that $\mathcal{A} \subseteq \mathcal{S}$ with probability tending to one is crucial to guarantee that

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

testing (2.1) is asymptotically equivalent to testing (1.3). However, in the sample level, some important variables may be unfortunately missed in the first screening stage due to limitation of the sample size, violation of some assumption or data randomness. In this case, the empirical Type-I error rates may be inflated. To deal with this issue, one may consider to utilize the iterated MDC-based screening to reduce the risk of missing important variables. Another efficient solution is the multi-splitting strategy in the light of Meinshausen, Meier and Bühlmann (2009). That is, one can divide the sample repeatedly (B times), and obtain one p-value from each sample splitting using the LOFTS procedure. For all p-values, denoted by $p_1, \ldots, p_B$, we define

$$Q(\gamma) = \min\left[1, q_\gamma\left(\{p_i/\gamma\}\right)\right],$$

for $\gamma \in (\gamma_{\min}, 1)$, where $q_\gamma\left(\{p_i/\gamma\}\right)$ is the γth quantile of $\{p_i/\gamma\}$ for $i = 1, \ldots, B$. Here, $\gamma_{\min} \in (0, 1)$ is a lower bound for γ, typically 0.05 or $1/B$ in practice. The adjusted p-value is then given by $Q(\gamma)$ for any fixed γ. However, a proper selection of γ may be difficult in practice. An adaptive version is defined as follows: Let $\gamma_{\min} \in (0, 1)$ be a lower bound for γ, typically 0.05, and

$$Q^*(\gamma) = \min\left\{1, (1 - \log\gamma_{\min}) \inf_{\gamma \in (\gamma_{\min}, 1)} Q(\gamma)\right\}.$$

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

With the adjusted p-value and the adaptive version of the p-value, the type-I error remains controlled at level α asymptotically. This result is presented in the following theorem.

Theorem 5. *Assume* $\lim_{n\to\infty} P(\mathcal{A} \subset \mathcal{S}_i) = 1$ *where* $\mathcal{S}_i$ *is the selected model in the screening stage based on the* i*th sample splitting, then*

$$\lim_{n\to\infty} \sup P\left\{Q(\gamma) \le \alpha\right\} \le \alpha, \qquad \lim_{n\to\infty} \sup P\left\{Q^*(\gamma) \le \alpha\right\} \le \alpha.$$

We also perform a toy example to illustrate how the the type-I error remains controlled at level α when some important covariates are missed with a nonignorable probability. We generate Y from the following regression model

$$Y = X_1 + X_2 + X_3 + 0.5X_4 + \varepsilon,$$

where $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ except X_4 are drawn from multivariate normal distribution with mean zero and covariance matrix $\mathbf{\Sigma} = (\rho_{kl})_{p\times p}$ with $\rho_{kl} = 0.5^{|k-l|}, k, l = 1, \ldots, p$, X_4 is generated from another regression model $X_4 = (0.5 - X_1 - X_3)^2 + \varepsilon_1$ and ε follows the standard normal distribution and is independent of $\mathbf{x}$, ε_1 is an independent copy of ε. The sample size is also set to be 200 while the dimensionality of the covariates is 1000, the reduced model size $|\mathcal{S}| = 5$ and the bootstrapped times in the LOFTS procedure is 300 for simplicity. In addition, the multi-splitting procedure is repeated 50

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Table 6: The empirical type-I errors for difference splitting techniques.

	single-splitting				multi-splitting			
nominal	0.01	0.02	0.05	0.10	0.01	0.02	0.05	0.10
empirical	0.054	0.077	0.135	0.220	0.001	0.002	0.037	0.095

times. In our simulations, X_4 are missed 204 times out of 1000 replicates, which makes the corresponding type-I error inflated. From Table 6, we can see that the multi-splitting strategy could improve the single-splitting technique and can well maintain the empirical type-I errors at the nominal levels $\alpha = 0.05$ and $\alpha = 0.10$.

Supplementary Materials

All technical proofs and the screening performance in the simulation are included in a separate online supplemental file.

Acknowledgements

The authors thank the editor, the associate editor, and the anonymous referees for their helpful comments that improved the article significantly. Zhang's work is supported by National Natural Science Foundation of P. R. China (11801349). Zhong's work is supported by National Natural Science Foundation of P. R. China (11671334 and 11301435), University Distinguished Young Researchers Program in Fujian Province and the Funda-

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

mental Research Funds for the Central Universities 20720181004. Zhu is the corresponding author and his work is supported by National Natural Science Foundation of P. R. China (11731011), Henry Fok Education Foundation Fund of Young College Teachers (141002), Project of Key Research Institute of Humanities and Social Sciences at Universities (16JJD910002) and National Youth Top-notch Talent Support Program, P. R. China. All authors equally contribute to this paper, and the authors are listed in the alphabetic order.

Appendix: Regularity Conditions

(C1) *(The Kernel Function)* The univariate kernel function $K(\cdot)$ is a density function with compact support. It is symmetric about zero and Lipschitz continuous. In addition, it satisfies

$$\int K(v)dv = 1, \quad \int v^i K(v)dv = 0, 1 \le i \le t-1, \quad 0 \ne \int v^t K(v)dv < \infty.$$

(C2) *(The Density)* The probability density function of $\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}}$, denoted by $f(\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}})$ is bounded away from 0 to infinity.

(C3) *(The Derivatives)* The $(t-1)$th derivatives of the mean function $\mathbf{m}(\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}})$, the density function $f(\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}})$ and $\mathbf{m}(\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}})f(\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}})$ are locally Lipschitz-continuous with respect to $\boldsymbol{\beta}_{\mathcal{S}}^{\mathrm{T}}\mathbf{x}_{\mathcal{S}}$.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

(C4) *(The Bandwidth)* The bandwidth h satisfies $h = O(n^{-\kappa})$, for some κ which satisfies $(2t)^{-1} < \kappa < (2d)^{-1}$.

(C5) *(The Moment)* The covariates used in the test stage statisfy that $E\left(\|\mathbf{x}_{\mathcal{S}}\|^2\right)/|\mathcal{S}| < \infty$.

References

Cook, D. (1998) *Regression Graphics: Ideas for Studying Regressions through Graphics.* Wiley, New York.

Cook, R. D. and Li, B. (2002) Dimension reduction for conditional mean in regression. *Annals of Statistics.* **30** 455–474.

Cook, R. D. and Li, B. (2004) Determining the dimension of iterative Hessian transformation. *Annals of Statistics.* **32** 2501–2531.

Chiang, A. P., Beck, J. S., Yen, H. J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Nishimura, D. Y., Braun, T. A., Kim, K.-Y. A., Huang, J., et al. (2006). Homozygosity mapping with SNP arrays identifies trim32, an e3 ubiquitin ligase, as a bardet–biedl syndrome gene (bbs11). *Proceedings of the National Academy of Sciences*, **103** 6287–6292.

Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society, Series B*, **74**, 37–65.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Fan, J. and Lv, J. (2008) Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B*. **70** 849–911.

González-Manteiga,W. and Crujeiras, R.M. (2013) An updated review of goodness-of-fit tests for regression models. *TEST*. **22** 361–411.

Guo, X., Wang, T. and Zhu, Lixing (2016) Model checking for parametric single-index models: a dimension-reduction model-adaptive approach. *Journal of the Royal Statistical Society, series B*. **78**: 1013–1035.

Huang, J., Ma, S., and Zhang, C. H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603–1618.

Li, L., (2007). Sparse sufficient dimension reduction. *Biometrika*. **94** 603–613.

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

Li, J., Zhong, W., Li, R. and Wu, R. (2014) A fast algorithm for detecting gene-gene interactions in genome-wide association studies, *The Annals of Applied Statistics*. **8** 2292C-2318.

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

Liu, J., Zhong, W. and Li, R. (2014), A selective overview of feature screening for ultrahigh-dimensional data, *Science China Mathematics* **58** 2033–2054.

Ma, Y. and Zhang, X. (2015) A validated information criterion to determine the structural

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

dimension in dimension reduction models, *Biomerika*, **102**, 409-420.

Ma, Y. and Zhu, L. P. (2012) A semiparametric approach to dimension reduction. *Journal of the American Statistical Association* **107** 168–179.

Ma, Y. and Zhu, L. P. (2013a) A review on dimension reduction. *International Statistics Review* **81** 134–150.

Ma, Y. and Zhu, L. P. (2013b) Efficiency loss and the linearity condition in dimension reduction. *Biometrika*. **100** 371–383.

Ma, Y. and Zhu, L. P. (2014) On estimation efficiency of the central mean subspace. *Journal of the Royal Statistical Society, Serie B*. **76**, 885–901.

Meinshausen, N., Meier L. and Bühlmann P. (2009) P-values for high-dimensional regression. *Journal of the American Statistical Association*. **104** 1671–1681.

Scheetz, T. E., Kim, K. Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., et al. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434.

Shah, R.D. and Buhlmann, P. (2018). Goodness of fit tests for high-dimensional linear models. *Journal of the Royal Statistical Society, Series B*, **80**: 113–135.

Shao, X. F. and Zhang, J. S. (2014) Martingale difference correlation and its use in high dimensional variable screening. *Journal of the American Statistical Association*. **109** 1302–1318.

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

Stute, W., Gonzáles-Manteiga, W. and Presedo-Quindimil, M. (1998) Bootstrap approximation in model checks for regression. *Journal of the American Statistical Association*. **93** 141–149.

Stute, W. and Zhu, L.X. (1998) Model checks for generalized linear models. *Scandinavian Journal of Statistics*. **29** 535–546.

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

Verzelen, N. and Villers, F. (2010) Goodness-of-fit tests for high-dimensional Gaussian linear models. *The Annals of Statistics*. **38** 704–752.

Wu, Y., Boos, D. D., and Stefanski L. A. (2007), Controlling variable selection by the addition of pseudo variables, *Journal of the American Statistical Association*. **102** 235–243.

Xia, Y., Li, W. K., Tong, H. and Zhang, D. (2004) A goodness-of-fit test for single-index models (with discussion). *Statistica Sinica*, **14** 1-39

Xia, Y., Tong, H., Li, W.K. and Zhu, L., (2002). An adaptive estimation of dimension reduction space (with discussions). *Journal of the Royal Statistical Society, series B*, **64** 363-410.

Zhang, X., Yao, S. and Shao, X., (2018). Conditional mean and quantile dependence testing in high dimension. *The Annals of Statistics*, **46**: 219–246

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

Zhu, L. and Ng, K. (2003). Checking the adequacy of a partial linear Model. *Statistica Sinica*,

Statistica Sinica: Newly accepted Paper
(accepted author-version subject to English editing)

113 763–781.

Zhu, L. P., Y, Z. and Zhu, L. X. (2003). A sparse eigen-decomposition estimation in semi-parametric regression. *Computational Statistics and Data Analysis*, **54** 976–986.

Zhu, L. P. and Zhong, W. (2015). Estimation and inference on central mean subspace for multivariate response data. *Computational Statistics and Data Analysis*, **92** 68–83.

Zhu, L. P., Zhu, L. X. and Feng, Z. (2010). Dimension reduction in regressions through cumulative slicing estimation *Journal of the American Statistical Association*, **105**, 1455–1466.

Yaowu Zhang, Research Institute for Interdisciplinary Science, School of Information Management and Engineering, Shanghai University of Finance and Economics, Shanghai 200433, China.

E-mail: zhang.yaowu@mail.shufe.edu.cn

Wei Zhong, Wang Yanan Institute for Studies in Economics, Department of Statistics, School of Economics, Fujian Key Laboratory of Statistical Science, Xiamen University, Xiamen 361005, China.

E-mail: wzhong@xmu.edu.cn

Liping Zhu, Research Center for Applied Statistical Science, Institute of Statistics and Big Data, Renmin University of China, Beijing 100872, China.

E-mail: zhu.liping@ruc.edu.cn