<table>
<thead>
<tr>
<th><strong>Title</strong></th>
<th>On Cumulative Slicing Estimation for High Dimensional Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manuscript ID</strong></td>
<td>SS-2018-0381</td>
</tr>
<tr>
<td><strong>URL</strong></td>
<td><a href="http://www.stat.sinica.edu.tw/statistica/">http://www.stat.sinica.edu.tw/statistica/</a></td>
</tr>
<tr>
<td><strong>DOI</strong></td>
<td>10.5705/ss.202018.0381</td>
</tr>
</tbody>
</table>
| **Complete List of Authors** | Cheng Wang  
Zhou Yu and  
Liping Zhu |
| **Corresponding Author** | Liping Zhu |
| **E-mail**           | zhulp1@hotmail.com                                       |

Notice: Accepted version subject to English editing.
ON CUMULATIVE SLICING ESTIMATION FOR
HIGH DIMENSIONAL DATA

Cheng Wang, Zhou Yu and Liping Zhu

Shanghai Jiao Tong University, East China Normal University,
Renmin University of China

Abstract: In the context of sufficient dimension reduction (SDR), sliced inverse regression (SIR) is the first and perhaps one of the most popular tools to reduce the covariate dimension for high-dimensional non-linear regressions. Despite the fact that the performance of SIR is very insensitive to the number of slices when the covariate is low or moderate dimensional, our empirical studies indicate that, the performance of SIR relies heavily upon the number of slices when the covariate is high- or ultrahigh-dimensional. How to select the optimal number of slices for SIR is still a longstanding problem in the SDR literature, which is a crucial issue for SIR to be effective in high- and ultrahigh-dimensional regressions. In this paper, we work with an improved version of SIR, the cumulative slicing estimation (CUME) method, which does not require selecting the optimal number of slices. We provide a general framework to analyze the phase transition phenomenon for the CUME method. We show that, without sparsity assumption, CUME is consistent if and only if \( p/n \to 0 \), where \( p \) stands for the covariate dimension and \( n \) stands for the sample size. If we make certain sparsity assump-
tions, then the thresholding estimate for the CUME method is consistent as long as \( \log(p)/n \to 0 \). We demonstrate the superior performance of our proposals through extensive numerical experiments.

**Key words and phrases:** Cumulative slicing estimation, sliced inverse regression, sparsity, sufficient dimension reduction.

1. Introduction

1.1. Background

Recent advances in information and technology allow us to collect big data in many scientific areas ranging from genome sequencing and biomedical imaging to social media analysis and high-frequency finance, etc. Big data are quite often characterized by high/ultrahigh-dimensionality (Fan et al., 2014). For instance, in health studies such as the Framingham heart study, many features related to certain heart disease and health status, such as genetic background, measurements from blood analysis, immune system status, nutrition, alcohol/tobacco/drug-consumption, operations, treatments, diagnosed diseases, are recorded. These features are high-dimensional and the high-dimensional pattern poses formidable and significant challenges for classic statistical methods.

Sufficient dimension reduction (Cook, 1998, SDR) is an effective paradigm that combines the idea of dimension reduction with the concept of suffi-
ciency for analyzing high-dimensional data. Suppose $Y \in \mathbb{R}$ is a response variable and $\mathbf{x} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$ is the associated covariate vector. Let $\perp$ stand for statistical independence. SDR seeks a $p \times d$ matrix $\mathbf{B} \in \mathbb{R}^{p \times d}$ such that

$$
Y \perp \mathbf{x} \mid (\mathbf{B}^T \mathbf{x}).
$$

Model (1.1) implies that, replacing the original high-dimensional $p$-vector $\mathbf{x}$ with $d$ linear combinations, denoted $(\mathbf{B}^T \mathbf{x})$, does not lose any regression information of $(Y \mid \mathbf{x})$. If $d = p$ and $\mathbf{B}$ is an arbitrary full-rank matrix, model (1.1) holds trivially. By the very purpose of dimension reduction, $d$ is often a small number. In real-world applications, quite often $d = 1$, 2 or at most 3. It is noticeable that $\mathbf{B}$ is not unique. If $\mathbf{B}$ satisfies model (1.1), for any nonsingular $\mathbf{C}$, $\mathbf{BC}$ will satisfy model (1.1) as well. Therefore, the parameter of interest is the column space of $\mathbf{B}$, denoted by $\text{span}(\mathbf{B})$. We refer to $\text{span}(\mathbf{B})$ with minimum column dimension as the central subspace, if it is uniquely defined. We denote the central subspace by $\mathcal{S}_{Y \mid \mathbf{x}}$. With slight abuse of notation, we still use $\mathbf{B}$ as a basis matrix of $\mathcal{S}_{Y \mid \mathbf{x}}$. We refer to the column dimension of $\mathbf{B}$ as the structural dimension of $\mathcal{S}_{Y \mid \mathbf{x}}$.

The very first tool for recovering $\mathcal{S}_{Y \mid \mathbf{x}}$ is sliced inverse regression (Li, 1991, SIR for short), which has gained wide recognition during the past two decades. Let $\boldsymbol{\Sigma} \overset{\text{def}}{=} \text{var}(\mathbf{x})$ and $\boldsymbol{\Lambda}_{\text{SIR}} \overset{\text{def}}{=} \text{var}\{E(\mathbf{x} \mid Y)\}$. SIR identi-
fies $S_{Y|x}$ through span$(\Sigma^{-1}A_{\text{sir}}\Sigma^{-1})$, the column space of $\Sigma^{-1}A_{\text{sir}}\Sigma^{-1}$. Li (1991) suggested a slicing procedure to estimate $A_{\text{sir}}$. The slicing estimation divides the range of the observed response values into $H$ slices, then calculates the sample average of the concomitant covariates within each slice. Due to its computational efficiency, simplicity and generality, the slicing estimation was later applied to many other SDR methods (Cook and Weisberg, 1991; Li and Wang, 2007). The slicing estimation is consistent for SIR when $p$ is fixed (Li, 1991; Hsing and Carroll, 1992; Zhu and Ng, 1995) and $H$ ranges from 2 to $n/2$. Zhu et al. (2006), Zhong et al. (2012) and Jiang and Liu (2014) proved the consistency of slicing estimation when $p = o(n^{1/2})$ and $H$ is a fixed number. Recently, Lin et al. (2018) showed that the slicing estimate of SIR is consistent when $p = o(n)$. However, the convergence rate derived in Lin et al. (2018) varies with the number of slices, which is often undesirable since selecting the optimal number of slices is still a longstanding problem in the SDR literature.

1.2. Issues with SIR

When the covariate is high/ultrahigh-dimensional, SIR faces several significant challenges, from both the theoretical and the practical perspectives. To the best of our knowledge, the consistency of slicing estimation for relatively large $p$ (say, log($p$) = $o(n)$) remains unknown in the SDR
ON CUMULATIVE SLICING ESTIMATION

literature. From the theoretical perspective, asymptotic study of the consistency of SIR with large $p$ is not straightforward, as the convergence rate also depends upon $H$ and a data-driven selection scheme for determining $H$ is yet unavailable. From the practical perspective, our empirical studies indicate that the performance of SIR may depend upon $H$ when $p$ is relatively large, which is not the case when $p$ is small. An anonymous referee pointed out that, if $H$ is too small, the pattern between $Y$ and $x$ may be averaged out within each slice. By contrast, if $H$ is too large, SIR may suffer from the large inner-slice variation. Therefore, how to choose an optimal $H$ delicately is an important issue when $p$ is relatively large.

We demonstrate through some simulated examples that SIR may be sensitive to the selection of $H$ when $p$ is relatively large. We adapt the following models which were commonly used in the SDR literature. In particular, models (1.2) and (1.5) were used by Li (1991), model (1.3) was used by Zhu et al. (2010), models (1.4) and (1.7) were used in Lin et al. (2018).

The covariates $x = (X_1, \ldots, X_p)^T$ are drawn independently from standard normal distribution, and $\varepsilon$ also follows a standard normal distribution. The
ON CUMULATIVE SLICING ESTIMATION

response variable $Y$ is generated from the following models:

\begin{align*}
Y &= X_1 + X_2 + X_3 + X_4 + 0.5\varepsilon; \\
Y &= \sin (X_1 + 0.5\varepsilon); \\
Y &= (X_1 + X_2 + X_3)^3/2 + 0.5\varepsilon; \\
Y &= (1 + X_1)/\{0.5 + (1.5 + X_2)^2\} + 0.2\varepsilon; \\
Y &= 4\sin (X_1 + X_2) + \exp (X_3 + X_4) + 0.2\varepsilon; \\
Y &= (X_1 + \cdots + X_7)\exp (X_8 + X_9) + 0.2\varepsilon. \tag{1.7}
\end{align*}

In general, estimating $S_{Y|x}$ for one-dimensional models is easier than that for multiple-dimensional ones. We measure the accuracy of the SIR estimate through the trace correlation, which is proposed by Ferré (1998). Let $\mathbf{B}$ be an underlying true basis of $S_{Y|x}$ and $\widehat{\mathbf{B}}$ be an estimated basis obtained by the SIR method. Define $\mathbf{P} \overset{\text{def}}{=} \mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top$ and $\widehat{\mathbf{P}} \overset{\text{def}}{=} \widehat{\mathbf{B}}(\widehat{\mathbf{B}}^\top\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^\top$. We denote $\text{trace}(\mathbf{A})$ as the trace of a matrix $\mathbf{A}$ and $d$ the structural dimension of $S_{Y|x}$. The trace correlation is defined as $r^2(d) \overset{\text{def}}{=} \text{trace}(\widehat{\mathbf{P}}\mathbf{P})/d$, which ranges from 0 to 1. Larger $r^2(d)$ values indicate more accurate estimation. In particular, $r^2(d) = 1$ if the estimated $S_{Y|x}$ and the true $S_{Y|x}$ are identical and $r^2(d) = 0$ if these two spaces are orthogonal to each other. We report how the mean values of $r^2(d)$ vary with the slice number $H$ for different dimension $p$. The simulation results based on 1000 repetitions are summarized in Figure 1.
(A)-(F) where the sample size is 200.

It is not surprising to see that when \( p \) is small, say \( p = 10 \), SIR has a very stable performance, as the \( r^2(d) \) values appear constant for a very wide range of slice numbers. However, when \( p \) is relatively large, the resulting pattern about \( H \) in SIR is quite different. The performance of SIR deteriorates quickly when \( p \) is large and \( H \) is too large or too small. For instance, in model (1.3) with \( p = 100 \), \( r^2(d) = 47.3\% \) when \( H = 2 \) and \( r^2(d) = 33.7\% \) when \( H = 40 \) while the peak occurs at \( H = 6 \) with \( r^2(d) = 59.21\% \). This indicates that our experiences for the small \( p \) case may not carry over to the large \( p \) case. When \( p \) is large, an SDR method which does not require selecting an appropriate slicing number is highly warranted.

We also observe that SIR deteriorates sharply as \( p \) increases in all models, which motivates us to suspect that the SIR estimate may not maintain consistency if \( p \) is relatively large. A consistent estimate is thus highly desired for a particular SDR method in ultrahigh-dimensional setting.

1.3. Our Contributions

In this paper, we work with an improved version of SIR, the so-called cumulative slicing estimation (Zhu et al., 2010, CUME for short). Unlike SIR, the CUME method is completely free of the slice number \( H \). It recovers \( S_{Y|x} \) through \( \text{span}(\Sigma^{-1} \Lambda \Sigma^{-1}) \), where \( \Lambda \overset{\text{def}}{=} E\{m(Y) m^T(Y)\} \) and \( m(y) \overset{\text{def}}{=} \ldots \)
cov\{x, I(Y \leq y)\}. Note that \(m(y) = \text{cov}\{E(x \mid Y), I(Y \leq y)\}\), indicating that both CUME and SIR use the inverse regression \(E(x \mid Y)\) to identify \(S_{Y\mid x}\). The difference is that, the slice number \(H\) is totally uninvolved in the CUME method. In the present paper, we provide a general framework to analyze the phase transition phenomenon for the CUME method. We show that, without any sparsity assumption, the CUME method is consistent if \(p = o(n)\). If both \(\Sigma^{-1}\) and \(\Sigma^{-1}\Lambda\Sigma^{-1}\) are sparse matrices, we suggest a thresholding estimate for \(\Sigma^{-1}\Lambda\Sigma^{-1}\). We will show that, the sparse estimate of \(\Sigma^{-1}\Lambda\Sigma^{-1}\) is consistent as long as \(\log(p) = o(n)\).

Recently, Lin et al. (2018) introduced Lasso-SIR algorithm to obtain a sparse estimate of the central subspace. The resulting estimate achieves the optimal convergence rate under certain sparsity conditions when \(p = o(n^2\lambda^2)\), where \(\lambda\) is the generalized signal-to-noise ratio. By contrast, our proposed thresholding estimate is consistent and achieves the same convergence rate when \(\log(p) = o(n)\). In addition, implementing the Lasso-SIR requires to specify the number of slices, whereas our method is completely free of slice number. In this sense, our results improve theirs significantly.

Next we demonstrate through some simulated examples that, for CUME to be consistent, \(p = o(n)\) is the largest divergence rate. We still adopt models (1.2)-(1.7) and generate the covariates and error terms in the same
ON CUMULATIVE SLICING ESTIMATION

manner. We consider different sample size $n$ and different covariate dimension $p$ such that their ratio $(p/n)$ equals 0.2, 0.1, 0.05, $(\log(n/5))^{-2}$ and $(\log(n))^{-2}$. The sample size $n$ ranges from 100 to 5000. We repeat our experiments 1000 times. Again we use the mean values of the trace correlation $r^2(d)$ to illustrate the performance of CUME. The simulation results are summarized in Figure 2 (A)-(F).

It can be clearly seen from Figure 2 that, if $p$ is proportional to $n$, say, $(p/n) = 0.05, 0.1$ and 0.2, the mean values of $r^2(d)$ appear very flat as $n$ increases. In this case, there is no hope for CUME to be consistent as $n$ diverges because otherwise the $r^2(d)$ values should get closer to one as $n$ increases. In contrast, if $(p/n) \to 0$, for example, if $(p/n) = (\log(n/5))^{-2}$ or $(\log(n))^{-2}$, the $r^2(d)$ values approach one gradually as $n$ increases. For example, in model (1.5) with $(p/n) = (\log(n/5))^{-2}$, the mean value of $r^2(d)$ is 74.88% when $n = 100$ and 94.17% when $n$ increases to 5000. This exhibits a clear pattern that, as long as $(p/n) \to 0$ when $n \to \infty$, the CUME method converges. Whereas $(p/n)$ is a constant as $n \to \infty$, there would be no hope to see such convergence behavior of the CUME method if we made no sparsity assumption. These simulation results demonstrate that CUME is consistent if and only if $p = o(n)$. For high-dimensional data where $p = O(n)$ or even $\log(p) = o(n)$, we need to regularize the CUME matrix...
to accommodate high-dimensionality under some sparsity assumptions.

The rest of this paper is organized as follows. In Section 2 we study
the consistency of the estimated CUME matrix. We show that the clas-
sical moment estimate of the CUME matrix is consistent for $p = o(n)$
without sparsity assumption, and the regularized estimate is consistent for
$log(p) = o(n)$ with sparsity assumption. In Section 3 we investigate the
finite-sample performance of our proposals through comprehensive simula-
tions and an application to a real-world problem. In Section 4 we give some
brief comments on our proposals and outline some future problems. All
technical details are given in the on-line Supplementary Material.

2. Main Results

2.1. Definitions and Notations

Suppose $\{(x_i, Y_i), i = 1, \ldots, n\}$ is a random sample of $(x, Y)$. For a
$p \times q$ matrix $A_{p \times q}$, let $\text{span}(A)$ be the space spanned by the columns of
$A$, $\text{trace}(A)$ stand for the trace of $A$, $\text{rank}(A)$ be the rank of $A$, and
$\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ stand for the maximum and the minimum eigenvalues
of $A$, respectively. Let $\lambda_{\max}(A) = \lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_q(A) =
\lambda_{\min}(A)$, where $\lambda_k(A)$ stands for the $k$-th largest principal eigenvalue of $A$.
Sometimes we may simply use $\lambda_k$ in place of $\lambda_k(A)$ when it is sufficiently
clear from the context. Let $\|A\|_F \overset{\text{def}}{=} \text{trace}(A^T A)^{1/2}$ be the Frobenius
norm and \(\|A\|\) be the spectral norm of \(A\). To be precise,

\[
\|A\| \overset{\text{def}}{=} \sup_{a^\top a = 1} (a^\top A^\top A a)^{1/2} = \lambda_{\max}^{1/2}(A^\top A).
\]

Let \(A_{k,l}\) be the \((k, l)\)-th entry of \(A\), i.e., \(A = (A_{k,l})_{p \times q}\). Define

\[
\|A\|_\infty \overset{\text{def}}{=} \max_{1 \leq k \leq p, 1 \leq l \leq q} |A_{k,l}|, \quad \text{and} \quad \|A\|_1 \overset{\text{def}}{=} \max_{1 \leq k \leq p} \sum_{l=1}^{q} |A_{k,l}|.
\]

Denote \(I_{p \times p}\) as the \(p \times p\) identity matrix. Let \(I(A)\) be an indicator function which equals one if the event \(A\) is true and zero otherwise, and \(\text{pr}(A) = E\{I(A)\}\) represents the probability that \(A\) is true. We denote \(c_0, C_0, c_1, C_1, \ldots\), as a sequence of generic constants which may take different values at various places.

The ultimate goal of SDR is to identify and recover \(S_{Y|X}\). Let \(B_{p \times d} \in \mathbb{R}^{p \times d}\) be a basis of \(S_{Y|X}\) and \(\hat{B}_{p \times d} \in \mathbb{R}^{p \times d}\) be an estimated basis, \(P \overset{\text{def}}{=} B (B^\top B)^{-1} B^\top\) and \(\hat{P} \overset{\text{def}}{=} \hat{B} (\hat{B}^\top \hat{B})^{-1} \hat{B}^\top\). The projection matrix \(P\), rather than its basis \(B\), is unique and identifical. Therefore, to quantify how accurate \(\hat{B}\) is in estimating \(S_{Y|X}\), it is reasonable to use the following three criteria:

1. The spectral norm \(\|\hat{P} - P\|\);

2. The Frobenius norm \(\|\hat{P} - P\|_F\);

3. The trace correlation \(r^2(d) \overset{\text{def}}{=} \text{trace}(\hat{P} P)/d\).
It is noticeable here that the Frobenius norm is equivalent to the trace correlation \( \| \hat{P} - P \|_F^2 = 2d\{1 - r^2(d)\} \). Both the spectral norm \( \| \hat{P} - P \| \) and the Frobenius norm \( \| \hat{P} - P \|_F \) are nonnegative and have upper bounds, with a smaller value indicating a more accurate estimate. The trace correlation \( r^2(d) \) ranges from zero to one, with a larger value indicating a better estimate. In the following section, we will study convergence rate of \( \hat{P} \) under the above three norms when \( p = o(n) \).

2.2. Usual Moment Estimate for the CUME Method when \( p = o(n) \)

In this section, we advocate using the CUME method to obtain \( \hat{B} \), an estimated basis of \( S_{Y|x} \), in that its estimation is completely free of tuning parameters. Recall that \( \Sigma = \text{var}(x) \), \( \Lambda = E\{m(Y)m^T(Y)\} \), where \( m(y) = \text{cov}\{x, I(Y \leq y)\} \). We estimate \( m(y) \) with

\[
\hat{m}(y) \overset{\text{def}}{=} n^{-1} \sum_{i=1}^{n} (x_i - \bar{x}) I(Y_i \leq y), \quad \bar{x} \overset{\text{def}}{=} n^{-1} \sum_{i=1}^{n} x_i,
\]

and estimate \( \Lambda \) and \( \Sigma \) with

\[
\hat{\Lambda} \overset{\text{def}}{=} n^{-1} \sum_{i=1}^{n} \hat{m}(Y_i) \hat{m}^T(Y_i), \quad \text{and} \quad \hat{\Sigma} \overset{\text{def}}{=} n^{-1} \sum_{i=1}^{n} (x_i - \bar{x}) (x_i - \bar{x})^T. \quad (1.8)
\]

The estimated basis \( \hat{B} \) is composed of the first \( d \) principal eigenvectors of \( \hat{\Sigma}^{-1} \hat{\Lambda} \hat{\Sigma}^{-1} \).

To state the consistency of \( \hat{B} \), we make the following assumptions.
Define \( d \overset{\text{def}}{=} \text{rank}(\Lambda) \) and \( \lambda_d(\Lambda) \) as the \( d \)-th principal eigenvalue of \( \Lambda \) (which is also the smallest nonzero principal eigenvalue of \( \Lambda \)). Assume \( \lambda_d(\Lambda) \geq c_0^{-1} \).

(2): Assume \( c_0^{-1} \leq \lambda_{\text{min}}(\Sigma) \leq \lambda_{\text{max}}(\Sigma) \leq c_0 \), where \( \lambda_{\text{min}}(\Sigma) \) and \( \lambda_{\text{max}}(\Sigma) \) are the respective smallest and largest eigenvalues of \( \Sigma \).

(3): Assume the covariate vector \( \mathbf{x} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p \) is sub-Gaussian.

That is, for any unit-length vector \( \mathbf{e} \), \( \text{pr}( \mid \mathbf{e}^T \mathbf{x} \mid \geq t) \leq \exp(1 - c_0 t^2) \), for all \( t \geq 0 \).

Assumption (A1) requires that the nonzero eigenvalues of \( \Lambda \) be bounded away from below. It ensures that the magnitude of signals, represented by nonzero eigenvalues of \( \Lambda \), are detectable. Assumption (A2) is widely assumed in the literature of high-dimensional covariance matrix estimation. See, for example, Bickel and Levina (2008) and Cai et al. (2011). It allows that the covariates are correlated, as long as their covariance matrix is non-singular. Assumption (A3) requires that the covariates are sub-Gaussian, which is weaker than the normality assumption. We require this technical condition to yield exponential inequalities.

**Theorem 1.** Assume conditions (A1)−(A3). If \( p = o(n) \), then

1. \( \| \hat{\Sigma} - \Sigma \|^2 = O_p(p/n) \) and \( \| \hat{\Lambda} - \Lambda \|^2 = O_p \{ \max(p, \log n)/n \} \);
ON CUMULATIVE SLICING ESTIMATION

2. $\|\hat{P} - P\|^2_F = (2d) \{1 - r^2(d)\} = O_p \{ \max(p, \log n)/n \}$;

3. $\|\hat{P} - P\|^2 = O_p \{ \max(p, \log n)/n \}$.

We first give some brief comments on Theorem 1. The sub-Gaussian assumption is widely used to study the consistency of the sample covariance matrix $\hat{\Sigma}$. See, for example, Vershynin (2012), Bunea and Xiao (2015) and Koltchinskii and Lounici (2017). In the present context, an important contribution we make is that we derive the convergence rate of both $\hat{\Lambda}$ and $\hat{P}$ under the spectral norm. Moreover, one may wonder why $\hat{\Sigma}$ and $\hat{\Lambda}$ have different convergence rates. The rate $(\log n/n)$ in $\|\hat{\Lambda} - \Lambda\|^2$ comes up when we derive the uniform convergence rate of $\hat{m}(y)$. We believe that the rate $O_p \{ \max(p, \log n)/n \}$ may be refined to $O_p(p/n)$ at the expenses of some delicate and yet much more tedious analysis. In high-dimensional data analysis, it is reasonable to expect that $p$ is greater than $\log n$. Accordingly, $\|\hat{\Lambda} - \Lambda\|^2 = O_p(p/n)$. In other words, the presence of $\log n/n$ does not have an essential impact on the convergence rate of $\|\hat{\Lambda} - \Lambda\|^2$ when the covariate dimension is high. The second statement connects the Frobenius norm with the trace correlation. In particular, $\|\hat{P} - P\|_F = (2d)^{1/2} \{1 - r^2(d)\}^{1/2}$. The last two statements also indicate that the Frobenius norm and the spectral norm of $\hat{P} - P$ have identical convergence rates.

Zhu et al. (2010) and Jiang and Liu (2014) derived the convergence
rate of \( \hat{\mathbf{B}} \) when \( p = o(n^{1/2}) \). We improve their results substantially in the present paper in that \( p = o(n) \), which is the largest \( p \) that one can handle without any sparsity condition. The key to improve the convergence rate of \( \| \hat{\mathbf{P}} - \mathbf{P} \|_F \) from \( O_p(p/n^{1/2}) \) (Zhu et al., 2006, 2010; Jiang and Liu, 2014) to \( O_p(p^{1/2}/n^{1/2}) \) is that we used an improved Davis-Kahan \( \sin \theta \) theorem (Yu et al., 2015). In particular,

\[
\| \hat{\mathbf{P}} - \mathbf{P} \|_F \leq 4 \min\{d^{1/2}\| \hat{\Sigma}^{-1} \hat{\Lambda} \hat{\Sigma}^{-1} - \Sigma^{-1} \Lambda \Sigma^{-1} \|_F, \| \hat{\Sigma}^{-1} \hat{\Lambda} \hat{\Sigma}^{-1} - \Sigma^{-1} \Lambda \Sigma^{-1} \|_F \}. 
\]

In general, \( \| \hat{\Sigma}^{-1} \hat{\Lambda} \hat{\Sigma}^{-1} - \Sigma^{-1} \Lambda \Sigma^{-1} \| \leq \| \hat{\Sigma}^{-1} \hat{\Lambda} \hat{\Sigma}^{-1} - \Sigma^{-1} \Lambda \Sigma^{-1} \|_F \), where \( d \) is a small number by the very purpose of dimension reduction. The improved Davis-Kahan \( \sin \theta \) theorem accounts for the significantly improved convergence rate.

Recently, Lin et al. (2018, Theorem 1) studied the consistency of SIR when \( p = o(n) \) and they showed that

\[
\| \hat{\mathbf{A}}_{\text{SIR}} - \mathbf{A}_{\text{SIR}} \| = O_p \left( \frac{1}{H^\vartheta} + \frac{H^2 p}{n} + \sqrt{\frac{H^2 p}{n}} \right),
\]

where \( \vartheta \) is a nonnegative constant. The above convergence rate indicates that the optimal number of slices is

\[
H = O \left\{ \left( \frac{p}{n} \right)^{-\frac{1}{2(\vartheta+1)}} \right\}. \quad (1.9)
\]

Accordingly, the resulting optimal convergence rate is

\[
\| \hat{\mathbf{A}}_{\text{SIR}} - \mathbf{A}_{\text{SIR}} \| = O_p \left\{ \left( \frac{p}{n} \right)^{-\frac{\vartheta}{2(\vartheta+1)}} \right\}. \quad (1.10)
\]
The above convergence rate is slower than that for the CUME method we derived in Theorem 1. Moreover, how to choose an optimal $H$ satisfying (1.9) is also an issue in the literature, partly because $\vartheta$ is unknown. It is thus encouraging to conclude that avoiding slicing estimation does not only overcome the longstanding computational issue involved in SIR, but also help CUME possess a better convergence rate than SIR.

In Theorem 1 we show that the usual moment estimate of the CUME matrix is consistent when $p = o(n)$. Next we demonstrate that such an estimate is inconsistent when $p/n \rightarrow \gamma$ for some $\gamma \in (0, 1)$. This indicates that $p = o(n)$ is a necessary and sufficient condition for the usual moment estimate of the CUME matrix to be consistent. In Section 1, we illustrate the inconsistency issue through some simulated examples when $p/n \rightarrow \gamma$ for some $\gamma \in (0, 1)$. Next we further demonstrate this inconsistency issue through the following analytical example.

**Example 1.** We assume $Y = X_1 + \sigma \varepsilon$, where $x = (X_1, \ldots, X_p)^\top$ follows a multivariate standard normal distribution and $\varepsilon$ is standard normal. In this example, $\Sigma \eqdef \text{var}(x) = I_{p \times p}$, where $I_{p \times p}$ stands for the $p \times p$ identity matrix. Because $x$ is standardized, we can simply estimate the basis matrix $B$ with the first $d$ principal eigenvectors of $\hat{\Lambda}$. With slight abuse of notation, we denote the estimated basis with $\hat{B}$. Define $P = B(B^\top B)^{-1}B^\top$ and
ON CUMULATIVE SLICING ESTIMATION

\( \hat{\mathbf{P}} \hat{\mathbf{B}} (\hat{\mathbf{B}}^T \hat{\mathbf{B}})^{-1} \hat{\mathbf{B}}^T \). In the Supplement Material, we show that

\[
\mathrm{pr}\left\{ \| \hat{\mathbf{P}} - \mathbf{P} \|^2_F \geq \frac{\gamma}{6\pi^2(1 + \sigma^2)(1 + \gamma)^2} \right\} \longrightarrow 1.
\]

This indicates that the usual moment estimate of the CUME matrix is no longer consistent when \( p/n \to \gamma \) for some \( \gamma \in (0, 1) \).

2.3. Regularized Estimate for the CUME Method when \( \log p = o(n) \)

In this section we derive the convergence rate for the CUME method when \( \log p = o(n) \). When \( p \) is greater than \( n \), \( \hat{\Sigma} \) is no longer invertible even when \( \Sigma \) is nonsingular. To address this issue, we turn to sparsity assumptions and propose a sparse solution, denoted as \( \hat{\Omega}_s \) and \( \hat{\Theta}_s \), to estimate \( \Omega \defeq \Sigma^{-1} \) and \( \Theta \defeq \Omega \Lambda \Omega \), respectively. Let \( \hat{\mathbf{B}}_s \) be composed of the first \( d \) principal eigenvectors of \( \hat{\Theta}_s \). In this section we derive the consistency of \( \hat{\mathbf{B}}_s \) under certain sparsity assumptions. Define \( \hat{\mathbf{P}}_s = \hat{\mathbf{B}}_s (\hat{\mathbf{B}}_s^T \hat{\mathbf{B}}_s)^{-1} \hat{\mathbf{B}}_s^T \).

We study the consistency of \( \hat{\mathbf{B}}_s \) under the Frobenius norm \( \| \hat{\mathbf{P}}_s - \mathbf{P} \|_F \), the spectral norm \( \| \hat{\mathbf{P}}_s - \mathbf{P} \| \) and the trace correlation \( r^2(d) = \text{trace}(\hat{\mathbf{P}}_s \mathbf{P})/d \), respectively.

We suggest an estimation of the precision matrix \( \Omega \) first and then propose a sparse estimation for \( \Theta \) based on the sparse solution \( \hat{\Omega}_s \).
**Estimation of \( \Omega \):** Estimation of the precision matrix has been extensively studied in the literature. See, for example, Meinshausen and Bühlmann (2006), Cai et al. (2011), Zhang and Zou (2014) and a recent review article by Fan et al. (2016). In this work, we will simply adapt the constrained \( \ell_1 \) minimization for inverse covariance matrix estimation (CLIME) proposed by Cai et al. (2011). The CLIME method is implemented as follows. For a given tuning parameter \( \lambda_{1n} \), let \( \hat{\Omega} \) be the solution set of the following optimization problem:

\[
\hat{\Omega} \in \arg \min_{\Omega} \| \Omega \|_1, \quad \text{subject to} \quad \| \hat{\Sigma} \Omega - I \|_\infty \leq \lambda_{1n},
\]

where \( \hat{\Sigma} \) is defined in (1.8). The above solution \( \hat{\Omega} \) is not symmetric in general. To obtain a symmetric estimate, the CLIME estimator \( \hat{\Omega}_s \) is defined as \( \hat{\Omega}_s \overset{\text{def}}{=} (\hat{\Omega}_{s,k,l}) \) where

\[
\hat{\Omega}_{s,k,l} = \hat{\Omega}_{s,l,k} = \hat{\Omega}_{k,l}I(\| \hat{\Omega}_{k,l} \| \leq \| \hat{\Omega}_{l,k} \|) + \hat{\Omega}_{l,k}I(\| \hat{\Omega}_{k,l} \| > \| \hat{\Omega}_{l,k} \|).
\]

In other words, we take the one with smaller magnitude between \( \hat{\Omega}_{k,l} \) and \( \hat{\Omega}_{l,k} \). The resultant estimate \( \hat{\Omega}_s \) is symmetric and more importantly, positive definite with high probability. By assuming that the covariates have exponential type tails and \( \lambda_{1n} = C_1(\log p/n)^{1/2} \) for some generic constant \( C_1 \), Cai et al. (2011) show that

\[
\| \hat{\Omega}_s - \Omega \| = O_p \left\{ M^{2-2q} s_1(p)(\log p/n)^{(1-q)/2} \right\},
\]
holds uniformly for

$$\Omega \in \mathcal{U}_1\{q, s_1(p)\}$$

\[
\begin{align*}
\text{def} = \left\{ \Omega : \Omega > 0, \|\Omega\|_1 \leq M \text{ and } \max_{1 \leq k \leq p} \sum_{l=1}^{p} |\Omega_{k,l}|^q \leq s_1(p) \right\}, \tag{1.11}
\end{align*}
\]

for some $0 \leq q < 1$. For brevity, we assume $\|\Omega\|_1 \leq c_0$ and then

$$\|\hat{\Omega}_s - \Omega\| = O_p \left\{ s_1(p)(\log p/n)^{(1-q)/2} \right\}. \tag{1.12}$$

Next we suggest a thresholding estimate for $\Theta$.

**Thresholding Estimation of $\Theta$:** For a given tuning parameter $\lambda_{2n}$, we propose the following sparse estimation

$$\hat{\Theta}_s \text{def} = (\hat{\Theta}_{s,k,l})_{p \times p} = \left\{ \hat{\Theta}_{k,l} I(|\hat{\Theta}_{k,l}| \geq \lambda_{2n}) \right\}_{p \times p},$$

where $\hat{\Theta}_{k,l}$ is the $(k,l)$-th element of $\hat{\Theta} \text{def} = \hat{\Theta}_s \hat{\Omega}_s$ and $\hat{\Omega}_s$ is the CLIME estimation. Assume

$$\Theta \in \mathcal{U}_2\{q, s_2(p)\} \text{def} \left\{ \Theta : \max_{1 \leq k \leq p} \sum_{l=1}^{p} |\Theta_{k,l}|^q \leq s_2(p) \right\}, \text{ for some } 0 \leq q < 1.$$ 

We remark here that $s_1(p)$ and $s_2(p)$ are constants that may depend on $p$.

We control the sparsity levels with $s_1(p)$ and $s_2(p)$ in the respective classes $\mathcal{U}_1\{q, s_1(p)\}$ and $\mathcal{U}_2\{q, s_2(p)\}$. In particular, when $q = 0$ we require that the number of nonzero entries in each row be not greater than $s_1(p)$ or $s_2(p)$. 
ON CUMULATIVE SLICING ESTIMATION

The class \( \mathcal{U}_1\{q, s_1(p)\} \) was introduced by Bickel and Levina (2008) and Cai et al. (2011). It is straightforward to verify that the band covariance matrices and the covariance matrices with power decay correlations satisfy the sparsity condition in \( \mathcal{U}_1\{q, s_1(p)\} \). The class \( \mathcal{U}_2\{q, s_2(p)\} \) is defined in a similar manner to the class \( \mathcal{U}_1\{q, s_1(p)\} \). It can also be verified that the sparsity condition in \( \mathcal{U}_2\{q, s_2(p)\} \) is satisfied if the number of truly important covariates is small. In particular, the matrix \( \Theta \) related to models (1.2)-(1.7) is sufficiently sparse with its upper-left block sub-matrix being nonzero. In effect, the class \( \mathcal{U}_2\{q, s_2(p)\} \) covers many common dimension-reduction models (Zhu et al., 2010, Theorem 1).

**Theorem 2.** Assume conditions (A1)–(A3) and (1.11). Let \( \lambda_{1n} = C_1(\log p/n)^{1/2} \) and \( \lambda_{2n} = C_2(\log p/n)^{1/2} \) for some generic nonnegative constants \( C_1 \) and \( C_2 \). Then, as \( n \to \infty \),

\[
\|\hat{\Theta}_s - \Theta\| = O_p \left\{ s_1^{-q}(p)s_2(p)(\log p/n)^{(1-q)^2/2} \right\}.
\]

(1.13)

Theorem 3 states the consistency of \( \hat{\Theta}_s \).

**Theorem 3.** Under the conditions of Theorem 2,

1. \( \|\hat{P}_s - P\|_F = O_p \left\{ s_1^{-q}(p)s_2(p)((\log p/n)^{(1-q)^2/2} \right\}, \)

2. \( \|\hat{P}_s - P\| = O_p \left\{ s_1^{-q}(p)s_2(p)((\log p/n)^{(1-q)^2/2} \right\}. \)
ON CUMULATIVE SLICING ESTIMATION

Theorems 3 ensures that the estimated central space is consistent even when \((\log p/n)\) vanishes slowly as long as \(s_1(p)\) and \(s_2(p)\) are small numbers. This generalizes the applicability of the CUME method to ultrahigh-dimensional data.

**Tuning Parameter Selection:** It remains to choose the appropriate \(\lambda_{1n}\) and \(\lambda_{2n}\) values for the thresholding regularized CUME method. The selector for \(\lambda_{1n}\) has been discussed extensively by Cai et al. (2011). Simply, \(\lambda_{1n}\) is decided under a likelihood loss function coupled with five-fold cross-validation. We also suggest to choose \(\lambda_{2n}\) with five-fold cross-validation such that the distance correlation (Székely et al., 2007; Székely and Rizzo, 2009) between \((\hat{B}_s^T \mathbf{x})\) and \(Y\) is maximized. The distance correlation retains the model free flavor of SDR in that it has the ability to measure non-linear dependence between \(Y\) and \((\hat{B}_s^T \mathbf{x})\). In our proposed five-fold cross-validation procedure, we randomly partition the original sample into five equal sized subsamples. We retain a single subsample as the test set and the remaining four subsamples as the training set. For each \(\lambda_{2n}\), we can obtain an estimate \(\hat{\Theta}_s\) and accordingly, \(\hat{\mathbf{B}}_s\), using the training set. We calculate the distance correlation between \((\hat{B}_s^T \mathbf{x})\) and \(Y\) using the test set. The cross-validation procedure is then repeated five times, with each of the five subsamples used exactly once as the test set. The five distance corre-
lations are averaged to produce a single estimation. We choose $\lambda_{2n}$ which maximizes the average of five distance correlations. Our limited experience indicates that this procedure is very effective.

3. Numerical Studies

3.1. Simulations

We illustrate the finite-sample performance of our proposed sparse estimate $\hat{B}_s$ through simulations. We also compare our proposal with the classical CUME method, the SIR method with different number of slices. We use the trace correlation $r^2(d)$ to assess the the finite-sample performance of different proposals. We still adapt models (1.2) to (1.7) in our simulations. Throughout we draw $\mathbf{x} = (X_1, \ldots, X_p)^T$ from multivariate normal distribution with mean zero and covariance matrix $\Sigma$, and draw $\varepsilon$ independently from standard normal distribution. We consider three scenarios. In the first two scenarios we fix $n = 200$ and $p = 10, 50, 100, 200$ and 300. In the last scenario, we fix $n = 400$ and let $p = 1000$ and 5000. We set $\Sigma = I_{p \times p}$, $\Sigma = (0.2^{|k-l|})_{p \times p}$ and $\Sigma = (0.5^{|k-l|})_{p \times p}$ in the first, the second and the last scenarios, respectively. In the first two scenarios, we directly implement our proposed sparse estimate procedure, the CUME and the SIR method, to estimate $S_{Y|X}$. In the third scenario, we first implement the sure independent ranking and screening method (Zhu Statistica Sinica: Newly accepted Paper (accepted author-version subject to English editing))
et al., 2011) to reduce the covariate dimension from $p$ to $p_0$ using the first 200 observations. In other words, we retain $p_0$ covariates after screening. We choose $p_0 = [n/\log n], 2[n/\log n], \ldots, 5[n/\log n]$, which correspond to 38, 76, 114, 152 and 190, respectively. Next we implement our proposed sparse estimate, CUME and SIR, using the remaining 200 observations and the retained $p_0$ covariates. We repeat each scenario 1000 times, and report the mean and standard deviation of the $r^2(d)$ values. Table 1 presents the results for $\Sigma = (0.2^{k-l})_{p \times p}$. Additional simulations are relegated to the on-line Supplement Material.

The simulation results indicate that SIR is sensitive to the number of slices when $p$ is relatively large. In the linear model (1.2) in scenario 2 with $p = 100$ and $\Sigma = (0.2^{k-l})_{p \times p}$, the $r^2(d)$ value obtained by SIR$_{20}$ is 0.894, whereas that obtained by SIR$_2$ is 0.519. In model (1.4) in scenario 2, the $r^2(d)$ value obtained by SIR$_{20}$ is 0.882, whereas that obtained by SIR$_2$ is 0.439. In the third scenario, the effect of the slice number appears more substantial than in the first two scenarios. For example, in model (1.2) with $p = 1000$ and $p_0 = 114$, the $r^2(d)$ value obtained by SIR$_{20}$ is 0.827, whereas that obtained by SIR$_5$ is only 0.659. In model (1.2) with $p = 5000$ and $p_0 = 152$, the $r^2(d)$ value obtained by SIR$_{20}$ is 0.612, whereas that obtained by SIR$_5$ is as small as 0.454. These simulation results indicate
that the slice number in SIR has non-ignorable impact on its performance when the covariate dimension is relatively large. While how to choose an optimal slice number is still unknown in the literature, the CUME method is free of the number of slices. However, the performance of the classical CUME method also deteriorates quickly when the covariate dimension $p$ increases. For example, in model (1.2) in scenario 1 the $r^2(d)$ value obtained by CUME is 0.992 when $p = 10$ and is 0.551 when $p = 150$. By contrast, our proposed sparse estimate of the CUME matrix are very stable across all scenarios. The $r^2(d)$ values obtained by our proposal are all larger than 0.950 in one dimensional models and are all greater than 0.700 in all two-dimensional models, which is in line with our anticipation in that estimating two-dimensional $S_{Y|x}$ is usually more difficult than estimating one-dimensional $S_{Y|x}$.

3.2. Real-Data Analysis

We demonstrate our proposed sparse estimate for the CUME method using the breast cancer data collected by Van’t Veer et al. (2002). In this study, 24481 gene expression levels were collected from 97 lymph node-negative breast cancer patients. We remove observations which contain missing values, leaving 24188 gene expression levels. We aim to predict the tumor size with these gene expression levels. Because the covariates are
ON CUMULATIVE SLICING ESTIMATION

Table 1: The averages (standard deviations) of trace correlation\((\times 100)\) for Scenario 2 where \(\text{SIR}_k\) denotes SIR with \(k\) slices and \(\Sigma = (0.2^{k-l})_{p \times p}\).

<table>
<thead>
<tr>
<th>(p)</th>
<th>CUME</th>
<th>SIR2</th>
<th>SIR5</th>
<th>SIR10</th>
<th>SIR20</th>
<th>NEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>98.9 (0.6)</td>
<td>95.7 (2.0)</td>
<td>98.8 (0.6)</td>
<td>99.3 (0.4)</td>
<td>99.5 (0.3)</td>
<td>98.4 (1.6)</td>
</tr>
<tr>
<td>50</td>
<td>93.0 (2.0)</td>
<td>76.4 (4.8)</td>
<td>92.5 (2.0)</td>
<td>95.5 (1.2)</td>
<td>96.5 (0.9)</td>
<td>98.7 (1.7)</td>
</tr>
<tr>
<td>100</td>
<td>79.0 (5.0)</td>
<td>51.9 (6.3)</td>
<td>79.6 (4.2)</td>
<td>87.1 (2.9)</td>
<td>89.4 (2.4)</td>
<td>98.6 (1.7)</td>
</tr>
<tr>
<td>200</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>98.4 (2.0)</td>
</tr>
<tr>
<td>300</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>98.4 (2.0)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>96.8 (1.8)</td>
<td>94.2 (3.1)</td>
<td>96.5 (1.9)</td>
<td>96.6 (2.0)</td>
<td>96.4 (2.1)</td>
</tr>
<tr>
<td>50</td>
<td>81.8 (5.2)</td>
<td>71.4 (6.8)</td>
<td>80.4 (5.5)</td>
<td>80.5 (5.6)</td>
<td>78.6 (6.4)</td>
<td>98.2 (5.6)</td>
</tr>
<tr>
<td>100</td>
<td>59.2 (8.6)</td>
<td>45.0 (8.3)</td>
<td>56.7 (9.0)</td>
<td>55.6 (9.8)</td>
<td>48.8 (11.9)</td>
<td>98.4 (6.1)</td>
</tr>
<tr>
<td>200</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>99.0 (5.1)</td>
</tr>
<tr>
<td>300</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>99.0 (4.9)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>98.6 (0.8)</td>
<td>94.2 (2.7)</td>
<td>98.7 (0.7)</td>
<td>99.2 (0.4)</td>
<td>99.4 (0.3)</td>
</tr>
<tr>
<td>50</td>
<td>90.3 (2.6)</td>
<td>69.8 (5.5)</td>
<td>91.5 (2.0)</td>
<td>95.0 (1.3)</td>
<td>96.0 (1.1)</td>
<td>99.0 (1.6)</td>
</tr>
<tr>
<td>100</td>
<td>72.6 (6.0)</td>
<td>43.9 (6.6)</td>
<td>77.8 (4.4)</td>
<td>85.8 (3.1)</td>
<td>88.2 (2.7)</td>
<td>98.9 (1.9)</td>
</tr>
<tr>
<td>200</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>98.9 (1.8)</td>
</tr>
<tr>
<td>300</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>98.9 (1.7)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>88.7 (5.4)</td>
<td>94.4 (3.9)</td>
<td>85.1 (7.3)</td>
<td>88.3 (6.2)</td>
<td>88.0 (6.3)</td>
</tr>
<tr>
<td>50</td>
<td>58.6 (5.5)</td>
<td>74.7 (6.2)</td>
<td>54.1 (6.4)</td>
<td>56.6 (7.2)</td>
<td>54.1 (7.4)</td>
<td>93.4 (8.9)</td>
</tr>
<tr>
<td>100</td>
<td>39.4 (4.3)</td>
<td>58.5 (9.0)</td>
<td>36.1 (4.4)</td>
<td>36.4 (4.8)</td>
<td>32.8 (4.9)</td>
<td>92.6 (9.0)</td>
</tr>
<tr>
<td>200</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>91.0 (10.2)</td>
</tr>
<tr>
<td>300</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>90.7 (11.4)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>85.5 (7.2)</td>
<td>58.8 (4.1)</td>
<td>84.2 (7.7)</td>
<td>85.1 (8.0)</td>
<td>82.8 (9.7)</td>
</tr>
<tr>
<td>50</td>
<td>48.7 (7.2)</td>
<td>42.3 (5.8)</td>
<td>49.9 (7.3)</td>
<td>49.7 (8.0)</td>
<td>45.4 (7.3)</td>
<td>85.3 (15.8)</td>
</tr>
<tr>
<td>100</td>
<td>27.9 (5.4)</td>
<td>27.7 (8.7)</td>
<td>28.7 (5.3)</td>
<td>28.8 (5.5)</td>
<td>25.9 (5.3)</td>
<td>82.4 (16.3)</td>
</tr>
<tr>
<td>200</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>81.7 (16.9)</td>
</tr>
<tr>
<td>300</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>81.2 (16.3)</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>96.8 (1.2)</td>
<td>48.1 (1.1)</td>
<td>95.7 (1.7)</td>
<td>96.5 (1.3)</td>
<td>96.6 (1.3)</td>
</tr>
<tr>
<td>50</td>
<td>80.4 (3.3)</td>
<td>38.0 (2.5)</td>
<td>73.9 (3.7)</td>
<td>79.1 (3.6)</td>
<td>78.7 (3.8)</td>
<td>95.0 (5.1)</td>
</tr>
<tr>
<td>100</td>
<td>55.9 (5.2)</td>
<td>26.1 (3.3)</td>
<td>50.8 (4.8)</td>
<td>54.3 (5.1)</td>
<td>50.9 (6.0)</td>
<td>93.6 (7.5)</td>
</tr>
<tr>
<td>200</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>87.0 (16.5)</td>
</tr>
<tr>
<td>300</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>80.0 (21.4)</td>
</tr>
</tbody>
</table>
ultrahigh-dimensional, we first apply Zhu et al. (2011)'s sure independent ranking and screening procedure to select the top 50 gene expression levels which are possibly the most predictive for the tumor size. We split the dataset into two sets: the training set contain 65 observations and the test set contains the remaining 32 observations. We estimate $\Theta$ using the training set. Figure 3 displays the scree plot of the eigenvalues of $\hat{\Theta}_x$. It can be clearly seen that there is an obvious nonzero eigenvalue. Figure 3 also presents the scatter plot between the response variable and the first linear combination on the test data. It seems that the first linear combination exhibits a clear monotone trend. We further conduct a distance correlation t-test (Székely et al., 2007) between each of the first two linear combinations and the response variable. The p-values are 0.005 and 0.275. All these suggest that the first linear combination may be sufficient to predict the tumor size. Therefore, it is reasonable to infer that the central subspace $S_{Y|x}$ is possibly one-dimensional.

Next we examine the performance of the first linear combination to predict the tumor size. We randomly partition the whole data set into a training and a test data set. We repeat this partition procedure 1000 times. We estimate $S_{Y|x}$ using different proposals based on the training set, and calculate the distance correlation (Székely et al., 2007) between
ON CUMULATIVE SLICING ESTIMATION

Table 2: Simulation results for Breast Cancer Data: The averages (standard deviations) of distance correlations and mean square errors based on test data.

<table>
<thead>
<tr>
<th></th>
<th>CUME</th>
<th>SIR₂</th>
<th>SIR₅</th>
<th>SIR₁₀</th>
<th>SIR₂₀</th>
<th>NEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>1.68(0.55)</td>
<td>1.31(0.34)</td>
<td>1.69(0.60)</td>
<td>1.86(0.70)</td>
<td>2.61(1.31)</td>
<td>0.78(0.22)</td>
</tr>
<tr>
<td>DC</td>
<td>0.37(0.10)</td>
<td>0.39(0.10)</td>
<td>0.34(0.09)</td>
<td>0.33(0.08)</td>
<td>0.33(0.08)</td>
<td>0.57(0.10)</td>
</tr>
</tbody>
</table>

the first linear combination and the response based on the test data. We also predict the tumor size based on the test set with nonparametric kernel regression. We evaluate the prediction performance using mean squared errors. The averages (the standard deviations) of the distance correlation and the mean squared errors are reported in Table 2 based on 1000 random partitions. The prediction performance of sliced inverse regression varies with the number of slices. In terms of both criteria, it can be seen from Table 2 that, our sparse estimate for the CUME method is again superior to both SIR and the classical CUME method.

4. Concluding Remarks

In the present article we first show that the classical CUME method is
ON CUMULATIVE SLICING ESTIMATION

consistent if and only if $p = o(n)$. This is the largest possible $p$ we can handle without sparsity assumption. When $p$ is greater than $n$, we introduce a sparse estimate for the CUME matrix, and show that the sparse estimate is consistent as long as $\log(p) = o(n)$. The sparse estimates involve two tuning parameters, $\lambda_{1n}$ and $\lambda_{2n}$. In this article, we suggest to select the optimal $\lambda_{1n}$ first, and then given the optimal $\lambda_{1n}$, select the optimal $\lambda_{2n}$. An alternative option is that we may choose them simultaneously when the computational complexity is not an issue. There are several other issues that deserve our further investigation. For example, for the CUME method to be consistent, the linearity condition is implicitly assumed. This assumption is violated if some covariates are categorical or discrete. How to relax the linearity assumption when $\mathbf{x}$ is ultrahigh-dimensional is not straightforward. In addition, how to decide the dimension of $S_{Y|x}$ for ultrahigh-dimensional semiparametric regressions is another important question and warrants for additional thorough investigation. Based on the asymptotic theory of CUME established in the present work, one may follow Luo and Li (2016) to combine eigenvalues and variation of eigenvectors for order determination in the high or ultrahigh-dimensional setting. Another interesting extension is to apply the thresholding idea to the functional data case.

Supplementary Material
ON CUMULATIVE SLICING ESTIMATION

The proofs of Example 1 and Theorems 1-2, together with additional simulations, are relegated to the on-line Supplementary Material.

Acknowledgements

We thank the Editor, an Associate Editor, and two anonymous reviewers for their insightful comments. Wang’s research is supported by Shanghai Sailing Program 16YF1405700 and National Natural Science Foundation of China 11701367. Yu’s research is supported by the National Natural Science Foundation of China 11571111, the 111 project B14019, the Program of Shanghai Subject Chief Scientist 14XD1401600, and the Shanghai Rising Star Program 16QA1401700. Zhu is the corresponding author and his research is supported by National Natural Science Foundation of China (11731011), Chinese Ministry of Education Project of Key Research Institute of Humanities and Social Sciences at Universities (16JJD910002) and National Youth Top-notch Talent Support Program of China.

References


REFERENCES

of reduced effective rank population matrices, with applications to fPCA. 

Bernoulli 21(2), 1200–1230.


REFERENCES


REFERENCES


School of Mathematical Sciences, MOE-LSC, Shanghai Jiao Tong University, Shanghai 200240, China.

E-mail: chengwang@sjtu.edu.cn

Key Laboratory of Advanced Theory and Application in Statistics and Data
REFERENCES

Science, Ministry of Education and School of Statistics, East China Normal University, Shanghai 200241, China.

E-mail: zyu@stat.ecnu.edu.cn

Institute of Statistics and Big Data and Center for Applied Statistics, Renmin University of China, Beijing 100872, China.

E-mail: zhu.liping@ruc.edu.cn
Figure 1: The mean values of the $r^2(d)$ values over 1000 repetitions for models (1.2)-(1.7). The squares stand for $p = 10$, the circles stand for $p = 50$ and the triangles stand for $p = 100$. The horizontal axis stands for the slice number $H$, which ranges from 2 to 40 in models (1.2)-(1.4) and from 3 to 40 in models (1.5)-(1.7), and the vertical axis stands for the mean values of the trace correlation $r^2(d)$. 

(A): model (1.2) 
(B): model (1.3) 
(C): model (1.4) 
(D): model (1.5) 
(E): model (1.6) 
(F): model (1.7)
Figure 2: The mean values of the $r^2(d)$ values over 1000 repetitions for models (1.2)-(1.7). The line marked with squares stands for $(p/n) = 0.2$; the line marked with hollow points stands for $(p/n) = 0.1$; the line marked with triangles stands for $(p/n) = 0.05$; the line marked with stars stands for $(p/n) = \{\log(n/5)\}^{-2}$ and the line marked with solid points stands for $(p/n) = \{\log(n)\}^{-2}$. The horizontal axis stands for the sample size $n$, which ranges from 100 to 5000, and the vertical axis stands for the mean values of the trace correlation $r^2(d)$. 
(A): Eigenvalues  (B): The 1st linear combination  (C): Residual plot

Figure 3: (A): The scree plot of the principal eigenvalues; (B): The scatter plots between the response in the vertical axis and the first linear combination \((x^T \beta_1)\) in the horizontal axis on the test data set; (C) The scatter plots between the residual of the nonparametric kernel regression using the first linear combination.