Statistica Sinica **31** (2021), 223-242
doi:https://doi.org/10.5705/ss.202018.0381

ON CUMULATIVE SLICING ESTIMATION FOR HIGH DIMENSIONAL DATA

Cheng Wang, Zhou Yu and Liping Zhu

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

Abstract: In the context of sufficient dimension reduction (SDR), the sliced inverse regression (SIR) successfully reduces the covariate dimension of a high-dimensional nonlinear regression. When the covariate is low or moderate dimensional, the performance of the SIR is insensitive to the number of slices. However, our empirical studies indicate that the performance of the SIR relies heavily on the number of slices when the covariate is high or ultrahigh dimensional. Determining the optimal number of slices remains an open problem in the SDR literature, despite its importance to the effectiveness of SIR in high- and ultrahigh-dimensional regressions. Thus, we propose an improved version of the SIR, called the cumulative slicing estimation (CUME) method, that does not require selecting an optimal number of slices. We provide a general framework in which to analyze the phase transitions of the CUME method. We show that, without the sparsity assumption, the CUME method is consistent if and only if $p/n \to 0$, where p denotes the covariate dimension, and n denotes the sample size. If we include certain sparsity assumptions, then the thresholding estimate for the CUME method is consistent as long as $\log(p)/n \to 0$. We demonstrate the superior performance of the proposed method using extensive numerical experiments.

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

1. Introduction

1.1. Background

Recent advances in information and technology now allow us to collect big data in scientific areas including genome sequencing, biomedical imaging, social media analysis, and high-frequency finance. Big data are often high or ultrahigh dimensional (Fan, Han and Liu (2014)). For instance, the Framingham heart study records many features related to heart disease and health status, including

Corresponding author: Liping Zhu, 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.

genetic background, measurements from blood analyses, immune system status, nutrition, alcohol/tobacco/drug consumption, operations, treatments, and diagnosed diseases. These features are high dimensional, which poses significant challenges for classic statistical methods.

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

$$Y \perp\!\!\!\perp \mathbf{x} \mid (\mathbf{B}^{\mathrm{T}}\mathbf{x}). \tag{1.1}$$

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

A popular tool used to recover $\mathcal{S}_{Y|\mathbf{x}}$ is the sliced inverse regression (Li (1991), SIR). Let $\mathbf{\Sigma} \overset{\text{def}}{=} \text{var}(\mathbf{x})$ and $\mathbf{\Lambda}_{\text{SIR}} \overset{\text{def}}{=} \text{var}\{E(\mathbf{x} \mid Y)\}$. The SIR identifies $\mathcal{S}_{Y|\mathbf{x}}$ using span($\mathbf{\Sigma}^{-1}\mathbf{\Lambda}_{\text{SIR}}\mathbf{\Sigma}^{-1}$), the column space of $\mathbf{\Sigma}^{-1}\mathbf{\Lambda}_{\text{SIR}}\mathbf{\Sigma}^{-1}$. Li (1991) proposed a slicing procedure to estimate $\mathbf{\Lambda}_{\text{SIR}}$ that divides the range of the observed response values into H slices, and then calculates the sample average of the concomitant covariates within each slice. Owing to its computational efficiency, simplicity, and generality, the slicing estimation was later applied to other SDR methods (Cook and Weisberg (1991); Li and Wang (2007)). The 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, Miao and Peng (2006), Zhong et al. (2012) and Jiang and Liu (2014) proved the consistency of the slicing estimation when $p = o(n^{1/2})$ and H is a fixed number. Recently, Lin, Zhao and Liu (2018b) showed that the slicing estimate of the SIR is consistent when $p = o(n)$. However, the convergence rate derived in Lin, Zhao and Liu (2018b) varies with the number of slices. This is often undesirable, giving that determining the optimal number of slices remains an open problem in the SDR literature.

1.2. Disadvantages of the SIR

When the covariate is high/ultrahigh dimensional, the SIR encounters several theoretical and practical problems. To the best of our knowledge, the consistency of the slicing estimation for relatively large p (e.g., $\log(p) = o(n)$) remains unknown in the SDR literature. From a theoretical perspective, an asymptotic study of the consistency of the SIR with large p is not straightforward, because the convergence rate also depends on H, and no data-driven selection schemes are yet available to determine H. From a practical perspective, our empirical results indicate that the performance of the SIR may depend on H when p is relatively large, but that this is not the case when p is small. However, if H is too small, the pattern between Y and $\mathbf{x}$ may be averaged out within each slice. In contrast, if H is too large, the SIR may suffer from significant inner-slice variation. Therefore, determine an optimal H is an important issue when p is relatively large.

We demonstrate using simulated examples that the SIR may be sensitive to the selection of H when p is relatively large. Here, we adapt the following models, common in the SDR literature. In particular, models (1.2) and (1.5) are used by Li (1991), model (1.3) is used by Zhu, Zhu and Feng (2010), and models (1.4) and (1.7) are used by Lin, Zhao and Liu (2018b). The covariates $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ are drawn independently from the standard normal distribution, and ε follows the standard normal distribution. The response variable Y is generated from the following models:

$$Y = X_1 + X_2 + X_3 + X_4 + 0.5\varepsilon; \tag{1.2}$$

$$Y = \sin(X_1 + 0.5\varepsilon); \tag{1.3}$$

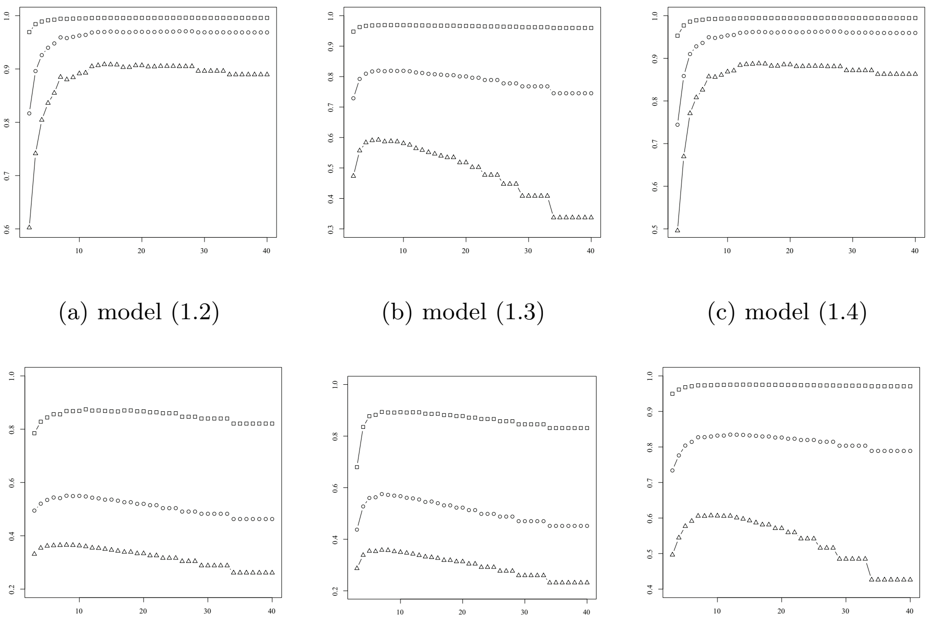
$$Y = \frac{(X_1 + X_2 + X_3)^3}{2} + 0.5\varepsilon; \tag{1.4}$$

$$Y = \frac{1 + X_1}{0.5 + (1.5 + X_2)^2} + 0.2\varepsilon; \tag{1.5}$$

$$Y = 4\sin(X_1 + X_2) + \exp(X_3 + X_4) + 0.2\varepsilon; \tag{1.6}$$

$$Y = (X_1 + \cdots + X_7)\exp(X_8 + X_9) + 0.2\varepsilon. \tag{1.7}$$

In general, estimating $\mathcal{S}_{Y|\mathbf{x}}$ for one-dimensional models is easier than doing so



(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 1. The mean values of the $r^2(d)$ values over 1,000 repetitions for models (1.2)−(1.7). The squares denotes $p = 10$, the circles denotes $p = 50$, and the triangles denotes $p = 100$. The horizontal axis shows the number of slices H, which ranges from 2 to 40 in models (1.2)−(1.4) and from 3 to 40 in models (1.5)−(1.7). The vertical axis shows the mean values of the trace correlation $r^2(d)$.

for multiple-dimensional models. We measure the accuracy of the SIR estimate using a trace correlation, as proposed by Ferré (1998). Let $\mathbf{B}$ be an underlying true basis of $\mathcal{S}_{Y|\mathbf{x}}$, and $\widehat{\mathbf{B}}$ be an estimated basis obtained using the SIR method. Define $\mathbf{P} \overset{\text{def}}{=} \mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}$ and $\widehat{\mathbf{P}} \overset{\text{def}}{=} \widehat{\mathbf{B}}(\widehat{\mathbf{B}}^{\mathrm{T}}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^{\mathrm{T}}$. We denote trace($\mathbf{A}$) as the trace of a matrix $\mathbf{A}$, and d as the structural dimension of $\mathcal{S}_{Y|\mathbf{x}}$. The trace correlation is defined as $r^2(d) \overset{\text{def}}{=} \text{trace}(\widehat{\mathbf{P}}\mathbf{P})/d$, and ranges from zero to one. Larger $r^2(d)$ values indicate a more accurate estimation. In particular, $r^2(d) = 1$ if the estimated $\mathcal{S}_{Y|\mathbf{x}}$ and the true $\mathcal{S}_{Y|\mathbf{x}}$ are identical, and $r^2(d) = 0$ if these two spaces are orthogonal to each other. Here, we examine how the mean $r^2(d)$ varies with the number of slices H for different dimensions p. The simulation results based on 1,000 repetitions are summarized in Figure 1 (A)−(F) where the sample size is 200.

Unsurprisingly, when p is small, say $p = 10$, the SIR exhibits a very stable performance, indicated by $r^2(d)$ values that appear constant for a wide range of slice numbers. However, when p is relatively large, the resulting pattern about H in the SIR is quite different. The performance of the 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 $H = 40$. However, the peak occurs at $H = 6$, with $r^2(d) = 59.21\%$. Thus, the results for small p may not carry over to the case of large p. Thus, when p is large, we need an SDR method that does not require selecting an appropriate number of slices.

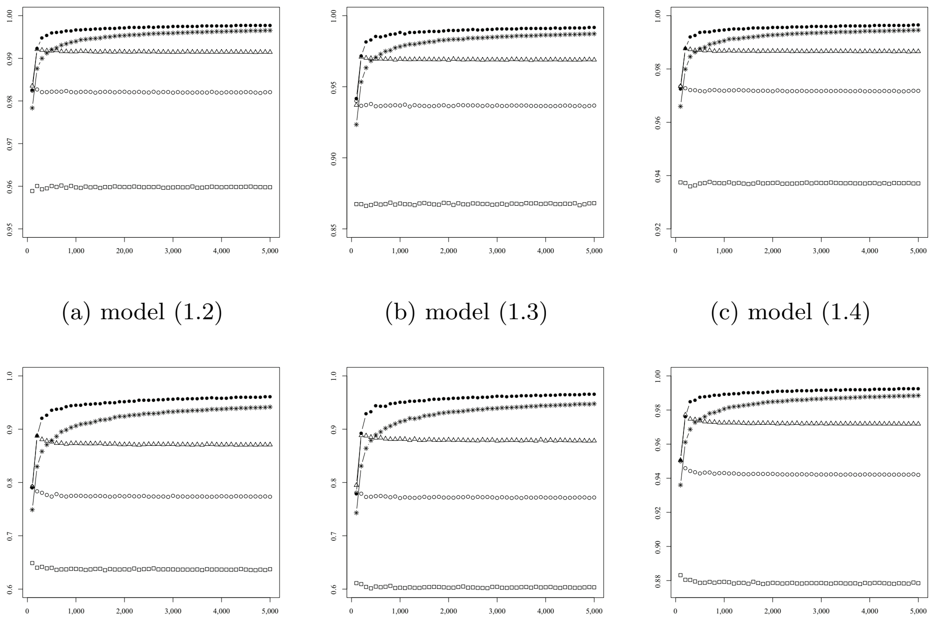
We also observe that the SIR deteriorates sharply as p increases in all models, indicating that the SIR estimate may not maintain consistency if p is relatively large. Thus, a consistent estimate is highly desirable for SDR methods in ultrahigh-dimensional settings.

1.3. Contributions to the literature

We propose an improved version of the SIR, called the cumulative slicing estimation (Zhu, Zhu and Feng (2010), CUME) method. Unlike the SIR, the CUME method is independent of the number of slices H. It recovers $\mathcal{S}_{Y|\mathbf{x}}$ using $\text{span}(\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1})$, where $\mathbf{\Lambda} \overset{\text{def}}{=} E\{\mathbf{m}(Y)\mathbf{m}^{\mathrm{T}}(Y)\}$ and $\mathbf{m}(y) \overset{\text{def}}{=} \text{cov}\{\mathbf{x}, I(Y \le y)\}$. Note that $\mathbf{m}(y) = \text{cov}\{E(\mathbf{x} \mid Y), I(Y \le y)\}$, indicating that both the CUME and the SIR methods use the inverse regression $E(\mathbf{x} \mid Y)$ to identify $\mathcal{S}_{Y|\mathbf{x}}$. The difference between the two is that the CUME method does not depend on the number of slices H. We provide a general framework in which to analyze the phase transitions of the CUME method. We show that, without the sparsity assumption, the CUME method is consistent if $p = o(n)$. If both $\mathbf{\Sigma}^{-1}$ and $\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}$ are sparse matrices, we suggest a thresholding estimate for $\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}$. We show that the sparse estimate of $\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}$ is consistent as long as $\log(p) = o(n)$.

Recently, Lin, Zhao and Liu (2018a) introduced the Lasso-SIR algorithm, which can be used to obtain a sparse estimate of the central subspace. The resulting estimate achieves an optimal convergence rate under certain sparsity conditions when $p = o(n^2\lambda^2)$, where λ is the generalized signal-to-noise ratio. In 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 specifying the number of slices, whereas our method does not. In this sense, our results improve upon theirs significantly.

Next, we use simulated examples to show that, for the CUME method to be consistent, $p = o(n)$ is the largest divergence rate. We still adopt models



(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 1,000 repetitions for models (1.2)−(1.7). The line marked with squares denotes $(p/n) = 0.2$; the line marked with hollow points denotes $(p/n) = 0.1$; the line marked with triangles denotes $(p/n) = 0.05$; the line marked with stars denotes $(p/n) = \{\log(n/5)\}^{-2}$; and the line marked with solid points denotes $(p/n) = \{\log(n)\}^{-2}$. The horizontal axis shows the sample size n, ranging from 100 to 5,000, and the vertical axis shows the mean values of the trace correlation $r^2(d)$.

(1.2)−(1.7), and generate the covariates and the error terms in the same manner. We consider different sample sizes n and covariate dimensions p, such that their ratio (p/n) is equal to 0.2, 0.1, 0.05, $\{\log(n/5)\}^{-2}$, or $\{\log(n)\}^{-2}$. The sample size n ranges from 100 to 5,000. We repeat our experiments 1,000 times. Again, we use the mean values of the trace correlation $r^2(d)$ to illustrate the performance of the CUME method. The simulation results are summarized in Figure 2 (A)−(F).

Figure 2 clearly shows that, if p is proportional to n (e.g., $(p/n) = 0.05, 0.1$, and 0.2), the mean values of $r^2(d)$ appear very flat as n increases. In this case, the CUME method cannot be consistent as n diverges, because the $r^2(d)$ values should get closer to one as n increases. In contrast, if $(p/n) \to 0$ (e.g., $(p/n) = \{\log(n/5)\}^{-2}$ 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 $r^2(d)$ is 74.88% when $n = 100$, and 94.17% when n increases to 5,000. This exhibits a clear pattern that, as long as $(p/n) \to 0$ when $n \to \infty$, the CUME method converges. Because (p/n) is a constant as $n \to \infty$, the CUME method cannot converge without the sparsity assumption. These simulation results demonstrate that the CUME method 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 examine the consistency of the estimated CUME matrix. We show that the classical moment estimate of the CUME matrix is consistent for $p = o(n)$ without the sparsity assumption, and that the regularized estimate is consistent for $\log(p) = o(n)$ with the sparsity assumption. In Section 3, we investigate the finite-sample performance of the proposed method using comprehensive simulations and real-world data. Section 4 concludes the paper. All technical details are relegated to the online Supplementary Material.

2. Main Results

2.1. Definitions and notation

Suppose $\{(\mathbf{x}_i, Y_i), i = 1, \ldots, n\}$ is a random sample of $(\mathbf{x}, Y)$. For a $p \times q$ matrix $\mathbf{A}_{p\times q}$, let $\mathrm{span}(\mathbf{A})$ be the space spanned by the columns of $\mathbf{A}$, $\mathrm{trace}(\mathbf{A})$ denote the trace of $\mathbf{A}$, $\mathrm{rank}(\mathbf{A})$ be the rank of $\mathbf{A}$, and $\lambda_{\max}(\mathbf{A})$ and $\lambda_{\min}(\mathbf{A})$ denote the maximum and the minimum eigenvalues of $\mathbf{A}$, respectively. Let $\lambda_{\max}(\mathbf{A}) = \lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots \geq \lambda_q(\mathbf{A}) = \lambda_{\min}(\mathbf{A})$, where $\lambda_k(\mathbf{A})$ denotes the kth-largest principal eigenvalue of $\mathbf{A}$. We may simply use λ_k in place of $\lambda_k(\mathbf{A})$ when it is sufficiently clear from the context. Let $\|\mathbf{A}\|_F \overset{\text{def}}{=} \{\mathrm{trace}(\mathbf{A}^{\mathrm{T}}\mathbf{A})\}^{1/2}$ be the Frobenius norm, and let $\|\mathbf{A}\|$ be the spectral norm of $\mathbf{A}$. Specifically,

$$\|\mathbf{A}\| \overset{\text{def}}{=} \sup_{\mathbf{a}^{\mathrm{T}}\mathbf{a}=1} (\mathbf{a}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\mathbf{A}\mathbf{a})^{1/2} = \lambda_{\max}^{1/2}(\mathbf{A}^{\mathrm{T}}\mathbf{A}).$$

Let $\mathbf{A}_{k,l}$ be the (k,l)th entry of $\mathbf{A}$; that is, $\mathbf{A} = (\mathbf{A}_{k,l})_{p\times q}$. Define

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

Denote $\mathbf{I}_{p\times p}$ as the $p \times p$ identity matrix. Let $I(A)$ be an indicator function, equal to one if event A is true, and zero otherwise; here, $\mathrm{pr}(A) = E\{I(A)\}$ 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 can take different values, depending on the context.

The goal of an SDR is to identify and recover $\mathcal{S}_{Y|\mathbf{x}}$. Let $\mathbf{B}_{p\times d} \in \mathbb{R}^{p\times d}$ be a basis of $\mathcal{S}_{Y|\mathbf{x}}$ and $\widehat{\mathbf{B}}_{p\times d} \in \mathbb{R}^{p\times d}$ be an estimated basis, where $\mathbf{P} \overset{\text{def}}{=} \mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}$ and $\widehat{\mathbf{P}} \overset{\text{def}}{=} \widehat{\mathbf{B}}(\widehat{\mathbf{B}}^{\mathrm{T}}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^{\mathrm{T}}$. The projection matrix $\mathbf{P}$, rather than its basis $\mathbf{B}$, is unique and identifiable. Therefore, to quantify how well $\widehat{\mathbf{B}}$ estimates $\mathcal{S}_{Y|\mathbf{x}}$, it is reasonable to use the following three criteria:

1. The spectral norm $\|\widehat{\mathbf{P}} - \mathbf{P}\|$;
2. The Frobenius norm $\|\widehat{\mathbf{P}} - \mathbf{P}\|_F$;
3. The trace correlation $r^2(d) \overset{\text{def}}{=} \text{trace}(\widehat{\mathbf{P}}\mathbf{P})/d$.

Note that the Frobenius norm is equivalent to the the trace correlation (Ferré (1998)), in that $\|\widehat{\mathbf{P}} - \mathbf{P}\|_F^2 = 2d\{1 - r^2(d)\}$. Both the spectral norm $\|\widehat{\mathbf{P}} - \mathbf{P}\|$ and the Frobenius norm $\|\widehat{\mathbf{P}} - \mathbf{P}\|_F$ are nonnegative and have upper bounds, where a smaller value indicates a more accurate estimate. The trace correlation $r^2(d)$ ranges from zero to one, where a larger value indicates a better estimate. In the following section, we examine the convergence rate of $\widehat{\mathbf{P}}$ 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 $\widehat{\mathbf{B}}$, an estimated basis of $\mathcal{S}_{Y|\mathbf{x}}$, because its estimation is free of tuning parameters. Recall that $\mathbf{\Sigma} = \text{var}(\mathbf{x})$ and $\mathbf{\Lambda} = E\{\mathbf{m}(Y)\mathbf{m}^{\mathrm{T}}(Y)\}$, where $\mathbf{m}(y) = \text{cov}\{\mathbf{x}, I(Y \le y)\}$. We estimate $\mathbf{m}(y)$ using

$$\widehat{\mathbf{m}}(y) \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})\, I(Y_i \le y), \quad \overline{\mathbf{x}} \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n} \mathbf{x}_i,$$

and estimate $\mathbf{\Lambda}$ and $\mathbf{\Sigma}$ using

$$\widehat{\mathbf{\Lambda}} \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n} \widehat{\mathbf{m}}(Y_i)\widehat{\mathbf{m}}^{\mathrm{T}}(Y_i), \quad \text{and} \quad \widehat{\mathbf{\Sigma}} \overset{\text{def}}{=} n^{-1}\sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})\,(\mathbf{x}_i - \overline{\mathbf{x}})^{\mathrm{T}}. \tag{2.1}$$

The estimated basis $\widehat{\mathbf{B}}$ is composed of the first d principal eigenvectors of $\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Sigma}}^{-1}$.

To state the consistency of $\widehat{\mathbf{B}}$, we make the following assumptions.

(A1) : Define $d \overset{\text{def}}{=} \text{rank}(\mathbf{\Lambda})$ and $\lambda_d(\mathbf{\Lambda})$ as the dth principal eigenvalue of $\mathbf{\Lambda}$ (which is also the smallest nonzero principal eigenvalue of $\mathbf{\Lambda}$). Assume $\lambda_d(\mathbf{\Lambda}) \ge c_0^{-1}$.

(A2) : Assume $c_0^{-1} \le \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) \le c_0$, where $\lambda_{\min}(\mathbf{\Sigma})$ and $\lambda_{\max}(\mathbf{\Sigma})$ are the smallest and largest eigenvalues, respectively, of $\mathbf{\Sigma}$.

(A3) : Assume the covariate vector $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ is subGaussian. That is, for any unit-length vector $\mathbf{e}$, $\mathrm{pr}\left(\mid \mathbf{e}^{\mathrm{T}}\mathbf{x} \mid \ge t\right) \le \exp(1 - c_0 t^2)$, for all $t \ge 0$.

Assumption (A1) requires that the nonzero eigenvalues of $\mathbf{\Lambda}$ be bounded from below. It ensures that the magnitudes of signals, represented by nonzero eigenvalues of $\mathbf{\Lambda}$, are detectable. Assumption (A2) is widely assumed in the literature on high-dimensional covariance matrix estimation; see, for example, Bickel and Levina (2008) and Cai, Liu and Luo (2011). This assumption allows covariates to be correlated, as long as their covariance matrix is nonsingular. Assumption (A3) requires that the covariates be subGaussian, 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. $\|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|^2 = O_p(p/n)$ *and* $\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\|^2 = O_p\left\{\max(p, \log n)/n\right\}$;
2. $\|\widehat{\mathbf{P}} - \mathbf{P}\|_F^2 = (2d)\left\{1 - r^2(d)\right\} = O_p\left\{\max\left(p, \log n\right)/n\right\}$;
3. $\|\widehat{\mathbf{P}} - \mathbf{P}\|^2 = O_p\left\{\max\left(p, \log n\right)/n\right\}$.

We first give some brief comments on Theorem 1. The subGaussian assumption is widely used in studies on the consistency of the sample covariance matrix $\widehat{\mathbf{\Sigma}}$; see, for example, Vershynin (2012), Bunea and Xiao (2015) and Koltchinskii and Lounici (2017). In the present context, an important contribution of this study is that we derive the convergence rates of both $\widehat{\mathbf{\Lambda}}$ and $\widehat{\mathbf{P}}$ under the spectral norm. Moreover, one may wonder why $\widehat{\mathbf{\Sigma}}$ and $\widehat{\mathbf{\Lambda}}$ have different convergence rates. The rate $(\log n/n)$ in $\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\|^2$ appears in the derivation of the uniform convergence rate of $\widehat{\mathbf{m}}(y)$. We believe the rate $O_p\{\max\left(p, \log n\right)/n\}$ may be refined to $O_p(p/n)$. In high-dimensional data analyses, it is reasonable to expect that p is greater than $\log n$. Accordingly, $\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\|^2 = O_p(p/n)$. In other words, the presence of $\log n/n$ does not have a significant effect on the convergence rate of $\|\widehat{\mathbf{\Lambda}} - \mathbf{\Lambda}\|^2$ when the covariate dimension is high. The second statement connects the Frobenius norm with the trace correlation. In particular, $\|\widehat{\mathbf{P}} - \mathbf{P}\|_F = (2d)^{1/2}\left\{1 - r^2(d)\right\}^{1/2}$. The last two statements indicate that the Frobenius norm and the spectral norm of $\widehat{\mathbf{P}} - \mathbf{P}$ have identical convergence rates.

Zhu, Zhu and Feng (2010) and Jiang and Liu (2014) derived the convergence rate of $\widehat{\mathbf{B}}$ when $p = o(n^{1/2})$. We improve upon their results substantially by finding the rate when $p = o(n)$, which is the largest p one can handle without a sparsity condition. The key to improving the convergence rate of $\|\widehat{\mathbf{P}} - \mathbf{P}\|_F$ from $O_p(p/n^{1/2})$ (Zhu, Miao and Peng (2006); Zhu, Zhu and Feng (2010); Jiang and Liu (2014)) to $O_p(p^{1/2}/n^{1/2})$ is that we use an improved Davis−Kahan $\sin\theta$ theorem (Yu, Wang and Samworth (2015)). In particular,

$$\|\widehat{\mathbf{P}} - \mathbf{P}\|_F \leq \frac{4\min\{d^{1/2}\|\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}\|, \|\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}\|_F\}}{\lambda_d(\mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1})}.$$

In general, $\|\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}\| \leq \|\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Lambda}}\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{\Lambda}\mathbf{\Sigma}^{-1}\|_F$, where d is a small number, given the purpose of dimension reduction. The improved Davis−Kahan $\sin\theta$ theorem accounts for the significantly improved convergence rate.

Recently, Lin, Zhao and Liu (2018b, Thm. 1) studied the consistency of the SIR when $p = o(n)$, showing that

$$\|\widehat{\mathbf{\Lambda}}_{\text{SIR}} - \mathbf{\Lambda}_{\text{SIR}}\| = O_p\left(\frac{1}{H^{\vartheta}} + \frac{H^2 p}{n} + \sqrt{\frac{H^2 p}{n}}\right),$$

where ϑ is a nonnegative constant. The above convergence rate indicates that the optimal number of slices is

$$H = O\left\{\left(\frac{p}{n}\right)^{-1/(2(\vartheta+1))}\right\}. \tag{2.2}$$

Accordingly, the resulting optimal convergence rate is

$$\|\widehat{\mathbf{\Lambda}}_{\text{SIR}} - \mathbf{\Lambda}_{\text{SIR}}\| = O_p\left\{\left(\frac{p}{n}\right)^{\vartheta/(2(\vartheta+1))}\right\}. \tag{2.3}$$

The above convergence rate is slower than that of the CUME method, derived in Theorem 1. Moreover, determining an optimal H that satisfies (2.2) is an issue, partly because ϑ is unknown. It is thus encouraging that avoiding a slicing estimation not only overcomes this longstanding computational issue, but also means the CUME method possesses a better convergence rate than that of the 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 $p/n \to \gamma$, 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 using simulated examples when $p/n \to \gamma$, for some $\gamma \in (0, 1)$. We further demonstrate this inconsistency issue by means of the following analytical example.

Example 1. Assume $Y = X_1 + \sigma\varepsilon$, where $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ follows a multivariate standard normal distribution, and ε is standard normal. In this example, $\mathbf{\Sigma} \overset{\text{def}}{=} \mathrm{var}(\mathbf{x}) = \mathbf{I}_{p\times p}$, where $\mathbf{I}_{p\times p}$ denotes the $p \times p$ identity matrix. Because $\mathbf{x}$ is standardized, we can simply estimate the basis matrix $\mathbf{B}$ using the first d principal eigenvectors of $\widehat{\mathbf{\Lambda}}$. With a slight abuse of notation, we denote the estimated basis as $\widehat{\mathbf{B}}$. Define $\mathbf{P} = \mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}$ and $\widehat{\mathbf{P}}\widehat{\mathbf{B}}(\widehat{\mathbf{B}}^{\mathrm{T}}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^{\mathrm{T}}$. In the Supplementary Material, we show that

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

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, $\widehat{\mathbf{\Sigma}}$ is no longer invertible, even when $\mathbf{\Sigma}$ is nonsingular. To address this issue, we turn to sparsity assumptions, and propose sparse solutions, denoted as $\widehat{\mathbf{\Omega}}_s$ and $\widehat{\mathbf{\Theta}}_s$, to estimate $\mathbf{\Omega} \overset{\text{def}}{=} \mathbf{\Sigma}^{-1}$ and $\mathbf{\Theta} \overset{\text{def}}{=} \mathbf{\Omega}\mathbf{\Lambda}\mathbf{\Omega}$, respectively. Let $\widehat{\mathbf{B}}_s$ be composed of the first d principal eigenvectors of $\widehat{\mathbf{\Theta}}_s$. In this section, we derive the consistency of $\widehat{\mathbf{B}}_s$ under certain sparsity assumptions. Define $\widehat{\mathbf{P}}_s = \widehat{\mathbf{B}}_s(\widehat{\mathbf{B}}_s^{\mathrm{T}}\widehat{\mathbf{B}}_s)^{-1}\widehat{\mathbf{B}}_s^{\mathrm{T}}$. We study the consistency of $\widehat{\mathbf{B}}_s$ under the Frobenius norm $\|\widehat{\mathbf{P}}_s - \mathbf{P}\|_F$, spectral norm $\|\widehat{\mathbf{P}}_s - \mathbf{P}\|$, and trace correlation $r^2(d) = \mathrm{trace}(\widehat{\mathbf{P}}_s\mathbf{P})/d$.

We suggest an estimation of the precision matrix $\mathbf{\Omega}$ first, and then propose a sparse estimation for $\mathbf{\Theta}$ based on the sparse solution $\widehat{\mathbf{\Omega}}_s$.

Estimation of $\mathbf{\Omega}$: Estimations of precision matrices have been studied extensively in the literature; see, for example, Meinshausen and Bühlmann (2006), Cai, Liu and Luo (2011) and Zhang and Zou (2014), and a recent review article by Fan, Liao and Liu (2016). In this work, we adapt the constrained ℓ_1 minimization for the inverse covariance matrix estimation (CLIME) proposed by Cai, Liu and Luo (2011). The CLIME method is implemented as follows. For a given tuning parameter λ_{1n}, let $\widehat{\mathbf{\Omega}}$ be the solution set of the following optimization problem:

$$\widehat{\mathbf{\Omega}} \in \arg\min_{\mathbf{\Omega}} \|\mathbf{\Omega}\|_1, \quad \text{subject to} \quad \|\widehat{\mathbf{\Sigma}}\mathbf{\Omega} - \mathbf{I}\|_\infty \le \lambda_{1n},$$

where $\widehat{\mathbf{\Sigma}}$ is defined in (2.1). The above solution $\widehat{\mathbf{\Omega}}$ is not symmetric, in general. To obtain a symmetric estimate, the CLIME estimator $\widehat{\mathbf{\Omega}}_s$ is defined as $\widehat{\mathbf{\Omega}}_s \overset{\text{def}}{=} \big(\widehat{\mathbf{\Omega}}_{s,k,l}\big)$, where

$$\widehat{\mathbf{\Omega}}_{s,k,l} = \widehat{\mathbf{\Omega}}_{s,l,k} = \widehat{\mathbf{\Omega}}_{k,l} I\big(|\widehat{\mathbf{\Omega}}_{k,l}| \le |\widehat{\mathbf{\Omega}}_{l,k}|\big) + \widehat{\mathbf{\Omega}}_{l,k} I\big(|\widehat{\mathbf{\Omega}}_{k,l}| > |\widehat{\mathbf{\Omega}}_{l,k}|\big).$$

In other words, we select $\widehat{\mathbf{\Omega}}_{k,l}$ or $\widehat{\mathbf{\Omega}}_{l,k}$ with the smallest magnitude. The resultant estimate $\widehat{\mathbf{\Omega}}_s$ is symmetric and, more importantly, positive definite with high probability. By assuming the covariates have exponential-type tails, and $\lambda_{1n} = C_1(\log p/n)^{1/2}$ for some generic constant C_1, Cai, Liu and Luo (2011) show that

$$\|\widehat{\mathbf{\Omega}}_s - \mathbf{\Omega}\| = O_p\left\{M^{2-2q} s_1(p)\left(\frac{\log p}{n}\right)^{(1-q)/2}\right\}$$

holds uniformly for

$$\begin{aligned}&\mathbf{\Omega} \in \mathcal{U}_1\{q, s_1(p)\}\\ &\overset{\text{def}}{=} \left\{\mathbf{\Omega} : \mathbf{\Omega} > 0, \|\mathbf{\Omega}\|_1 \le M \text{ and } \max_{1\le k\le p} \sum_{l=1}^{p} |\mathbf{\Omega}_{k,l}|^q \le s_1(p)\right\},\end{aligned} \tag{2.4}$$

for some $0 \le q < 1$. For brevity, we assume $\|\mathbf{\Omega}\|_1 \le c_0$. Then

$$\|\widehat{\mathbf{\Omega}}_s - \mathbf{\Omega}\| = O_p\left\{s_1(p)\left(\frac{\log p}{n}\right)^{(1-q)/2}\right\}. \tag{2.5}$$

Next, we suggest a thresholding estimate for $\mathbf{\Theta}$.

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

$$\widehat{\mathbf{\Theta}}_s \overset{\text{def}}{=} \big(\widehat{\mathbf{\Theta}}_{s,k,l}\big)_{p\times p} = \big\{\widehat{\mathbf{\Theta}}_{k,l}\ I(|\widehat{\mathbf{\Theta}}_{k,l}| \ge \lambda_{2n})\big\}_{p\times p},$$

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

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

Note 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 no greater than $s_1(p)$ or $s_2(p)$. The class $\mathcal{U}_1\{q, s_1(p)\}$ was introduced by Bickel and Levina (2008) and Cai, Liu and Luo (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 $\mathbf{\Theta}$ in models (1.2)−(1.7) is sufficiently sparse, with its upper-left block submatrix being nonzero. In effect, the class $\mathcal{U}_2\{q, s_2(p)\}$ includes many common dimension-reduction models (Zhu, Zhu and Feng (2010, Thm. 1)).

Theorem 2. *Assume conditions* (A1)−(A3) *and* (2.4)*. 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* $n \to \infty$*,*

$$\|\widehat{\mathbf{\Theta}}_s - \mathbf{\Theta}\| = O_p \left\{ s_1^{1-q}(p) s_2(p) \left(\frac{\log p}{n} \right)^{(1-q)^2/2} \right\}. \tag{2.6}$$

Theorem 3 states the consistency of $\widehat{\mathbf{B}}_s$.

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

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

Theorem 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. This generalizes the applicability of the CUME method to ultrahigh-dimensional data.

Tuning Parameter Selection: It remains to choose the appropriate λ_{1n} and λ_{2n} values for the thresholding regularized CUME method. The selector for λ_{1n} is discussed extensively by Cai, Liu and Luo (2011). Simply, λ_{1n} is decided under a likelihood loss function coupled with five-fold cross-validation. We suggest choosing λ_{2n} using five-fold cross validation such that the distance correlation (Székely, Rizzo and Bakirov (2007); Székely and Rizzo (2009)) between $(\widehat{\mathbf{B}}_s^{\mathrm{T}}\mathbf{x})$ and Y is maximized. The distance correlation retains the model$-$free flavor of the SDR, because it can measure the nonlinear dependence between Y and $(\widehat{\mathbf{B}}_s^{\mathrm{T}}\mathbf{x})$. 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 use the remaining four subsamples as the training set. For each λ_{2n}, we obtain an estimate $\widehat{\boldsymbol{\Theta}}_s$ and, accordingly, $\widehat{\mathbf{B}}_s$, using the training set. We calculate the distance correlation between $(\widehat{\mathbf{B}}_s^{\mathrm{T}}\mathbf{x})$ and Y using the test set. The cross-validation procedure is repeated five times, where each subsample is used exactly once as a test set. The five distance correlations are averaged to produce a single estimation. We choose λ_{2n}, which maximizes the average of the five distance correlations. Our limited experience indicates that this procedure is effective.

3. Numerical Studies

3.1. Simulations

We illustrate the finite-sample performance of our proposed sparse estimate $\widehat{\mathbf{B}}_s$ using simulations. We also compare the proposed method with the classical CUME method, that is, the SIR method, using different numbers of slices. We use the trace correlation $r^2(d)$ to assess the the finite-sample performance of the methods. We adapt models (1.2)$-$(1.7) in our simulations. Throughout, we draw $\mathbf{x} = (X_1, \ldots, X_p)^{\mathrm{T}}$ from a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$, and draw ε independently from a 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 $p = 1{,}000$ and 5,000. We set $\boldsymbol{\Sigma} = \mathbf{I}_{p\times p}$, $\boldsymbol{\Sigma} = (0.2^{|k-l|})_{p\times p}$, and $\boldsymbol{\Sigma} = (0.5^{|k-l|})_{p\times p}$ in the first, second, and last scenarios, respectively. In the first two scenarios, we directly implement our proposed sparse estimate procedure, the CUME and the SIR method, to estimate $\mathcal{S}_{Y|\mathbf{x}}$. In the third scenario, we first implement the sure independent ranking and screening method (Zhu 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 $p_0 = [n/\log n]$, $2[n/\log n], \ldots, 5[n/\log n]$, corresponding 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 1,000 times, and report the mean and standard deviation of the $r^2(d)$ values. Table 1 presents the results for $\mathbf{\Sigma} = (0.2^{|k-l|})_{p\times p}$. Additional simulations are relegated to the Supplementary Material.

The simulation results indicate that the 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 $\mathbf{\Sigma} = (0.2^{|k-l|})_{p\times p}$, 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 number of slices appears more substantial than in the first two scenarios. For example, in model (1.2), with $p = 1{,}000$ 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 $p = 5{,}000$ 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 number of slices in the SIR has a nonignorable effect on its performance when the covariate dimension is relatively large. Recall that the CUME method does not rely on choosing an optimal number of slices. Nevertheless, 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 the CUME method is 0.992 when $p = 10$, and is 0.551 when $p = 150$. In contrast, our proposed sparse estimate of the CUME matrix is stable across all scenarios. The $r^2(d)$ values obtained using the proposed method are all larger than 0.950 in the one-dimensional models, and are all greater than 0.700 in the two-dimensional models. This is in line with our expectation that estimating a two-dimensional $\mathcal{S}_{Y|\mathbf{x}}$ is more difficult than estimating a one-dimensional $\mathcal{S}_{Y|\mathbf{x}}$.

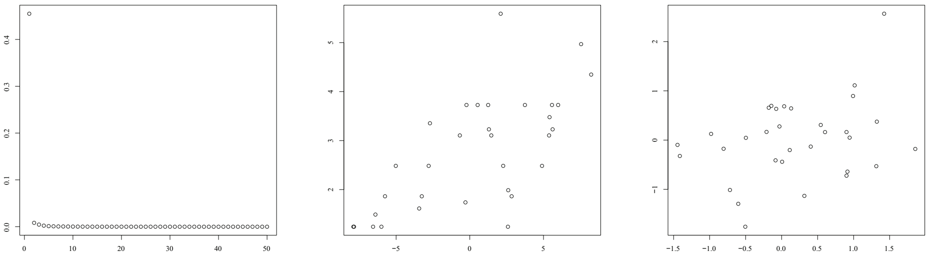
3.2. Real-data analysis

We demonstrate our proposed sparse estimate for the CUME method using breast cancer data collected by Van't Veer et al. (2002). In this study, 24,481 gene expression levels were collected from 97 lymph node-negative breast cancer patients. We remove observations that contain missing values, leaving 24,188 gene expression levels. We aim to predict the tumor size based on levels. Because the covariates are ultrahigh dimensional, we first apply the sure independent ranking and screening procedure of Zhu et al. (2011) to select the top 50 gene expression levels, which we expect will best predict the tumor size. We split the data set into two sets: a training set containing 65 observations, and a test set

Table 1. The averages (standard deviations) of the trace correlations (× 100) for Scenario 2, where SIR$_k$ denotes SIR with k slices, and $\mathbf{\Sigma} = (0.2^{|k-l|})_{p\times p}$.

	p	CUME	SIR$_2$	SIR$_5$	SIR$_{10}$	SIR$_{20}$	NEW
(1.2)	10	98.9(0.6)	95.7(2.0)	98.8(0.6)	99.3(0.4)	99.5(0.3)	98.4(1.6)
	50	93.0(2.0)	76.4(4.8)	92.5(2.0)	95.5(1.2)	96.5(0.9)	98.7(1.7)
	100	79.0(5.0)	51.9(6.3)	79.6(4.2)	87.1(2.9)	89.4(2.4)	98.6(1.7)
	200	-	-	-	-	-	98.4(2.0)
	300	-	-	-	-	-	98.4(2.0)
(1.3)	10	96.8(1.8)	94.2(3.1)	96.5(1.9)	96.6(2.0)	96.4(2.1)	96.3(3.9)
	50	81.8(5.2)	71.4(6.8)	80.4(5.5)	80.5(5.6)	78.6(6.4)	98.2(5.6)
	100	59.2(8.6)	45.0(8.3)	56.7(9.0)	55.6(9.8)	48.8(11.9)	98.4(6.1)
	200	-	-	-	-	-	99.0(5.1)
	300	-	-	-	-	-	99.0(4.9)
(1.4)	10	98.6(0.8)	94.2(2.7)	98.7(0.7)	99.2(0.4)	99.4(0.3)	98.3(1.9)
	50	90.3(2.6)	69.8(5.5)	91.5(2.0)	95.0(1.3)	96.0(1.1)	99.0(1.6)
	100	72.6(6.0)	43.9(6.6)	77.8(4.4)	85.8(3.1)	88.2(2.7)	98.9(1.9)
	200	-	-	-	-	-	98.9(1.8)
	300	-	-	-	-	-	98.9(1.7)
(1.5)	10	88.7(5.4)	94.4(3.0)	85.1(7.3)	88.3(6.2)	88.0(6.3)	93.8(7.2)
	50	58.6(5.5)	74.7(6.2)	54.1(6.4)	56.6(7.2)	54.1(7.4)	93.4(8.9)
	100	39.4(4.3)	58.5(9.0)	36.1(4.4)	36.4(4.8)	32.8(4.9)	92.6(9.6)
	200	-	-	-	-	-	91.0(10.2)
	300	-	-	-	-	-	90.7(11.4)
(1.6)	10	85.5(7.2)	58.8(4.1)	84.2(7.7)	85.1(8.0)	82.8(9.7)	91.7(9.3)
	50	48.7(7.2)	42.3(5.8)	49.9(7.3)	49.7(8.0)	45.4(7.3)	85.3(15.8)
	100	27.9(5.4)	27.7(8.7)	28.7(5.3)	28.8(5.5)	25.9(5.3)	82.4(16.3)
	200	-	-	-	-	-	81.7(16.9)
	300	-	-	-	-	-	81.2(16.3)
(1.7)	10	96.8(1.2)	48.1(1.1)	95.7(1.7)	96.5(1.3)	96.6(1.3)	96.4(2.8)
	50	80.4(3.3)	38.0(2.5)	75.9(3.7)	79.1(3.6)	78.7(3.8)	95.0(5.1)
	100	55.9(5.2)	26.1(3.3)	50.8(4.8)	54.3(5.1)	50.9(6.0)	93.6(7.5)
	200	-	-	-	-	-	87.0(16.5)
	300	-	-	-	-	-	80.0(21.4)

containing the remaining 32 observations. We estimate $\mathbf{\Theta}$ using the training set. Figure 3 displays a scree plot of the eigenvalues of $\widehat{\mathbf{\Theta}}_s$. The figure clearly shows there is an obvious nonzero eigenvalue. Figure 3 also presents scatter plots for the response variable and the first linear combination on the test data. The first linear combination exhibits a clear monotone trend. We further conduct a distance correlation t-test (Székely, Rizzo and Bakirov (2007)) between each of the first



(a) Eigenvalues

(b) The first linear combination.

(c) Residual plot

Figure 3. (A): Scree plot of the principal eigenvalues; (B): Scatter plot of the response on the vertical axis and the first linear combination ($\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}_1$) on the horizontal axis on the test data set; (C) Scatter plot of the residuals of the nonparametric kernel regression using the first linear combination.

Table 2. Simulation results for the breast cancer data: The averages (standard deviations) of the distance correlations and mean squared errors based on the test data.

	CUME	SIR_2	SIR_5	SIR_{10}	SIR_{20}	NEW
MSE	1.68(0.55)	1.31(0.34)	1.69(0.60)	1.86(0.70)	2.61(1.31)	0.78(0.22)
DC	0.37(0.10)	0.39(0.10)	0.34(0.09)	0.33(0.08)	0.33(0.08)	0.57(0.10)

two linear combinations and the response variable. The p-values are 0.005 and 0.275, respectively. These results suggest that the first linear combination may be sufficient to predict the tumor size. Therefore, it is reasonable to infer that the central subspace $\mathcal{S}_{Y|\mathbf{x}}$ may be one dimensional.

Next, we examine the performance of the first linear combination in terms of predicting the tumor size. We randomly partition the whole data set into a training and a test data set. We repeat this partition procedure 1,000 times. We estimate $\mathcal{S}_{Y|\mathbf{x}}$ using the different training sets, and calculate the distance correlation (Székely, Rizzo and Bakirov (2007)) between the first linear combination and the response based on the test data. We also predict the tumor size based on the test set using a nonparametric kernel regression. We evaluate the prediction performance using the mean squared errors. The averages (the standard deviations) of the distance correlations and mean squared errors are reported in Table 2, based on 1,000 random partitions. The prediction performance of the SIR varies with the number of slices. Table 2 shows that, in terms of both criteria, our sparse estimate for the CUME method is superior to the SIR and the classical CUME method.

4. Conclusion

We have shown that the classical CUME method is consistent if and only if $p = o(n)$. This is the largest possible p we can handle without a sparsity assumption. When p is greater than n, we introduce a sparse estimate for the CUME matrix, showing that the estimate is consistent as long as $\log(p) = o(n)$. The sparse estimates involve two tuning parameters, λ_{1n} and λ_{2n}. Here, we suggest selecting the optimal λ_{1n} first, and then using this to select the optimal λ_{2n}. Alternatively, we can choose the two simultaneously when the computation is not complex. Several other issues deserve further investigation. For example, for the CUME method to be consistent, we implicitly assume the linearity condition. This assumption is violated if some covariates are categorical or discrete. However, relaxing the linearity assumption when $\mathbf{x}$ is ultrahigh-dimensional is not straightforward. In addition, how to decide the dimension of $\mathcal{S}_{Y|\mathbf{x}}$ for ultrahigh-dimensional semiparametric regressions remains unsolved. Based on the asymptotic theory of the proposed CUME method, one may follow Luo and Li (2016) to combine the eigenvalues and the variation of the eigenvectors to determine the order in high- or ultrahigh-dimension. Another interesting extension would be to apply the thresholding idea to the functional data case.

Supplementary Material

The online Supplementary Material provides proofs for Example 1 and Theorems 1−2, together with additional simulations.

Acknowledgments

We thank the Editor, Associate Editor, and two anonymous reviewers for their insightful comments. Wang's research was supported by the Shanghai Sailing Program (16YF1405700) and the National Natural Science Foundation of China (11701367). Yu's research was supported by the National Natural Science Foundation of China (11571111), 111 project (B14019), Program of Shanghai Subject Chief Scientist (14XD1401600), and Shanghai Rising Star Program (16QA1401700). Zhu is the corresponding author, and his research was supported by the National Natural Science Foundation of China (11731011 and 11931014), 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

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *The Annals of Statistics* **36**, 2577–2604.

Bunea, F. and Xiao, L. (2015). On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fPCA. *Bernoulli* **21**, 1200–1230.

Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106**, 594–607.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. John Wiley & Sons.

Cook, R. D. and Weisberg, S. (1991). Discussion of 'sliced inverse regression for dimension reduction'. *Journal of the American Statistical Association* **86**, 328–332.

Fan, J., Han, F. and Liu, H. (2014). Challenges of big data analysis. *National Science Review* **1**, 293–314.

Fan, J., Liao, Y. and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal* **19**, 1–32.

Ferré, L. (1998). Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association* **93**, 132–140.

Hsing, T. and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics* **20**, 1040–1061.

Jiang, B. and Liu, J. S. (2014). Variable selection for general index models via sliced inverse regression. *The Annals of Statistics* **42**, 1751–1786.

Koltchinskii, V. and Lounici, K. (2017). Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli* **23**, 110–133.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**, 997–1008.

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

Lin, Q., Zhao, Z. and Liu, J. S. (2018a). Sparse sliced inverse regression via Lasso. *Journal of the American Statistical Association* **114**, 1726–1739.

Lin, Q., Zhao, Z. and Liu, J. S. (2018b). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics* **46**, 580–610.

Luo, W. and Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika* **103**, 875–887.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* **34**, 1436–1462.

Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. *Annals of Applied Statistics* **3**, 1236–1265.

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.

Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M. J., Witteveen, A. T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530.

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, 210–268. Cambridge Univ. Press, Cambridge.

Yu, Y., Wang, T. and Samworth, R. (2015). A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika* **102**, 315–323.

Zhang, T. and Zou, H. (2014). Sparse precision matrix estimation via Lasso penalized D-trace loss. *Biometrika* **101**, 103–120.

Zhong, W., Zhang, T., Zhu, Y. and Liu, J. S. (2012). Correlation pursuit: forward stepwise variable selection for index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**, 849–870.

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

Zhu, L.-X., Miao, B. and Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association* **101**, 630–643.

Zhu, L.-X. and Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica* **5**, 727–736.

Cheng Wang

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

E-mail: chengwang@sjtu.edu.cn

Zhou Yu

Key Laboratory of Advanced Theory and Application in Statistics and Data Science, Ministry of Education and School of Statistics, East China Normal University, Shanghai 200241, China.

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

Liping Zhu

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

(Received September 2018; accepted February 2019)