On Sliced Inverse Regression With High-Dimensional Covariates

Lixing ZHU, Baiqi MIAO, and Heng PENG

Sliced inverse regression is a promising method for the estimation of the central dimension-reduction subspace (CDR space) in semiparametric regression models. It is particularly useful in tackling cases with high-dimensional covariates. In this article we study the asymptotic behavior of the estimate of the CDR space with high-dimensional covariates, that is, when the dimension of the covariates goes to infinity as the sample size goes to infinity. Strong and weak convergence are obtained. We also suggest an estimation procedure of the Bayes information criterion type to ascertain the dimension of the CDR space and derive the consistency. A simulation study is conducted.

KEY WORDS: Central dimension-reduction subspace; Convergence rate; Dimensionality determination; Sliced inverse regression.

1. INTRODUCTION

Large-scale data analysis has recently received much attention. Reducing the dimensions of data without loss of information is a natural manner in which to proceed with such analysis. In this area, Li (1991) considered a semiparametric regression model in which the response variable y depends on the covariates $\mathbf{x} = (x_1, \ldots, x_p)^T$ through K linear combinations of the components, x_i's. His model can be generalized to a conditional independence statement. Let $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ be K orthogonal $p \times 1$ vectors, the norms of which are $\|\boldsymbol{\beta}_i\| = \sqrt{\beta_{i1}^2 + \cdots + \beta_{ip}^2} = 1$ for $i = 1, \ldots, K$, where β_{ij}, $j = 1, \ldots, p$, are the components of $\boldsymbol{\beta}_i$. When $(\boldsymbol{\beta}_1^T\mathbf{x}, \ldots, \boldsymbol{\beta}_K^T\mathbf{x})$ is given, y is independent of $\mathbf{x}$, that is,

$$y \perp\!\!\!\perp \mathbf{x} | (\boldsymbol{\beta}_1^T\mathbf{x}, \ldots, \boldsymbol{\beta}_K^T\mathbf{x}), \tag{1}$$

where "$\perp\!\!\!\perp$" represents independence.

A dimension-reduction subspace (Cook 1994, 1998) is defined as the column space of any $p \times K$ ($K \le p$) matrix $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ such that (1) holds. For uniqueness, we are interested in a subspace with minimal dimensions. Under mild conditions, the minimal subspace is often uniquely defined in practice and coincides with the intersection of all subspaces that satisfy (1) (Cook 1994, 1996). This intersection is called the *central dimension-reduction* (CDR) *space* and is written as $\mathcal{S}_{y|\mathbf{x}}$. In this article we assume that $\mathcal{S}_{y|\mathbf{x}}$ exists.

Sliced inverse regression (SIR) (Li 1991) can be used to estimate $S_{y|\mathbf{x}}$. This method is popular in the literature. Li (1991) derived the asymptotic results under the assumption that $\mathbf{x}$ is normally distributed, Hsing and Carroll (1992) proved the asymptotic normality of the slicing estimator when each slice contains two points, Zhu and Ng (1995) obtained the asymptotic normality for general cases, and Zhu and Fang (1996) studied the asymptotic behavior of a kernel estimator. Other important estimation approaches have also been treated in the literature, including sliced average variance estimates (SAVE) (Cook and Weisberg 1991; Cook 2000), parametric inverse regression (PIR) (Bura and Cook 2001a,b), and the proposed hybrid methods of Zhu, Ohtaki, and Li (2006) that are convex combinations of SIR and SAVE. Another relevant work is that of Cook and Li (2002).

All of the existing studies of asymptotics for model (1) are related to cases with a fixed dimension p and a sample size, say n, that is comparably high. However, it is often the case that the number of covariates is also large. Li (1991) first raised this issue and suggested the importance of studying it further. Recent research with covariates of a very high dimension also demonstrates this necessity. These kinds of datasets include gene expression data, DNA microarray data, and consumer financial history data. Golub et al. (1999) brought to life a whole new branch of data analysis under the name of microarray analysis. Data with high-dimensional covariates also appear in other areas. Zhu, Zhu, and Li (2006) analyzed a dataset of agricultural meteorological disasters. They explored the relationship between the yields of three crops and meteorological conditions. The dimension p of the covariate was also comparably high. Greenshtain and Ritov (2004) studied the asymptotics of lasso methods when p goes to infinity as n tends to infinity. Bickel and Levina (2004) investigated the asymptotic properties for Fisher's linear discriminant analysis (for relevant data analysis see Levina 2002). Conventional methods are limited for this kind of data analysis, and dimension-reduction methods may be helpful. Among others, Antoniadis, Lambert-Lacroix, and Leblanc (2003) applied a relevant method, MAVE, proposed by Xia, Tong, Li, and Zhu (2002), to analyze gene expression data for classification, and Bura and Pfeiffer (2003) used SIR and SAVE for DNA microarray data. Related work was done by Chiaromonte and Martinelli (2002).

Furthermore, to apply SIR to estimate $\mathcal{S}_{y|\mathbf{x}}$, the linearity condition as designed by Li (1991) must be checked. Note that the linearity condition cannot be always met in cases in which the dimensions are small or moderate. Hall and Li (1993) proved that when $p \to \infty$ as $n \to \infty$, the linear combinations of the covariates are approximately normally distributed. This ensures that the linearity condition for SIR is asymptotically satisfied. Therefore, for cases in which the dimensions are high, SIR can work well. This stimulates us to study the asymptotic behavior of SIR with high-dimensional covariates.

Lixing Zhu is Professor of Statistics, Department of Mathematics, Hong Kong Baptist University, and Cheung Kong Chair Professor at Renmin University of China under the Cheung Kong Scholars Program of the Ministry of Education and Li Ka Shing Foundation, Hong Kong, People's Republic of China (E-mail: *lzhu@hkbu.edu.hk*). Baiqi Miao is Professor, University of Science and Technology of China. Heng Peng is Post-Doctoral Fellow, Princeton University, Princeton, NJ 08544. This research was supported by a grant from the University Grants Council of Hong Kong (HKU7181/02H). The authors are grateful to the editor, an associate editor, and the two referees for their constructive comments and suggestions, which led to the great improvement of earlier draft. They also thank the associate editor and one of the referees for their generous help in improving the presentation of the article.

Journal of the American Statistical Association
June 2006, Vol. 101, No. 474, Theory and Methods
DOI 10.1198/016214505000001285

The asymptotic research for the $p \to \infty$ case is very challenging, however. The literature contains some relevant studies on covariance matrices. It is impossible to uniformly formulate the asymptotic results for different ratios $\lambda = p/n$ when p diverges to infinity as n tends to infinity (see Bai and Saranadasa 1996; Bai 1999 for details). This demonstrates the significant difference in convergence between the case with a fixed p and that with a varying p.

The estimate for a matrix based on SIR has a much more complicated structure than the sample covariance matrix. In this article we explore the asymptotic structure of SIR matrix estimate, then investigate the strong and weak convergence when $p \to \infty$ as $n \to \infty$.

Another important issue is determinating the dimension of $\mathcal{S}_{y|\mathbf{x}}$. For SIR, Li (1991) suggested a sequential chi-squared test procedure to determine the dimensions. Schott (1994) extended this procedure to the situation in which the vector of explanatory variables is sampled from an elliptically symmetric distribution. Velilla (1998) also considered a sequential test. Bura and Cook (2001a) suggested a general weighted chi-squared sequential test that does not require the assumption that covariates follow a normal distribution. It should be noted that the significance level at each step in a sequential test procedure does not determine the significance level of the entire procedure, and that the retained dimension depends on the choice of significance level. Taking this into account, Ferré (1998) proposed an approach to determine the dimension of $\mathcal{S}_{y|\mathbf{x}}$. Although his method is somewhat similar to that of Li (1991), it does not need a sequential test, and it selects the dimension automatically.

We suggest a new procedure of the Bayes information criterion (BIC) (Schwarz 1978) type for determinating dimensions. Consistency of the estimator of the dimensions is obtained.

To the best of our knowledge, this is the first article to address the problem of high-dimensional covariates in the context of SIR, and it also presents a useful new methodology for dimension estimation. For example, the results can be used to deal with the problem of agricultural meteorological disasters, as studied by Zhu, Ohtaki, and Li (2006).

The article is organized as follows. In Section 2 we give a brief description of SIR. We provide the consistency results when $p \to \infty$ as $n \to \infty$ in Section 3. In Section 4 we present our BIC-type procedure for determining the dimension of $\mathcal{S}_{y|\mathbf{x}}$ and the consistency result for this procedure. We report simulations in Section 5 and give technical proofs of the theorems in the Appendix.

2. A BRIEF DESCRIPTION OF SLICED INVERSE REGRESSION

Denote by $\mathbf{P}_{(\cdot)}$ the projection operator in the standard inner product (see Cook 1998). When $\mathbf{x}$ is standardized, assume that the linearity condition is satisfied, that is,

$$E\big(\mathbf{x}\big|\mathbf{P}_{\mathcal{S}_{y|\mathbf{x}}}\mathbf{x}\big) = \mathbf{P}_{\mathcal{S}_{y|\mathbf{x}}}\mathbf{x}. \tag{2}$$

Then, under (1), the centered "inverse regression," $E(\mathbf{x}|y) - E(\mathbf{x})$, is confined to $\mathcal{S}_{y|\mathbf{x}}$. This means that

$$\mathcal{S}_{E(\mathbf{x}|y)} \subset \mathcal{S}_{y|\mathbf{x}},$$

where $\mathcal{S}_{E(\mathbf{x}|y)}$ represents the space spanned by the centered "inverse regression" $E(\mathbf{x}|y) - E(\mathbf{x})$. As Chiaromonte, Cook, and Li (2002) discussed, if we assume that

$$y \perp\!\!\!\perp \mathbf{x}|\mathbf{P}_{\mathcal{S}_{E(\mathbf{x}|y)}}\mathbf{x}, \tag{3}$$

and if (2) holds, then $\mathcal{S}_{E(\mathbf{x}|y)} = \mathcal{S}_{y|\mathbf{x}}$. In other words, an estimate of $\mathcal{S}_{E(\mathbf{x}|y)}$ would provide a good approximation to $\mathcal{S}_{y|\mathbf{x}}$. If condition (3) is not satisfied, then, following the suggestion of Li (1991) and Cook and Weisberg (1991), we can explore higher conditional moments of $\mathbf{x}$ given y or consider the so-called "central kth-moment subspace" (CKMS) (Yin and Cook 2002). We assume that $\mathcal{S}_{E(\mathbf{x}|y)} = \mathcal{S}_{y|\mathbf{x}}$ and restrict our attention to SIR.

In the general case, $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\mathcal{S}_{E(\mathbf{x}|y)} = \mathcal{S}_{y|\mathbf{x}}$. Note that $\{\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\eta}_1, \ldots, \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\eta}_K\}$ is a basis of $S_{y|\mathbf{x}}$, where $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_K$ are the eigenvectors associated with the K largest eigenvalues of the matrix $\operatorname{cov}\{E(\mathbf{x}|y)\}$. Therefore, to estimate $S_{y|\mathbf{x}}$, we need only construct an estimate of $\operatorname{cov}\{E(\mathbf{x}|y)\}$. In contrast, it is well known that

$$\boldsymbol{\Sigma}_{\mathbf{x}} = \operatorname{cov}\{E(\mathbf{x}|y)\} + E\{\operatorname{cov}(\mathbf{x}|y)\} =: \operatorname{cov}\{E(\mathbf{x}|y)\} + \boldsymbol{\Lambda}_p.$$

Thus, by estimating $\boldsymbol{\Lambda}_p$ and $\boldsymbol{\Sigma}_{\mathbf{x}}$, we can also obtain an estimate of $\operatorname{cov}\{E(\mathbf{x}|y)\}$, and then the estimate of $\mathcal{S}_{y|\mathbf{x}}$.

Alternatively, we can derive the estimate of $S_{y|\mathbf{x}}$ using the standardized variable $\mathbf{x}, \mathbf{z} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}\{\mathbf{x} - E(\mathbf{x})\}$. As is known, $\mathcal{S}_{y|\mathbf{z}} = \boldsymbol{\Sigma}_{\mathbf{x}}^{1/2}\mathcal{S}_{y|\mathbf{x}}$ (see Li 1991; Cook 1998, chaps. 10 and 11). Hence $\{\boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}\boldsymbol{\eta}_1', \ldots, \boldsymbol{\Sigma}_{\mathbf{x}}^{-1/2}\boldsymbol{\eta}_K'\}$ is also a basis of $S_{y|\mathbf{x}}$, where $\boldsymbol{\eta}_1', \ldots, \boldsymbol{\eta}_K'$ are the eigenvectors associated with the K largest eigenvalues of matrix $\operatorname{cov}\{E(\mathbf{z}|y)\}$.

Li (1991) suggested two ways to estimate the basis of $S_{y|\mathbf{x}}$. Once we have an estimate $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}$ of $\boldsymbol{\Sigma}_{\mathbf{x}}$, we can either estimate $\operatorname{cov}\{E(\mathbf{x}|y)\}$ directly or estimate $\boldsymbol{\Lambda}_p$ first and then estimate $\operatorname{cov}\{E(\mathbf{x}|y)\}$ by $\boldsymbol{\Sigma}_{\mathbf{x}} - \boldsymbol{\Lambda}_p$. The same can be applied to $\mathbf{z}$. The following estimate is for $\boldsymbol{\Lambda}_p$.

Let $\{(y_i, \mathbf{x}_i), i = 1, \ldots, n\}$ be an iid sample and, according to the value of y_i, denote the order statistics by $(y_{(i)}, \mathbf{x}_{(i)}), i = 1, \ldots, n$, where $y_{(1)} \le \cdots \le y_{(n)}$ and $\mathbf{x}_{(i)}$ are called the concomitant of $y_{(i)}$ (see Yang 1977). We introduce a double subscript (h, j) in which the first element refers to the slice number and the second number refers to the order number of an observation in the given slice, that is,

$$y_{(h,j)} = y_{(c(h-1)+j)}, \qquad \mathbf{x}_{(h,j)} = \mathbf{x}_{(c(h-1)+j)},$$

where $c > 0$ is the number of $y_{(i)}$ in every slice. The estimate $\widehat{\boldsymbol{\Lambda}}_p$ of $\boldsymbol{\Lambda}_p$ has the form

$$\widehat{\boldsymbol{\Lambda}}_p = \frac{1}{H}\sum_{h=1}^{H}\left\{\frac{1}{c-1}\sum_{j=1}^{c}\left(\mathbf{x}_{(h,j)} - \frac{1}{c}\sum_{\ell=1}^{c}\mathbf{x}_{(h,\ell)}\right) \times \left(\mathbf{x}_{(h,j)} - \frac{1}{c}\sum_{\ell=1}^{c}\mathbf{x}_{(h,\ell)}\right)^T\right\}, \tag{4}$$

where $H = [(n + c - 1)/c]$ is the number of slices where $[a]$ is the largest integer part of a. In practice, the number of $y_{(i)}$ in the last slice may be less than c, but this has little effect on SIR, because the number of slices H is very large.

When the dimension p of $\mathbf{x}$ is fixed, Hsing and Carroll (1992) proved the root-n consistency of the estimate $\widehat{\boldsymbol{\Lambda}}_p$ for the case in which each slice contains two of the ordered $y_{(i)}$, that is, $c = 2$.

Zhu and Ng (1995) established the root-n consistency for the case where c is an arbitrarily fixed number and the case where $c \to \infty$ as $n \to \infty$. We study the asymptotic properties of $\widehat{\mathbf{\Lambda}}_p$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{x}} - \widehat{\mathbf{\Lambda}}_p$.

3. ASYMPTOTIC PROPERTIES OF SLICED INVERSE REGRESSION WITH LARGE p

Let $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, be an independent and identically distributed (iid) sample. Denote the inverse regression curve and the associated residual by $\mathbf{m}(y) = E(\mathbf{x}|y)$ and

$$\varepsilon_i = \mathbf{x}_i - \mathbf{m}(y_i) \qquad \text{and} \qquad \varepsilon_{(i)} = \mathbf{x}_{(i)} - \mathbf{m}\big(y_{(i)}\big), \quad i = 1, \ldots, n. \tag{5}$$

Here $\mathbf{m}(y)$ is a p-dimensional vector with components $m_i(y) = E(x_i|y)$, $i = 1, \ldots, p$. Note that the $\varepsilon_{(i)}$'s are also concomitants of the $y_{(i)}$. It is clear that the ε_i's are iid and that (see Yang 1977) they are independent with mean 0 when the order statistics, $y_{(i)}$'s, are given. Because $p \to \infty$ as $n \to \infty$, all $p, \mathbf{x}_i, \mathbf{\Sigma}_{\mathbf{x}}, y_{(i)}, \varepsilon_{(i)}$, and $\mathbf{m}(y_{(i)})$ depend on n and should be denoted by $p_n, x_i^{(n)}, \mathbf{\Sigma}_{\mathbf{x}}^{(n)}, y_{(i)}^{(n)}, \varepsilon_{(i)}^{(n)}$, and $\mathbf{m}_n(y_{(i)}^{(n)})$. For notational simplicity, we omit "n" from the subscripts and superscripts unless stated otherwise.

To study the asymptotic behavior, we need smoothness conditions on the inverse regression curve $\mathbf{m}(y)$. Let $\Pi_n(B)$ be the collection of all n-point partitions, $-B \le y_{(1)}^* \le \cdots \le y_{(n)}^* \le B$, of the interval $[-B, B]$, where $B > 0$ and $n \ge 1$. Any vector-valued or real-valued function, $\mathbf{m}(y)$, is said to have a *total variation of order r* if, for any fixed $B > 0$,

$$\lim_{n\to\infty} \frac{1}{n^r \sqrt{p}} \sup_{\Pi_n(B)} \sum_{i=1}^{n-1} \big\|\mathbf{m}\big(y_{(i+1)}^*\big) - \mathbf{m}\big(y_{(i)}^*\big)\big\| = 0, \tag{6}$$

where $\|\cdot\|$ is the Euclidean norm—namely, $\|\mathbf{A}\|$ is the root of the squared sum of all elements of $\mathbf{A}$. A similar definition of total variation has been given by Hsing and Carroll (1992) and Zhu and Ng (1995), except for the value $\sqrt{p}$ in the denominator. We have this value because in the Euclidean norm of $\mathbf{m}(y_{(i+1)}^*) - \mathbf{m}(y_{(i)}^*)$ there are p terms to be summed. We use $\sqrt{p}$ to average these terms. There are many functions of total variation; a special case is the function that satisfies the Lipschitz condition. This can be easily seen because for some $C > 0$, $\|\mathbf{m}(y_{(i+1)}^*) - \mathbf{m}(y_{(i)}^*)\| \le C\sqrt{p}\|y_{(i+1)}^* - y_{(i)}^*\|$ and $(1/\sqrt{p}) \sum_{i=1}^n \|\mathbf{m}(y_{(i+1)}^*) - \mathbf{m}(y_{(i)}^*)\| \le C\|y_{(n)}^* - y_{(1)}^*\| \le 2CB$. Thus, such a function is of a total variation of any order $r > 0$.

Furthermore, for the components $m_i(\cdot)$ of $\mathbf{m}(\cdot)$, if there is a nondecreasing real-valued function M and a real number B_0 such that for any two points, say y_1 and y_2, both in $(-\infty, -B_0]$ or $[B_0, \infty)$,

$$|m_i(y_1) - m_i(y_2)| \le |M(y_1) - M(y_2)|, \qquad i = 1, \ldots, p, \tag{7}$$

then we say that the function $\mathbf{m}(y)$ is *nonexpansive in the metric of M_{B_0} on both sides of B_0*.

The following theorem states the strong and weak convergence of $\|\widehat{\mathbf{\Lambda}}_p - \mathbf{\Lambda}_p\|$.

Theorem 1. Assume that the following conditions are satisfied:

(A) $\sup_{i\le p} E(|x_i|^l) < C_1$ for some constants $l \ge 4$ and C_1, where $\mathbf{x} = (x_1, \ldots, x_p)^T$.

(B) The inverse regression function $\mathbf{m}(y)$ has a total variation of order r, $0 < r < 1/4$. When p converges to infinity, y_i are dependent on n, and we assume the following condition.

(C) Let $F_n(\cdot)$ be the distribution from which the sample y_i's are drawn. Let Y be the random variable following the distribution $F_n(\cdot)$. Suppose that there are two random variables, W_1 and W_2, with distributions G_1 and G_2 such that $G_1(x) \le F_n(x) \le G_2(x)$; that is, $W_2 \prec Y \prec W_1$, where "$\prec$" represents "stochastically less than." Similarly "$\succ$" represents " stochastically greater than."

(D) For any B_0, $\mathbf{m}(y)$ is nonexpansive in the metric of $M_{B_0}(y)$ on both sides of a positive number B_0, such that $l \times r > 2$ for the l of condition (A),

$$E\big(|M(W_1)|^l\big) < \infty \qquad \text{and} \qquad E\big(|M(W_2)|^l\big) < \infty.$$

Then we have for any fixed $c \ge 2$, when $p = o(n^{\min\{1/2, 1-2/l-2r\}})$ as $n \to \infty$,

$$\|\widehat{\mathbf{\Lambda}}_p - \mathbf{\Lambda}_p\| = O_p\big(pn^{-1/2}\big)$$

and

$$\|(\widehat{\mathbf{\Sigma}}_{\mathbf{x}} - \widehat{\mathbf{\Lambda}}_p) - (\mathbf{\Sigma}_{\mathbf{x}} - \mathbf{\Lambda}_p)\| = O_p\big(pn^{-1/2}\big).$$

When $p = o(n^{1/4}/(\log n))$, the foregoing two terms almost surely have the convergence rate $o(pn^{-1/4}(\log n))$.

Remark 1. Condition (A) is almost necessary for obtaining the rate of weak convergence with root n. When p is fixed, the existence of a fourth moment is a necessary condition (see Petrov 1995). Condition (B) assumes the smoothness of the inverse regression $\mathbf{m}(y)$. It is slightly stronger than that of Zhu and Ng (1995) when p is fixed; therefore, it is quite mild. As for conditions (C) and (D), they are special for our problem. In the case where p is fixed and the distribution F_n of y_i is also independent of n, conditions (C) and (D) are reduced to those of Zhu and Ng (1995). Hence the conditions are needed to uniformly bind the distribution series $F_n(\cdot)$.

Remark 2. For high-dimension p, it is a natural concern whether $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}$ will be ill-conditioned, because the inverse matrix $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}$ is needed when standardized variable $\mathbf{z}$ is used to construct the estimate of the CDR space. It is noteworthy that for a fixed p, when n is large, $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}$ cannot be ill-conditioned if $\mathbf{\Sigma}_{\mathbf{x}}$ is of full rank. When p is large, $\mathbf{x}$ is normally distributed, and $\mathbf{\Sigma}_{\mathbf{x}}$ is of full rank, the smallest and largest eigenvalues of $\mathbf{\Sigma}_{\mathbf{x}}^{-1/2}\widehat{\mathbf{\Sigma}}_{\mathbf{x}}\mathbf{\Sigma}_{\mathbf{x}}^{-1/2} - \mathbf{I}_p$ converge to 0 almost surely as $n \to \infty$ and $p/n \to 0$. That is, for large p, once n is large enough, $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{x}}^{-1}$ are also of full rank. Thus the ill-conditioning problem can also be avoided (see Bai 1999). In our case, $p = o(n^{\min\{1/2, 1-2/l-2r\}})$ satisfies this requirement.

Remark 3. In the cases where all moments of $\mathbf{x}$ exist and the total variation of order r is small, $p = o(n^{\min\{1/2, 1-2/l-2r\}})$ can be close to $O(n^{1/2})$. It is noteworthy that $p = O(n^{1/2})$ is the fastest possible rate if we do not assume extra conditions. This is because each element of $\widehat{\mathbf{\Lambda}}_p - \mathbf{\Lambda}_p$ has an optimal rate $n^{-1/2}$ and $\|\widehat{\mathbf{\Lambda}}_p - \mathbf{\Lambda}_p\|$ is the root of the squared sum of p^2 elements. The resulting rate of $\|\widehat{\mathbf{\Lambda}}_p - \mathbf{\Lambda}_p\|$ must have a factor p. But some data, such as DNA microarray data, have even higher dimension than the sample size. In this case the asymptotic behavior of the estimate is very difficult to study unless $\widehat{\mathbf{\Lambda}}_p - \mathbf{\Lambda}_p$ has a special structure. This is an important issue that merits further study.

Remark 4. The choice of the number, c, of data points in each slice is of practical importance. When p is fixed, Zhu and Ng (1995) derived that for a wide range of c, the convergence of the estimate can be achieved. When p varies with n, we have not obtained a similar result for the case with $c \to \infty$ and $c/n \to 0$ as $n \to \infty$ due to technical difficulties. However, we guess that if the convergence rate of c is much slower than that of p, then the consistency may still hold. We will consider this problem in a future study. When c is proportional to n (i.e., the number of slices is fixed), the CDR space can still be estimated, because the space that is determined by an approximation of $\boldsymbol{\Sigma}_{\mathbf{x}} - \boldsymbol{\Lambda}_p$ is a subspace contained in the CDR space although the estimator is not consistent with the matrix $\boldsymbol{\Sigma}_{\mathbf{x}} - \boldsymbol{\Lambda}_p$. In this case the asymptotics are much easier to study. The basic idea is as follows. In each slice there are n/H data points where H is a fixed number. Hence we can immediately use central limit theorems to a H-dimensional vector, the components of which are the sample covariance matrices based on the data in each slice. This is because each sample covariance matrix can be rewritten asymptotically as a sum of iid random variables; see the relevant work of Li and Zhu (2004).

Remark 5. By the known result of strong convergence about the sum of independent random variables (see Petrov 1995, thm. 9.20, p. 278), we can easily derive that almost surely $\|\boldsymbol{\Sigma}_{\mathbf{x}} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}\| = O(pn^{-1/2}\log n)$. The proof is given in the Appendix. Similar to Corollary 1, we can also derive the convergence of the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}$. Let $\{\gamma_i\}$ and $\{\widehat{\gamma}_i\}$ be the eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{x}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}$. Under the condition that $\lim_{p\to\infty} \min_{1\le i\le p}\{\gamma_i\} > b > 0$, we also have that when n is large, $\lim_{p\to\infty} \min_{1\le i\le p}\{\widehat{\gamma}_i\} > b/2 > 0$. Note that

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{x}}^{-1} = \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}(\boldsymbol{\Sigma}_{\mathbf{x}} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}})\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1}.$$

By Lemma A.2 in the Appendix, the symmetry of both $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}$ and $\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}$ and Corollary 1, it is not difficult to show that

$$\begin{aligned}\|\boldsymbol{\Sigma}_{\mathbf{x}}^{-1} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1}\| &\le \frac{1}{(\min_{1\le i\le p}\{\gamma_i\}\min_{1\le i\le p}\{\widehat{\gamma}_i\})}\|\boldsymbol{\Sigma}_{\mathbf{x}} - \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}\| \\ &= O\big(pn^{-1/2}\log n\big) \quad \text{a.s.}\end{aligned}$$

Together with Theorem 1 and Corollary 1, we can obtain the same convergence rates as those of Theorem 1 for the estimate of the basis, $\{\boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\eta}_1, \ldots, \boldsymbol{\Sigma}_{\mathbf{x}}^{-1}\boldsymbol{\eta}_K\}$. Hence, when we use the standardized variable $\mathbf{z} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1/2}(\mathbf{x} - E(\mathbf{x}))$, the asymptotic results of Theorem 1 and Corollary 1 are also similar.

Consider the convergence of the eigenvectors associated with the largest K eigenvalues of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}} - \widehat{\boldsymbol{\Lambda}}_p$ and the projection spaces of these eigenvectors. Let $\lambda_{p1} \ge \cdots \ge \lambda_{pp}$ denote the eigenvalues of $\boldsymbol{\Sigma}_{\mathbf{x}} - \boldsymbol{\Lambda}_p$ and $\widehat{\lambda}_{p1} \ge \cdots \ge \widehat{\lambda}_{pp}$ denote the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}} - \widehat{\boldsymbol{\Lambda}}_p$. Furthermore, let $P_{\lambda_{pi}}$ and $P_{\widehat{\lambda}_{pi}}$ be the projection spaces associated with λ_{pi} and $\widehat{\lambda}_{pi}$. Note that if an eigenvalue is distinct from others, then the corresponding projection space is the space spanned by the associated eigenvector. Theorem 1 implies the following corollary.

Corollary 1. Assume that the conditions of Theorem 1 hold. For each $i = 1, \ldots, K$, $|\lambda_{pi} - \widehat{\lambda}_{pi}|$ converges in probability to 0 at the same rates as in Theorem 1. In addition, assume that $\lambda_{p1} \ge \cdots \ge \lambda_{pK}$ are distinct in the sense that

$$\liminf_{n\to\infty} |\lambda_{p(i-1)} - \lambda_{pi}| > 0, \qquad i = 2, \ldots, K.$$

Then, for $i = 1, \ldots, K$, $\|P_{\lambda_{pi}} - P_{\widehat{\lambda}_{pi}}\|$ also converge in probability to 0 at the same rate as in Theorem 1.

4. DETERMINATION OF THE DIMENSION OF $\mathcal{S}_{y|\mathbf{x}}$

Differing from the methods in the literature, we suggest a procedure of the BIC type for determining the dimension. It can be used to handle the problems with high-dimensional covariates. The procedure is easy to implement and the consistency of the estimate is established.

Zhao, Krishnaiah, and Bai (1986a,b) studied a problem of detecting the number of signals from noise. By analyzing the eigenvalues of the covariance matrix of the sample, they proposed a BIC-type model selection procedure to determine the number of signals and proved the consistency of the estimate. There is some similarity between their problem and the determination of the dimension of $\mathcal{S}_{y|\mathbf{x}}$. We now borrow their idea to construct a determination criterion.

Let $\boldsymbol{\Psi} = \boldsymbol{\Sigma}_{\mathbf{x}} - \boldsymbol{\Lambda}_p = \operatorname{cov}\{E(\mathbf{x}|y)\}$ and $\widehat{\boldsymbol{\Psi}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}} - \widehat{\boldsymbol{\Lambda}}_p$. Because the smallest $p - K$ eigenvalues of $\boldsymbol{\Psi}$ are 0, we can consider the largest K eigenvalues the signals and the value of K the number of signals. To apply the method used by Zhao et al. (1986a,b), we artificially add an identity matrix $\mathbf{I}_p$, which acts as a covariance matrix of white noise to $\boldsymbol{\Psi}$ and $\widehat{\boldsymbol{\Psi}}$.

Let $\boldsymbol{\Omega} = \boldsymbol{\Psi} + \mathbf{I}_p$ and $\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Psi}} + \mathbf{I}_p$. Let $\theta_1 \ge \theta_2 \ge \cdots \ge \theta_p$ be the eigenvalues of $\boldsymbol{\Omega}$ and $\widehat{\theta}_1 \ge \widehat{\theta}_2 \ge \cdots \ge \widehat{\theta}_p$ be the eigenvalues of $\widehat{\boldsymbol{\Omega}}$. It is clear that $\theta_i = \lambda_i + 1$, where λ_i are the eigenvalues of $\boldsymbol{\Psi}$. Determining the dimension of $S_{y|\mathbf{x}}$ now becomes estimating K, the number of the eigenvalues of $\boldsymbol{\Omega}$ >1.

Borrowing the idea of maximum likelihood estimation in the normal distribution case, we define

$$\log L(\boldsymbol{\theta}) = -\frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{n}{2}\operatorname{tr}\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\Omega}}, \tag{8}$$

because it is a function of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Let Θ_k be the set consisting of all values such that $\theta_1 \ge \theta_2 \ge \cdots \ge \theta_k > 1$ and $\theta_{k+1} = \cdots = \theta_p = 1$. In addition, let τ denote the number of $\widehat{\theta}_i$'s that are >1. According to Zhao et al. (1986a,b), we can have a specific form of $\sup_{\boldsymbol{\theta}\in\Theta_k} \log L(\boldsymbol{\theta})$,

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta_k} \log L(\boldsymbol{\theta}) = -\frac{n}{2}\sum_{i=1}^{p}\log\widehat{\theta}_i - \frac{np}{2} \\ + \frac{n}{2}\sum_{i=1+\min(\tau,k)}^{p}(\log\widehat{\theta}_i + 1 - \widehat{\theta}_i).\end{aligned} \tag{9}$$

Note that this supreme does not involve the unknowns relating to the matrix $\boldsymbol{\Omega}$ and its eigenvalues, θ_i's. For defining the estimator $\widehat{K}$ of the true dimension K, (9) is equivalent to

$$\frac{n}{2}\sum_{i=1+\min(\tau,k)}^{p}(\log\widehat{\theta}_i + 1 - \widehat{\theta}_i).$$

and thus we can define a procedure based on it. A criterion of BIC type is defined as follows. Let

$$\log L_k = \frac{n}{2}\sum_{i=1+\min(\tau,k)}^{p}(\log\widehat{\theta}_i + 1 - \widehat{\theta}_i) \tag{10}$$

and

$$G(k) = \log L_k - \frac{C_n k(2p-k+1)}{2}, \tag{11}$$

where the second term is the penalty term, C_n is a penalty constant, and $k(2p-k+1)/2$ equals the number of free parameters of (8) needed to estimate when $\boldsymbol{\theta} \in \Theta_k$. It can be justified as follows. When k is fixed, the SIR matrix $\boldsymbol{\Psi}$ is $p \times p$ being of rank k. Hence the last $p-k$ columns of this matrix are the linear combinations of the first k columns. The total number of the parameters in these k columns is kp. Furthermore, because of the symmetry of the matrix, in the left-upper $k \times k$ submatrix $k(k-1)/2$ elements are the same as those above the diagonal. Hence the total number of the free parameters is $kp - k(k-1)/2 = k(2p-k+1)/2$.

The estimator of K is defined as the maximizer $\widehat{K}$ of $G(k)$ over $k \in \{0, \ldots, p-1\}$, that is,

$$G(\widehat{K}) = \max_{0 \le k \le p-1} G(k). \tag{12}$$

The following result states the weak and strong consistency of $\widehat{K}$.

Theorem 2. Assume that $p = O(n^s)$, K is a constant independent of n, $\|\boldsymbol{\Omega} - \widehat{\boldsymbol{\Omega}}\| = O_P(n^{-t})$ [or $O(n^{-t})$ a.s.], $t > 0$, $2t > s$, and C_n satisfies the following:

(a) $\lim_{n\to+\infty} C_n/n^{1-s} = 0$
(b) $\lim_{n\to+\infty} C_n/n^{1-2t} = \infty$.

Then $\widehat{K} - K = o_P(1)$ or $= o(1)$ a.s.

Remark 6. From the proof of Theorem 1, we know that there are s and t satisfying the designed conditions of Theorem 2. If $p = O(n^s)$ and $C_n = O(n^a)$, then, from the results of Theorem 1, we derive that for any small $\eta > 0$,

$$\|\boldsymbol{\Omega} - \widehat{\boldsymbol{\Omega}}\| = O_p\big(n^{2s-1/2}\big) \text{ or } O\big(n^{s+\eta-1/4}\big) \quad \text{a.s.}$$

Thus, from Theorem 2, the power a of C_n can be chosen within the range $1-2t = 4s < a < 1-s$ for the weak convergence and $1/2 + 2s < a < 1-s$ for the strong convergence. If p is fixed, then $s = 0$, and we can select a in the interval $[0, 1)$ and $[1/2, 1)$ for the weak and strong convergence. Therefore, the range is fairly wide.

Remark 7. In our procedure, we artificially add a covariance matrix, $\mathbf{I}_p$, to the matrix $\boldsymbol{\Psi}$. There is an alternative to constructing $\boldsymbol{\Omega}$. Note that $\boldsymbol{\Sigma}_{\mathbf{x}} = \boldsymbol{\Psi} + \boldsymbol{\Lambda}_p$. Then, if $\boldsymbol{\Lambda}_p$ is nonsingular, we can replace the previous $\boldsymbol{\Omega}$ and $\widehat{\boldsymbol{\Omega}}$ by

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Lambda}_p^{-1} = \boldsymbol{\Psi}\boldsymbol{\Lambda}_p^{-1} + \mathbf{I}_p \qquad \text{and}$$

$$\widehat{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}\widehat{\boldsymbol{\Lambda}}_p^{-1} = \widehat{\boldsymbol{\Psi}}\widehat{\boldsymbol{\Lambda}}_p^{-1} + \mathbf{I}_p.$$

Remark 8. Note that the BIC-type method consistently estimates the dimension of the CDR space under the condition that the estimator of SIR matrix is consistent at some rate. The limit distribution is not necessary. Therefore, the method is easy to use and is readily extended to other dimension-reduction methods, such as SAVE and MAVE. In contrast, the popular sequential test method can be easily used when the limit distribution of the estimator of the SIR matrix is available and its limiting variance is easy to estimate; otherwise, we must use a Monte Carlo approximation to simulate its distribution. Permutation (e.g., Cook and Weisberg 1991) or the bootstrap may be applicable if the consistency of these approximations can be verified.

5. SIMULATION STUDY

To investigate the performance of estimation and dimensionality determination by the BIC-type method, we carried out a set of simulations. In this section we report part of the results of these simulations. We used five models. To measure the distance between the CDR space and its estimator, we first obtained the eigenvectors $\{\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_K\}$ associated with the first K largest eigenvalues of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}} - \widehat{\boldsymbol{\Lambda}}_p$, then multiplied these vectors from the left by $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}^{-1}$ to form an estimator $\mathcal{B}$ of $\mathcal{S}_{y|\mathbf{x}}$. As suggested by Li (1991), for any i, we use the squared multiple correlation coefficient $R^2(\widehat{\boldsymbol{\beta}}_i)$ to measure the distance between $\widehat{\boldsymbol{\beta}}_1^T\mathbf{x}$ and $\mathcal{B}$, where

$$R^2(\widehat{\boldsymbol{\beta}}_i) = \max_{\boldsymbol{\beta} \in \mathcal{S}_{y|\mathbf{x}}} \frac{(\widehat{\boldsymbol{\beta}}_i \boldsymbol{\Sigma}_{\mathbf{x}} \boldsymbol{\beta}^T)^2}{\widehat{\boldsymbol{\beta}}_i \boldsymbol{\Sigma}_{\mathbf{x}} \widehat{\boldsymbol{\beta}}_i^T \cdot \boldsymbol{\beta} \boldsymbol{\Sigma}_{\mathbf{x}} \boldsymbol{\beta}^T}, \tag{13}$$

and use the average of $R^2(\widehat{\boldsymbol{\beta}}_i)$'s to measure the distance between $\mathcal{B}$ and $\mathcal{S}_{y|\mathbf{x}}$. We can also use the squared trace correlation, the average of the squared canonical correlation coefficients between $\widehat{\boldsymbol{\beta}}_1^T\mathbf{x}, \ldots, \widehat{\boldsymbol{\beta}}_K^T\mathbf{x}$ and $\boldsymbol{\beta}_1^T\mathbf{x}, \ldots, \boldsymbol{\beta}_K^T\mathbf{x}$, denoted by $R^2(\widehat{\mathcal{B}})$, as our criterion (see also Hooper 1959). These two criteria are similar, but in the simulation we report all values of $R^2(\widehat{\boldsymbol{\beta}}_i)$'s, so that the performance of the estimation can be described more clearly.

Example 1. Consider the model

$$y = x_1 \cdot (x_2 + x_3 + 1) + \varepsilon, \tag{14}$$

where ε is normally distributed with mean 0 and variance σ^2. In the simulation, $\sigma = .5$. The variables x_1, x_2, and x_3 are the components of $\mathbf{x} = \{x_1, \ldots, x_p\}$ $(p \ge 3)$ and are independent of ε. $\mathbf{x}$ is assumed to have a multivariate normal distribution $\mathrm{N}(\mathbf{0}, \mathbf{I}_p)$, and the CDR space is spanned by two vectors, $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)$ and $\boldsymbol{\beta}_2 = (0, 1, 1, \ldots, 0)$.

We report the simulation results with $n = 400$ and 800 in Table 1. To examine the sensitivity of the estimator to the selection of the number of slices, we considered $c = 10, 20, 40$, and 80. All of the reported results are the averages of 500 independent replicates, and the values in parentheses are the standard deviations.

Looking at Table 1, we see that when n is fixed and p gets larger, the values of $R^2(\boldsymbol{\beta}_i)$'s get smaller. Clearly, dimensionality has an impact on the performance of the estimation. As expected, when n gets larger, the estimation improves. Comparing the case with $(p = 10, n = 400)$ with the case with $(p = 20, n = 800)$ shows that the magnitudes of R^2 are similar, as does comparising the case $(p = 20, n = 400)$ and the case $(p = 40, n = 800)$ also shows this. Thus these examples indicate that the ratio p/n plays an important role in the convergence of the estimator (see the discussion in Remark 2).

For selecting the number of slices, which is equivalent to selecting the number c of data points in each slice, we found that generally the range of the slice numbers for selection is fairly wide. As Zhu and Ng (1995) proved, the root-n consistency

Table 1. Estimation of the CDR Space With Model (14)

	$c = 10$		$c = 20$		$c = 40$		$c = 80$	
p	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$
$n = 400$								
10	.90	.71	.91	.78	.93	.80	.92	.76
	(.07)	(.22)	(.05)	(.16)	(.04)	(.10)	(.05)	(.14)
20	.78	.51	.84	.61	.83	.66	.82	.59
	(.12)	(.21)	(.06)	(.18)	(.07)	(.13)	(.07)	(.15)
30	.69	.35	.78	.52	.75	.53	.74	.49
	(.14)	(.21)	(.08)	(.14)	(.09)	(.16)	(.09)	(.14)
40	.60	.29	.71	.42	.72	.43	.70	.39
	(.13)	(.17)	(.10)	(.16)	(.09)	(.16)	(.9)	(.13)
50	.53	.26	.61	.32	.63	.34	.63	.33
	(.17)	(.18)	(.11)	(.16)	(.10)	(.15)	(.10)	(.12)
$n = 800$								
10	.96	.87	.96	.91	.96	.92	.97	.92
	(.02)	(.09)	(.02)	(.04)	(.02)	(.05)	(.02)	(.04)
20	.90	.72	.92	.81	.92	.83	.93	.83
	(.04)	(.14)	(.04)	(.10)	(.03)	(.06)	(.03)	(.06)
30	.85	.62	.88	.73	.89	.74	.88	.74
	(.05)	(.16)	(.04)	(.11)	(.04)	(.08)	(.04)	(.08)
40	.80	.51	.83	.64	.85	.69	.84	.67
	(.06)	(.18)	(.04)	(.12)	(.05)	(.09)	(.5)	(.10)
50	.75	.41	.80	.57	.81	.62	.80	.60
	(.09)	(.18)	(.07)	(.13)	(.05)	(.12)	(.05)	(.10)

can be maintained in a rather wide range of slice numbers, although the asymptotic variance is affected. The simulation results given herein support this theoretical conclusion. Another finding of interest is that the choice of c is not seriously affected by the dimension p. From $p = 10$–50, the values of R^2 with $c = 20, 40, 80$ are very similar.

To determine the dimension, we must consider the selection of C_n when we use the BIC proposed in this article. For a different problem, the choice of C_n has been discussed by Bai, Krishnaiah, and Zhao (1989). In our simulations, we considered $C_n = c^{-1}W_n$ where c is the number of data points in each slice. We used the term c^{-1} because Zhu and Ng (1995) have proven that the asymptotic variance of $\widehat{\mathbf{\Lambda}}_p$ depends on $(c-1)^{-1}$, which then plays the role of a variance σ^2 that is used in the BIC for variable selection (see Eubank 1999). Selecting C_n is then equivalent to selecting W_n. We tried several values. From Theorem 2, W_n can be selected in a fairly wide range. Note that Schwarz (1978) used $\log n$, and from Theorem 2, a rate of n^α for some α satisfies the conditions. Therefore, we started with a rate of $\log n$ and a rate of $n^{1/3}$, and tried several combinations of these two values: $W_n = .1\log n$, $.5\log(n)$, $.1n^{1/3}$, $.5n^{1/3}$, $(.5\log(n) + .1n^{1/3})$, and $(.5\log(n) + .1n^{1/3})/2$. Table 2 gives the results of $W_n = (.5\log(n) + .1n^{1/3})/2$, which performs better overall than the other W_n's.

A competitor in the literature is the now-standard sequential test (ST) method. ST includes the chi-squared test (CST) suggested by Li (1991) when the distribution of X is normal, as well as the weighted chi-squared test (WCST) proposed by Bura and Cook (2001a) when X is not normally distributed. These two methods were compared. It is noteworthy that when $p \to \infty$, no asymptotic results are available for sequential tests. But in practice both n and p are fixed, and we can still use this method. The proportions of the correct decisions based on the CST reported in Table 2 were obtained in the following way. Let the nominal level be .05. Set the hypothesis $H_0 : d = 1$ versus $H_1 : d > 1$. We recorded the simulation experiments that reject H_0. Then set the hypothesis $H_0 : d = 2$ versus $H_1 : d > 2$. Among the recorded experiments in the previous test, we recorded the decisions that reject the null hypothesis. This procedure proceeds until $H_0 : d = K$ versus $H_1 : d > K$ for $K \le p - 1$, where K is the true dimension of the CDR space. The results reported in Table 2 are the proportions of correct decisions at step K of the 500 simulation experiments.

Because X is normally distributed in this example, we use the CST. Interestingly, from Table 2, we can see that a large c (or, equivalently, a small number of slices) favors the CST, as does high dimensions. The results reported in Table 2 based on the BIC do not show this trend. When we choose the respective c that favors its corresponding methodology, the BIC outperforms the CST for $p \le 30$, whereas the ST works better when $p = 40, 50$.

Table 2. Proportion of Correct Decisions About the Dimension of the CDR Space With Model (14)

	BIC with $W_n = (.5\log(n) + .1n^{1/3})/2$								Chi-squared test							
	$n = 400$				$n = 800$				$n = 400$				$n = 800$			
p	$c = 10$	$c = 20$	$c = 40$	$c = 80$	$c = 10$	$c = 20$	$c = 40$	$c = 80$	$c = 10$	$c = 20$	$c = 40$	$c = 80$	$c = 10$	$c = 20$	$c = 40$	$c = 80$
10	.78	.92	.93	.94	.94	.99	.99	.98	.39	.71	.82	.82	.70	.94	.94	.95
20	.73	.82	.81	.87	.85	.95	.95	.90	.24	.48	.54	.67	.46	.80	.88	.93
30	.69	.62	.54	.65	.81	.82	.90	.70	.17	.34	.40	.51	.31	.65	.75	.89
40	.43	.26	.20	.32	.63	.76	.79	.51	.11	.26	.34	.39	.27	.54	.70	.86
50	.12	.06	.12	.13	.42	.54	.72	.19	.10	.21	.28	.35	.18	.50	.60	.75

Table 3. Estimation of the CDR Space With Model (15)

	$c=5$		$c=10$		$c=20$		$c=40$	
n	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$
Distribution (A)								
100	1.00	.55	1.00	.61	.99	.57	.99	.50
	(0)	(.28)	(0)	(.27)	(.01)	(.28)	(.01)	(.29)
200	1.00	.65	1.00	.75	1.00	.79	1.00	.74
	(0)	(.28)	(0)	(.23)	(0)	(.20)	(0)	(.22)
300	1.00	.737	1.00	.82	1.00	.88	1.00	.86
	(0)	(.25)	(0)	(.18)	(0)	(.11)	(0)	(.16)
400	1.00	.79	1.00	.87	1.00	.90	1.00	.92
	(0)	(.21)	(0)	(.15)	(0)	(.11)	(0)	(.08)
Distribution (B)								
100	.98	.56	.98	.64	.99	.67	.98	.66
	(.03)	(.30)	(.02)	(.31)	(.01)	(.28)	(.02)	(.30)
200	.99	.59	.99	.65	.99	.78	.99	.71
	(.01)	(.32)	(.01)	(.31)	(.01)	(.30)	(.01)	(.22)
300	.99	.64	1.00	.73	1.00	.79	1.00	.82
	(.01)	(.32)	(0)	(.27)	(0)	(.24)	(0)	(.19)
400	1.00	.67	1.00	.76	1.00	.82	1.00	.87
	(.01)	(.31)	(0)	(.27)	(0)	(.23)	(0)	(.17)

Example 2. In this example we consider a model used by Velilla (1998),

$$y = (4 + x_1)(x_2 + x_3 + 2) + \varepsilon, \tag{15}$$

where ε has a normal distribution with mean 0 and variance σ^2 with $\sigma = .5$ in the simulations. The variables x_1, x_2, and x_3 are the components of $\mathbf{x} = \{x_1, \ldots, x_5\}$ and are independent of ε. This model is similar to (14) but has different distributions of $\mathbf{x}$, as follows:

(A) $\mathbf{x}$ has a normal distribution $\mathrm{N}_5(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = \mathrm{diag}(2, 2, 2, 4, 2)$.

(B) $\mathbf{x}$ is obtained by

$$\mathbf{x} = \mathbf{C}\mathbf{v} + \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{w},$$

where

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Here $\mathbf{v} = (v_1, v_2, v_3)'$ consists of three iid uniform random variables on $(-4, 4)$, and $\mathbf{w} = (w_1, w_2)'$ has two independent components, of which w_1 is distributed with a mixture $.5\mathrm{N}(0, 4) + .5\mathrm{N}(0, 16)$ of two normal distributions and w_2 has the $\mathrm{U}(-4, 4)$ distribution, where $\mathrm{U}(-4, 4)$ is the uniform distribution on $(-4, 4)$. Velilla (1998) proved that $\mathbf{x}$ in case (B) also satisfies the linearity condition designed by Li (1991), although it is not elliptically distributed.

Because $p = 5$ is relatively small, we can consider small sample sizes. In the simulation, the sample size was $n = 100, 200, 300, 400$. Tables 3 and 4 report the results with model (15). Case (A) is used to ascertain whether difference between the variances of the components of $\mathbf{x}$ affects the estimation efficiency. Looking at Table 3, we find that in a normal case, the effect is not significant. For the selection of c, the pairs ($n = 100$, $c = 10$), ($n = 200, 300$, $c = 20$), and ($n = 400$, $c = 40$) give large values of R^2. It seems that, overall, $H = 10$ may work well. For case (B), SIR still works, although its performance is not as good as that of case (A) with normal covariates. Again, the estimation is not very sensitive to the choice of the slice number; $c = 20$ works well, and $H = 10$ may still be considered as a working value for practical use.

For the determination of dimensionality, Table 4 reports the results of the BIC and the CST for (A) and WCST for (B) with $c = 5, 10, 20$. The normal distribution of $\mathbf{x}$ clearly favors the two methods, but for the very nonnormal case (B), estimation for K is more difficult. A comparison of these two methods indicates that the BIC is clearly better than the CST/WCST regardless of whether or not the distribution of $\mathbf{x}$ is normal.

Example 3. In this example we consider a model in which the CDR space is spanned by $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)^T$ in mean and $\boldsymbol{\beta}_2 = (0, 1, 1, \ldots, 0)^T$ in variance as

$$y = (4 + x_1) + (x_2 + x_3 + 2) \cdot \varepsilon. \tag{16}$$

Here ε has a normal distribution with mean 0 and variance σ^2 with $\sigma = .5$ in the simulation. The variables $x_1, \ldots, x_3$ are

Table 4. Proportion of Correct Decisions About the Dimension of the CDR Space With Model (15)

	BIC with $W_n = (.5\log(n) + .1n^{1/3})/2$						Chi-squared test			Weighted chi-squared test		
	Distribution (A)			Distribution (B)			Distribution (A)			Distribution (B)		
n	$c=5$	$c=10$	$c=20$	$c=5$	$c=10$	$c=20$	$c=5$	$c=10$	$c=20$	$c=5$	$c=10$	$c=20$
100	.53	.57	.60	.33	.39	.34	.17	.31	.36	.24	.32	.37
200	.68	.81	.88	.44	.41	.53	.33	.57	.75	.28	.36	.50
300	.79	.92	.95	.46	.51	.68	.44	.73	.90	.30	.50	.58
400	.88	.95	.98	.53	.63	.75	.57	.81	.93	.38	.55	.65

Table 5. Estimation of the CDR Space With Model (16)

	$c=5$		$c=10$		$c=20$		$c=40$	
n	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$
Distribution (A)								
100	.95	.51	.97	.53	.97	.50	.95	.33
	(.06)	(.29)	(.03)	(.29)	(.03)	(.30)	(.04)	(.25)
200	.98	.52	.98	.61	.99	.64	.98	.64
	(.02)	(.31)	(.03)	(.30)	(.01)	(.28)	(.01)	(.27)
300	.99	.61	.99	.70	.99	.76	.99	.72
	(.01)	(.28)	(.01)	(.28)	(.01)	(.23)	(.01)	(.30)
400	.99	.63	.99	.76	.99	.81	.99	.83
	(.01)	(.30)	(.01)	(.23)	(.01)	(.21)	(.01)	(.17)
Distribution (B)								
100	.99	.30	.99	.32	.99	.40	.97	.30
	(.01)	(.29)	(.01)	(.29)	(.01)	(.31)	(.03)	(.26)
200	.99	.30	.99	.36	.99	.41	.99	.31
	(.01)	(.28)	(.01)	(.30)	(.01)	(.32)	(.01)	(.33)
300	1.00	.30	.99	.36	.99	.43	.99	.41
	(0)	(.28)	(0)	(.31)	(0)	(.32)	(0)	(.27)
400	1.00	.34	1.00	.36	1.00	.46	1.00	.48
	(0)	(.29)	(0)	(.31)	(0)	(.32)	(0)	(.33)

the components of $\mathbf{x} = \{x_1, \ldots, x_p\}$ $(p = 5)$ and are independent of ε, and $\mathbf{x}$ has the same distributions (A) and (B) as in model (15). For this model, the situation is similar to that of Example 2; the methods work better in the normal case (A). As to the selection of c, we come to a similar conclusion that SIR is not sensitive to the number of slices and $c = 20$ works well with the sample size that we used. In addition, $H = 10$ may be a good choice for practical use.

For determining dimensionality, we considered $c = 5, 10, 20$ when the sample size was $\leq$300. Looking at Table 6, we clearly see that the dimension is difficult to estimate, especially in the nonnormal case (B). This is because, in contrast to the model of Example 2, one dimension is included in the variance. Hence this model does not favor the SIR. Therefore, we considered a larger sample size to explore how large a sample size is needed to obtain a good estimator; $n = 400, 800, 1{,}600, 3{,}200$ were tested. For sequential test method, similar to those of Example 2, in the normal case (A) we use the CST and in the nonnormal case (B) we use the WCST. In a normal case, the estimation will be satisfactory when $n \geq 800$; however, for the nonnormal case (B), $n = 3{,}200$ is still not large enough. This observation demonstrates that we may need to use other means to obtain a better structure of data such as that suggested by Cook and Nachtsheim (1994). Between the BIC and the CST/WCST, the former generally performs better, although in case (B) the WCST performs slightly better when the sample size is large.

A comparison of models (15) and (16) provides evidence that SIR has difficulty estimating the CDR space when basis is in the variance part.

Example 4. Finally, we consider a model with a three-dimensional CDR space,

$$y = x_1(x_2 + x_3 + 2) + (x_4 + x_5 + 2)^3 + \varepsilon, \qquad (17)$$

where ε is normally distributed with mean 0 and variance σ^2. In the simulation, $\sigma = .5$. The variables x_1, x_2, x_3, x_4, and x_5 are the components of $\mathbf{x} = \{x_1, \ldots, x_p\}$ $(p \geq 5)$ and are independent of ε. $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ is normally distributed with $\mathrm{N}_p(\mathbf{0}, \mathbf{I}_p)$. The CDR space is spanned by $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)^T$, $\boldsymbol{\beta}_2 = (0, 1, 1, \ldots, 0)^T$, and $\boldsymbol{\beta}_3 = (0, 0, 0, 1, 1, \ldots, 0)^T$.

Based on our experience, we believe that a small sample size cannot provide a good estimation of the CDR space of this model. Here we report the results with $n = 400$ and $n = 3{,}200$ to demonstrate the difficulty of estimating the CDR space and the large sample size needed to obtain an acceptable estimation.

Table 7 shows that the $R^2(\widehat{\boldsymbol{\beta}}_3)$'s are small, even when $p = 10$. When n is increased to 3,200, the $R^2(\widehat{\boldsymbol{\beta}}_3)$'s become acceptably large when $p \leq 20$. Increasing the dimension p significantly reduces the performance of the estimation. Clearly, for larger p,

Table 6. Proportion of Correct Decisions About the Dimension of the CDR Space With Model (16)

	BIC with $W_n = (.5\log(n) + .1n^{1/3})/2$							*Chi-squared test*			*Weighted chi-squared test*		
	Distribution (A)			*Distribution (B)*				*Distribution (A)*			*Distribution (B)*		
n	$c=5$	$c=10$	$c=20$	$c=5$	$c=10$	$c=20$	n	$c=5$	$c=10$	$c=20$	$c=5$	$c=10$	$c=20$
100	.28	.28	.23	.20	.16	.13	100	.08	.15	.16	.02	.06	.09
200	.29	.43	.48	.13	.13	.13	200	.12	.24	.38	.04	.08	.08
300	.34	.49	.61	.09	.11	.13	300	.21	.34	.51	.04	.05	.10
	Distribution (A)			*Distribution (B)*				*Distribution (A)*			*Distribution (B)*		
	$c=20$	$c=40$	$c=80$	$c=20$	$c=40$	$c=80$		$c=20$	$c=40$	$c=80$	$c=20$	$c=40$	$c=80$
400	.72	.79	.69	.11	.17	.19	400	.60	.71	.63	.09	.16	.19
800	.92	.98	.99	.11	.20	.28	800	.87	.91	.93	.14	.24	.31
1,600	.99	1.00	.99	.12	.27	.47	1,600	.95	.95	.94	.20	.33	.52
3,200	.99	1.00	1.00	.15	.38	.76	3,200	.96	.93	.96	.32	.52	.77

Table 7. Estimation of the CDR Space With Model (17)

	$c = 10$			$c = 20$			$c = 40$			$c = 80$		
p	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$	$R^2(\widehat{\beta}_3)$	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$	$R^3(\widehat{\beta}_3)$	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$	$R^2(\widehat{\beta}_3)$	$R^2(\widehat{\beta}_1)$	$R^2(\widehat{\beta}_2)$	$R^2(\widehat{\beta}_3)$
$n = 400$												
10	1.00	.70	.30	1.00	.77	.33	1.00	.80	.33	1.00	.74	.24
	(.001)	(.19)	(.22)	(.001)	(.15)	(.23)	(.001)	(.13)	(.23)	(.002)	(.14)	(.18)
20	.99	.46	.20	1.00	.56	.21	.99	.60	.21	.99	.53	.14
	(.002)	(.22)	(.16)	(.002)	(.19)	(.15)	(.002)	(.17)	(.16)	(.003)	(.06)	(.12)
30	.99	.30	.16	.99	.40	.16	.99	.47	.16	.99	.41	.12
	(.003)	(.18)	(.13)	(.003)	(.18)	(.13)	(.003)	(.17)	(.13)	(.004)	(.14)	(.11)
40	.99	.23	.13	.99	.31	.13	.99	.37	.13	.98	.31	.10
	(.004)	(.14)	(.11)	(.003)	(.17)	(.11)	(.004)	(.15)	(.10)	(.005)	(.14)	(.09)
$n = 3{,}200$												
10	1.00	.94	.47	1.00	.95	.60	1.00	.95	.70	1.00	.95	.69
	(.0003)	(.03)	(.26)	(.0003)	(.03)	(.25)	(.0003)	(.02)	(.20)	(.0003)	(.03)	(.19)
20	1.00	.87	.26	1.00	.89	.37	1.00	.90	.43	1.00	.89	.47
	(.0004)	(.05)	(.19)	(.0004)	(.04)	(.21)	(.0004)	(.04)	(.21)	(.0005)	(.04)	(.19)
30	1.00	.79	.17	1.00	.83	.22	1.00	.85	.31	1.00	.84	.36
	(.0005)	(.07)	(.14)	(.0005)	(.05)	(.17)	(.0005)	(.04)	(.18)	(.0006)	(.05)	(.16)
40	1.00	.72	.12	1.00	.78	.18	1.00	.79	.24	1.00	.79	.29
	(.0006)	(.09)	(.11)	(.0005)	(.06)	(.14)	(.0006)	(.05)	(.15)	(.0007)	(.05)	(.14)

we should use larger datasets to obtain a reasonable estimation. The selection of c seems similar to that of the previous examples. Within a large range of c (from 20 to 80), the performance of SIR is similar.

The determination of the dimension is also difficult. To provide insights into how the dimension of the CDR space can be estimated, in Table 8 we report the proportions of the estimated dimensions of the CDR space in the 500 replications. Neither the BIC nor the CST works for $n = 400$ (see Table 8); a large sample size is needed. We also report the simulation results with $n = 3{,}200$. Table 8 clearly shows that when the sample size is equal to 400, the BIC tends to estimate K as 2 and the CST seems to estimate K even smaller. This becomes clearer when the dimension p of X is large. When $n = 3{,}200$, both methods can obtain a high proportion of correct decisions, and the BIC outperforms the CST, especially when p is large. Another interesting observation is that, in contrast to the performance of estimation with model (14), the CST gets worse when p gets larger.

In summary, estimation of the SIR matrix is not very sensitive to the choice of c or, equivalently, to the number of slices. The choice of c is not greatly affected by the dimension p, and thus we can choose the number of slices regardless of the dimension of the covariates. Based on our simulations, the number of slices chosen can fall between 10 and 20. Compared with the sequential test method for determining the dimension of the CDR space, the advantages of the BIC are that it does not rely on the limit distribution of the estimator of the SIR matrix and that it can be readily applied to other dimension-reduction methods, such as the SAVE and the MAVE. However, the need to select C_n is a disadvantage. Whether there is a data-driven selection for C_n deserves further study.

APPENDIX: PROOFS OF THEOREMS

A.1 Notation

For the sake of convenience, we have assumed that n/c is an integer. Here we define some notation. Recalling (4) and (5), we define $\mathbf{Y}_n = (y_{(1)}, \ldots, y_{(n)})^T$, $\mathbf{X}_n = (\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(n)})_{p\times n}$ as a $p \times n$ matrix, $\mathbf{X}_h = (\mathbf{x}_{(h,1)}, \ldots, \mathbf{x}_{(h,c)})_{p\times c}, h = 1, 2, \ldots, H$, as $p \times c$ matrices, and for any integer t, $\mathbf{e}_t = (1, \ldots, 1)^T$ as a $t \times 1$ column vector and $\boldsymbol{\mathcal{E}}_t = (\mathbf{e}_t, \ldots, \mathbf{e}_t)^T_{t\times p}$ as a $t \times p$ matrix. We further define more matrices. Set $\mathbf{I}_c$ to be the $c \times c$ identity matrix and

$$\mathbf{P}_c = \mathbf{I}_c - \frac{1}{c}\mathbf{e}_c\mathbf{e}_c^T, \qquad \mathbf{Q}_c = \mathbf{P}_c - \left(1 - \frac{1}{c}\right)\mathbf{I}_c = \frac{1}{c}(\mathbf{I}_c - \mathbf{e}_c\mathbf{e}_c^T);$$

$$\mathbf{M}_n = \big(\mathbf{m}\big(y_{(1)}\big), \ldots, \mathbf{m}\big(y_{(n)}\big)\big)_{p\times n},$$

$$\mathbf{M}_h = \big(\mathbf{m}\big(y_{(h,1)}\big), \ldots, \mathbf{m}\big(y_{(h,c)}\big)\big)_{p\times c};$$

$$\boldsymbol{\varepsilon}_n = \big(\varepsilon_{(1)}, \ldots, \varepsilon_{(n)}\big)_{p\times n}, \qquad \boldsymbol{\varepsilon}_{h,c} = \big(\varepsilon_{(h,1)}, \ldots, \varepsilon_{(h,c)}\big)_{p\times c},$$

and

$$\widetilde{\mathbf{M}}_n = \big[\big\{\mathbf{m}\big(y_{(1)}\big) - \mathbf{m}\big(y_{(2)}\big)\big\}, \ldots, \big\{\mathbf{m}\big(y_{(n-1)}\big) - \mathbf{m}\big(y_{(n)}\big)\big\}, \mathbf{m}\big(y_{(n)}\big)\big]_{p\times n},$$

Table 8. Proportions of the Estimated Dimensions of the CDR Space With Model (17), $H = 10$

	BIC with $W_n = (.5\log(n) + .1n^{1/3})/2$								Chi-squared test							
	$n = 400$				$n = 3{,}200$				$n = 400$				$n = 3{,}200$			
p	10	20	30	40	10	20	30	40	10	20	30	40	10	20	30	40
$K = 1$	.05	.07	.05	.04	0	0	0	.04	.14	.35	.44	.58	0	0	0	0
$K = 2$	.86	.84	.85	.85	.02	.05	.07	0	.75	.59	.51	.40	.02	.15	.30	.39
$K = 3$	.09	.09	.10	.11	.98	.95	.91	.89	.11	.06	.05	.02	.94	.83	.67	.59
$K = 4$	0	0	0	0	0	0	.02	.07	0	0	0	0	.04	.02	.03	.02

NOTE: The true dimension is three.

where the $\varepsilon_{(i)}$'s are as defined in (5) and $\varepsilon_{(h,j)} = \varepsilon_{(c(h-1)+j)}$. Furthermore, for any t,

$$\mathbf{T}_t = \begin{pmatrix} 1 & -1 & & \\ & 1 & \ddots & \\ & & \ddots & -1 \\ & & & 1 \end{pmatrix}_{t\times t} \quad \text{and}$$

$$\mathbf{S}_t = \mathbf{T}_t^{-1} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ & 1 & \cdots & \vdots \\ & & \ddots & 1 \\ & & & 1 \end{pmatrix}_{t\times t}.$$

From the definitions of $\widetilde{\mathbf{M}}_n$ and $\mathbf{T}_n$, it is clear that $\widetilde{\mathbf{M}}_n^T = \mathbf{T}_n \mathbf{M}_n^T$.

Now we can rewrite $\widehat{\mathbf{\Lambda}}_p$ in terms of the foregoing notation in a matrix form. Because $\mathbf{X}_n = \mathbf{M}_n + \boldsymbol{\varepsilon}_n$,

$$\begin{aligned} \widehat{\mathbf{\Lambda}}_p &= \frac{1}{H(c-1)} \sum_{h=1}^{H} \mathbf{X}_h \mathbf{P}_c \mathbf{X}_h^T \\ &= \frac{1}{H(c-1)} \mathbf{X}_n (\mathbf{I}_H \otimes \mathbf{P}_c) \mathbf{X}_n^T \\ &= \frac{1}{H(c-1)} \boldsymbol{\varepsilon}_n (\mathbf{I}_H \otimes \mathbf{P}_c) \boldsymbol{\varepsilon}_n^T \\ &\quad + \frac{1}{H(c-1)} \{\boldsymbol{\varepsilon}_n (\mathbf{I}_H \otimes \mathbf{P}_c) \mathbf{M}_n^T + \mathbf{M}_n (\mathbf{I}_H \otimes \mathbf{P}_c) \boldsymbol{\varepsilon}_n^T\} \\ &\quad + \frac{1}{H(c-1)} \mathbf{M}_n (\mathbf{I}_H \otimes \mathbf{P}_c) \mathbf{M}_n^T \\ &=: \mathbf{J}_1 + \mathbf{J}_2 + \mathbf{J}_3, \end{aligned} \tag{A.1}$$

where "$\otimes$" represents the Kronecker product. Note that $\mathbf{P}_c + \mathbf{Q}_c = (1 - 1/c)\mathbf{T}_c$. Then $\mathbf{I}_H \otimes \mathbf{P}_c + \mathbf{I}_H \otimes \mathbf{Q}_c = (1 - 1/c)\mathbf{I}_H$. $\mathbf{J}_1$ is then decomposed as

$$\begin{aligned} \mathbf{J}_1 &= \frac{1}{H(c-1)} \left\{ \left(1 - \frac{1}{c}\right) \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^T \right\} + \frac{1}{H(c-1)} \{\boldsymbol{\varepsilon}_n (\mathbf{I}_H \otimes \mathbf{Q}_c) \boldsymbol{\varepsilon}_n^T\} \\ &=: \mathbf{K}_1 + \mathbf{K}_2. \end{aligned} \tag{A.2}$$

It is easy to see that

$$\|\mathbf{\Lambda}_p - \widehat{\mathbf{\Lambda}}_p\| \le \|\mathbf{\Lambda}_p - \mathbf{K}_1\| + \|\mathbf{K}_2\| + \|\mathbf{J}_2\| + \|\mathbf{J}_3\|. \tag{A.3}$$

We need only derive the convergence rate of the foregoing four terms on the right side of (A.3). For this, we introduce the following lemmas.

A.2 Lemmas

Throughout this and the following sections, b is a constant that may take different values with each appearance.

Lemma A.1. Under the conditions of Theorem 1,

$$\lim_{n\to\infty} \frac{1}{n^r \sqrt{p}} \sum_{i=2}^{n} \{\|\mathbf{m}(y_{(i)}) - \mathbf{m}(y_{(i-1)})\|\} = 0 \quad \text{a.s.}$$

Proof. If all $y_i \in [-B, B]$ for some $B > 0$, then the result follows from condition (B) of Theorem 1. Now we need only consider the case where the support of $y_i^{(n)}$ is unbounded. It suffices to show that the sums

$$S_{n1} = \frac{1}{n^r \sqrt{p}} \sum_{i=[n\widehat{\theta}]+1}^{[n(1-\widehat{\theta})]} \{\cdots\},$$

$$S_{n2} = \frac{1}{n^r \sqrt{p}} \sum_{i=2}^{[n\widehat{\theta}]} \{\cdots\},$$

and

$$S_{n3} = \frac{1}{n^r \sqrt{p}} \sum_{i=[n(1-\widehat{\theta})+1]}^{n} \{\cdots\}$$

converge to 0 almost surely as $n \to \infty$, where $[a]$ is the largest integer part of a.

We deal with S_{n1} first. Recall the definition of $G_2(x)$ and $G_1(x)$ in condition (C) of Theorem 1. For any fixed δ, we define two events: For any $\beta > 0$,

$$A_n = \{y_{([n\delta])} > G_2^{-1}(\beta)\} \quad \text{and}$$

$$B_n = \{y_{([n(1-\delta)])} < G_1^{-1}(1-\beta)\}.$$

For $\delta < 1/2$, select $0 < \beta < \delta(1/2)^{1/\delta}$ and a B_0 such that $B_0 > G_1^{-1}(1-\beta)$ and $-B_0 < G_2^{-1}(\beta)$. Then for any small η, for n sufficiently large, we can derive that, combining (6) with condition (D) of Theorem 1,

$$\begin{aligned} P\{S_{n1} > \eta\} &= P\left\{ \frac{1}{n^r \sqrt{p}} \sum_{i=[n\delta]+1}^{[n(1-\delta)]} \|\mathbf{m}(y_{(i)}) - \mathbf{m}(y_{(i-1)})\| > \eta \right\} \\ &\le P\left\{ \frac{1}{n^r \sqrt{p}} \sum_{i=[n\delta]+1}^{[n(1-\delta)]} \|\mathbf{m}(y_{(i)}) - \mathbf{m}(y_{(i-1)})\| > \eta, A_n B_n \right\} \\ &\quad + P\{(A_n B_n)^c\} \\ &\le P(A_n^c) + P(B_n^c) + o(1), \end{aligned}$$

where $A_n B_n$ represents the intersection of two events A_n and B_n.

Now we deal with $P(A_n^c)$ and $P(B_n^c)$. Suppose that $w_{(i)}$ is the ith-order statistic of iid random variables $w_i, i = 1, \ldots, n$, that has the distribution $G_2(x)$ of W_2 because by condition (C), $Y \succ W_2$, and then $y_{([n\delta])} \succ w_{([n\delta])}$. Thus,

$$P(A_n^c) = P\{y_{([n\delta])} \le G_2^{-1}(\beta)\} \le P\{w_{([n\delta])} \le G_2^{-1}(\beta)\}.$$

From the properties of order statistics, we know that

$$\begin{aligned} &P\{w_{([n\delta])} \le G_2^{-1}(\beta)\} \\ &\qquad \le \frac{n!}{([n\delta]-1)!(n-[n\delta])!} \int_0^{\beta} t^{[n\delta]-1} (1-t)^{n-[n\delta]}\, dt. \end{aligned}$$

By Stirling's formula, $n!$ is asymptotically equal to $(\frac{n}{e})^n \sqrt{2\pi n}$; then we have

$$P(A_n^c) = O\left\{ \left(\frac{\beta}{\delta}\right)^{[n\delta]} \frac{1}{(1-\delta)^{n-[n\delta]}} \right\}.$$

Recall that $\frac{\beta}{\delta} < (\frac{1}{2})^{1/\delta}$ and $\delta < \frac{1}{2}$. Then there is a ρ_1 with $0 < \rho_1 < 1$,

$$P(A_n^c) = O(\rho_1^n). \tag{A.4}$$

Similarly, we can obtain that for some ρ_2 with $0 < \rho_2 < 1$,

$$P(B_n^c) = O(\rho_2^n). \tag{A.5}$$

From (A.4) and (A.5), the strong convergence of S_{n1} follows the Borel–Cantelli lemma.

Because the arguments for proving the convergence of S_{n2} and S_{n3} are similar, we present only the proof for S_{n3} here. Invoking condition (B) of Theorem 1 and the monotonic property of $\mathbf{M}(\cdot)$ assumed in (7), we have

$$\begin{aligned} S_{n3} &= \frac{1}{n^r \sqrt{p}} \sum_{i=[n(1-\delta)+1]}^{n} \|\mathbf{m}(y_{(i)}) - \mathbf{m}(y_{(i-1)})\| \\ &\le n^{-r} \sum_{i=[n(1-\delta)+1]}^{n} |\mathbf{M}(y_{(i)}) - \mathbf{M}(y_{(i-1)})| \end{aligned}$$

$$= n^{-r}\left|\mathbf{M}\left(y_{[n(1-\delta)]}\right) - \mathbf{M}\left(y_{(n)}\right)\right|$$
$$\le n^{-r}\left|\mathbf{M}\left(y_{(n)}\right)\right|.$$

Without loss of generality, we assume that $\mathbf{M}(y_{(n)})$ is positive. Then we need only prove that $n^{-r}\mathbf{M}(y_{(n)}) \to 0$ a.s. $n \to \infty$. Note that for any $\eta > 0$,

$$P\left\{n^{-r}\mathbf{M}\left(y_{(n)}\right) > \eta\right\} = 1 - \left[P\{n^{-r}\mathbf{M}(y) \le \eta\}\right]^n$$
$$= 1 - \left[1 - P\{n^{-r}\mathbf{M}(y) > \eta\}\right]^n.$$

By condition (C), there is a W_1 such that $Y \prec W_1$, and from condition (D), it is easily derived that for a constant $b > 0$,

$$P\{n^{-r}\mathbf{M}(y) > \eta\} \le P\{n^{-r}\mathbf{M}(W_1) > \eta\} \le \frac{b}{(\varepsilon n^r)^l}.$$

Using the inequality $(1-\alpha)^n \ge 1 - n\alpha$, $0 \le \alpha \le 1$, we obtain that

$$P\left\{n^{-r}\mathbf{M}\left(y_{(n)}\right) > \eta\right\} = 1 - \left[1 - P\{n^{-r}\mathbf{M}(y) > \eta\}\right]^n \le n\frac{b}{(\eta n^r)^l}.$$

As $l \times r > 2$, the Borel–Cantelli lemma yields $n^{-r}\mathbf{M}(y_{(n)}) \to 0$ a.s. $n \to \infty$, and then the convergence of S_{n3} is proved. The proof of the lemma is finished.

Lemma A.2. If $\mathbf{M}_1$ is an $n \times p$ real-valued matrix, $\mathbf{M}_2$ is a $p \times n$ real-valued matrix, and $\mathbf{M}_3$ and $\mathbf{M}_4$ are two $n \times n$ real-valued matrices, then

$$\|\mathbf{M}_1\mathbf{M}_2\| \le \lambda_{\max}^{1/2}(\mathbf{M}_2\mathbf{M}_2^T)\|\mathbf{M}_1\|,$$
$$\lambda_{\max}(\mathbf{M}_1\mathbf{M}_2) = \lambda_{\max}(\mathbf{M}_2\mathbf{M}_1),$$

and

$$\lambda_{\max}(\mathbf{M}_3\mathbf{M}_4) \le \lambda_{\max}(\mathbf{M}_3)\lambda_{\max}(\mathbf{M}_4),$$

where $\lambda_{\max}(\cdot)$ represents the maximum eigenvalue (see Bai 1999).

Lemma A.3. Under the conditions of Theorem 1, we have

$$\max_j \max_i \left|\varepsilon_{(i)}(j)\right| = \max_j \max_i |\varepsilon_i(j)|$$
$$= O_p\left(n^{1/l}p\right) \text{ or } O\left(n^{1/l}p(\log n)^{2/l}\right) \quad \text{a.s.},$$

where $j = 1, \ldots, p$ and $i = 1, \ldots, n$, $\varepsilon_{(i)}(j)$ and $\varepsilon_i(j)$ are the jth components of random vector $\varepsilon_{(i)}$ and ε_i defined in (5).

Proof. For any $j = 1, \ldots, p$, let $|\varepsilon_{nj}| = \max_{1\le i\le n}|\varepsilon_{(i)}(j)|$. It is clear that $|\varepsilon_{nj}| = \max_i |\varepsilon_i(j)|$ and $\max_j \max_i |\varepsilon_{(i)}(j)| = \max_i |\varepsilon_{nj}|$. We investigate only the almost-sure convergence, because the convergence in probability is much easier to obtain. Choosing $\eta = bn^{1/l}p(\log n)^{1/2}$, together with the iid property of variables ε_i, we have that

$$P\left(\max_j |\varepsilon_{nj}| \ge \eta\right)$$
$$\le \max_j P\left\{|\varepsilon_{nj}| \ge bn^{1/l}(\log n)^{2/l}\right\}$$
$$\le \max_j \left[1 - P\left\{|\varepsilon_{nj}| \le bn^{1/l}(\log n)^{2/l}\right\}\right]$$
$$= \max_j \left(1 - \left[1 - P\left\{|\varepsilon_1(j)| \ge bn^{1/l}(\log n)^{2/l}\right\}\right]^n\right).$$

Furthermore, invoking condition (A), we can easily derive the finiteness of the lth moment of $|\varepsilon_i(j)|$, and

$$P\left\{|\varepsilon_1(j)| \ge bn^{1/l}(\log n)^{2/l}\right\} \le \frac{b}{n(\log n)^2}.$$

Hence

$$\max_j \left(1 - \left[1 - P\left\{|\varepsilon_1(j)| \ge bn^{1/l}(\log n)^{2/l}\right\}\right]^n\right)$$
$$\le 1 - \left\{1 - \frac{b}{n(\log n)^2}\right\}^n \le 1 - \exp\left\{-\frac{b}{(\log n)^2}\right\}$$
$$\le \frac{b}{(\log n)^2}.$$

The Borel–Cantelli lemma implies strong convergence of the subsequence $\{\max_{1\le j\le p}|\varepsilon_{e^m j}| : m \ge 1\}$ of $\{\max_{1\le j\le p}|\varepsilon_{nj}| : n \ge 1\}$. Using the subsequence approach, we need only show that for any n with $e^m \le n \le e^{m+1}$,

$$P\left\{\max_{e^m\le n\le e^{m+1}}\left(\max_j |\varepsilon_{nj}| - \max_j |\varepsilon_{e^m j}|\right) \ge \eta\right\}$$
$$= P\left\{\max_j |\varepsilon_{e^{m+1}j}| - \max_j |\varepsilon_{e^m j}| \ge \eta\right\}$$
$$\le \max_j P\left\{\max_{e^m\le n\le e^{m+1}} |\varepsilon_{nj}| \ge be^{m/l}m^{2/l}\right\}$$
$$\le \frac{b}{m^2}.$$

The last inequality comes from the same argument as the foregoing with the upper bound $b/(\log\ n)^2$. This means that the difference between ε_{nj} and its subsequence also converges almost surely to 0 at the rate $n^{-1/l}p(\log n)^{2/l}$.

For the convergence in probability, we do not need to use the subsequence approach. The rate is clearly $O(n^{-1/l}p)$. The proof is completed.

A.3 Proof of Theorem 1

The Proof for Strong Convergence. As described earlier, to prove the strong convergence of $\widehat{\mathbf{\Lambda}}_p$ we need only prove convergence of the right side of (A.3). The proof is divided into four steps:

a. $|\mathbf{J}_2| = o(n^{-1+1/l+r}p^{3/2}(\log n)^{2/l})$ a.s. as $n \to \infty$.

From the definition of $\mathbf{J}_2$ in (A.1), we need only consider the term

$$\left\|\frac{1}{H(c-1)}\boldsymbol{\varepsilon}_n(\mathbf{I}_H \otimes \mathbf{P}_c)\widetilde{\mathbf{M}}_n^T\right\|.$$

Recalling the definitions of $\mathbf{P}_c$ and $\mathbf{e}_c\mathbf{e}_c^T$ in Section A.1, we have $\mathbf{P}_c\mathbf{e}_c\mathbf{e}_c^T = 0$. Invoking the equations $\widetilde{\mathbf{M}}_n^T = \mathbf{T}_n\widetilde{\mathbf{M}}_n^T$ and $\mathbf{S}_n = \mathbf{T}_n^{-1}$, we can derive that, together with the properties of the Kronecker product and recalling that $\mathbf{S}_n = \mathbf{T}_n^{-1}$,

$$\|\boldsymbol{\varepsilon}_n(\mathbf{I}_H \otimes \mathbf{P}_c)\widetilde{\mathbf{M}}_n^T\| = \|\boldsymbol{\varepsilon}_n(\mathbf{I}_H \otimes \mathbf{P}_c)\mathbf{S}_n\widetilde{\mathbf{M}}_n^T\| = \left\|\boldsymbol{\varepsilon}_n\{\mathbf{I}_H \otimes (\mathbf{P}_c\mathbf{S}_c)\}\widetilde{\mathbf{M}}_n^T\right\|.$$

From the definitions of $\boldsymbol{\varepsilon}_n$ and $\boldsymbol{\mathcal{E}}_n$ in Section A.1, we have

$$\left\|\boldsymbol{\varepsilon}_n(\mathbf{I}_H \otimes (\mathbf{P}_c\mathbf{S}_c))\widetilde{\mathbf{M}}_n^T\right\| \le \max_{1\le i\le n, 1\le j\le p}\left|\varepsilon_{(i)}(j)\right|\left\|\mathcal{E}_n\{\mathbf{I}_H \otimes (\mathbf{P}_c\mathbf{S}_c)\}\widetilde{\mathbf{M}}_n^T\right\|.$$

Also note that in matrix $\mathbf{I}_H \otimes (\mathbf{P}_c\mathbf{S}_c)$, all elements of the last column are 0 and the elements such as $a_{k_1k_2} = 0$ if $|k_1 - k_2| > c$ and $a_{k_1k_2} \le 1$ otherwise. Then, by the first inequality of Lemma A.2, we have

$$\left\|\mathcal{E}_n\{\mathbf{I}_H \otimes (\mathbf{P}_c\mathbf{S}_c)\}\widetilde{\mathbf{M}}_n^T\right\| \le 2c\sum_{i=2}^n \left\|\mathbf{m}\left(y_{(i)}\right) - \mathbf{m}\left(y_{(i-1)}\right)\right\|.$$

Applying Lemma A.3 to $\varepsilon_{(i)}(j)$ and Lemma A.1 to $\widetilde{\mathbf{M}}_n^T$, we have

$$\max_{1\le j\le p}\max_{1\le i\le n}\left|\varepsilon_{(i)}(j)\right| = O\left(n^{1/l}p(\log n)^{2/l}\right) \quad \text{a.s.}$$

and

$$\sum_{i=2}^n \left\|\mathbf{m}\left(y_{(i)}\right) - \mathbf{m}\left(y_{(i-1)}\right)\right\| = O(n^r\sqrt{p}) \quad \text{a.s.}$$

Hence we almost surely have

$$\frac{1}{H(c-1)}\left\|\boldsymbol{\varepsilon}_n\{\mathbf{I}_H\otimes(\mathbf{P}_c\mathbf{S}_c)\}\widetilde{\mathbf{M}}_n^T\right\| = O\big(n^{-1+1/l+r}p^{3/2}(\log n)^{1/l}\big).$$

b. $\|\mathbf{J}_3\| = O(n^{2r-1}p)$ a.s. as $n\to\infty$.

By Lemma A.2 and the property of the Kronecker product, we can obtain that, from the definition of $\mathbf{M}_n^T$ and $\widetilde{\mathbf{M}}_n^T$,

$$\|\mathbf{M}_n(\mathbf{I}_H\otimes\mathbf{P}_c)\mathbf{M}_n^T\| = \|\widetilde{\mathbf{M}}_n(\mathbf{I}_H\otimes\mathbf{S}_c^T\mathbf{P}_c\mathbf{S}_c)\widetilde{\mathbf{M}}_n^T\|.$$

Note that the last row and last column of the matrix $\mathbf{I}_H\otimes\mathbf{S}_c^T\mathbf{P}_c\mathbf{S}_c$ are both 0. Let $\widetilde{\mathbf{M}}_{n1}$ be a matrix with elements equal to those of $\widetilde{\mathbf{M}}_n$ except that the elements of the last column are 0. Then the matrix $\widetilde{\mathbf{M}}_n(\mathbf{I}_H\otimes\mathbf{S}_c^T\mathbf{P}_c\mathbf{S}_c)\widetilde{\mathbf{M}}_n^T$ can be rewritten as $\widetilde{\mathbf{M}}_{n1}(\mathbf{I}_H\otimes\mathbf{S}_c^T\mathbf{P}_c\mathbf{S}_c)\widetilde{\mathbf{M}}_{n1}^T$. It is also clear that $\lambda_{\max}^{1/2}(\mathbf{S}_c^T\mathbf{P}_c\mathbf{S}_c)\le b$ for some $b>0$. Thus, invoking Lemma A.1,

$$\begin{aligned}\|\mathbf{J}_3\| &= \frac{1}{H(c-1)}\|\widetilde{\mathbf{M}}_n(\mathbf{I}_H\otimes\mathbf{P}_c)\widetilde{\mathbf{M}}_{n1}^T\|\\ &\le \frac{bc}{n(c-1)}\|\widetilde{\mathbf{M}}_{n1}^T\|^2 = O(n^{2r-1}p)\quad\text{a.s.}\end{aligned}$$

The proof is finished.

c. $\|\mathbf{K}_2\| = O(n^{-1/4}p(\log n)^{1/2})$ a.s. as $n\to\infty$.

Note that $\|\mathbf{K}_2\|^4$ can be written as

$$\|\mathbf{K}_2\|^4 = \frac{1}{H^4(c-1)^4c^4}\left\|\sum_{h=1}^{H}\boldsymbol{\varepsilon}_{h,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h,c}^T\right\|^4,$$

where $\boldsymbol{\varepsilon}_{h,c}$ is defined in Section A.1 and

$$\mathbf{Q}_c' = c\mathbf{Q}_c = \begin{pmatrix} 0 & -1 & \cdots & -1\\ -1 & \ddots & -1 & -1\\ \vdots & & & \vdots\\ -1 & \cdots & -1 & 0\end{pmatrix}_{c\times c}.$$

Furthermore,

$$\begin{aligned}&E\left\|\sum_{h=1}^{H}\boldsymbol{\varepsilon}_{h,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h,c}^T\right\|^4\\ &\quad= E\left[\operatorname{tr}\left(\sum_{h_1=1}^{H}\boldsymbol{\varepsilon}_{h_1,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h_1,c}^T\sum_{h_2=1}^{H}\boldsymbol{\varepsilon}_{h_2,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h_2,c}^T\right)\right]^2\\ &\quad= \sum_{h_1=1}^{H}\sum_{h_2=1}^{H}\sum_{h_3=1}^{H}\sum_{h_4=1}^{H}E\big[\operatorname{tr}\big(\boldsymbol{\varepsilon}_{h_1,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h_1,c}^T\boldsymbol{\varepsilon}_{h_2,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h_2,c}^T\big)\\ &\qquad\qquad\times\operatorname{tr}\big(\boldsymbol{\varepsilon}_{h_3,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h_3,c}^T\boldsymbol{\varepsilon}_{h_4,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h_4,c}^T\big)\big].\end{aligned}$$

Together with the special structure of $\mathbf{Q}_c'$ and the conditional independence of $\varepsilon_{(i)}$ given the $y_{(i)}$'s, we know that $\mathrm{E}\{\operatorname{tr}(\boldsymbol{\varepsilon}_{h_i,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h_i,c}^T\}=0$ for $i=1,2,3,4$. Then it is easy to derive that if $h_1\ne h_j$, where $j=2,3,4$,

$$E\operatorname{tr}\big(\boldsymbol{\varepsilon}_{h_1,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h_1,c}^T\boldsymbol{\varepsilon}_{h_2,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h_2,c}^T\big)\operatorname{tr}\big(\boldsymbol{\varepsilon}_{h_3,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h_3,c}^T\boldsymbol{\varepsilon}_{h_4,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h_4,c}^T\big)=0.$$

For the other cases, invoking the conditions of Theorem 1 and the conditional independence of the $\varepsilon_{(i)}$'s given the $y_{(i)}$'s, we have that for some $b>0$,

$$\begin{aligned}&\big|E\big[\operatorname{tr}\big(\boldsymbol{\varepsilon}_{h_1,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h_1,c}^T\boldsymbol{\varepsilon}_{h_2,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h_2,c}^T\big)\operatorname{tr}\big(\boldsymbol{\varepsilon}_{h_3,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h_3,c}^T\boldsymbol{\varepsilon}_{h_4,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h_4,c}^T\big)\big]\big|\\ &\qquad\le p^4c^8E\big[\big(|\varepsilon_i(j)|^4\big)^2\big]\le bp^4c^8.\end{aligned}$$

Thus

$$E\left(\left\|\sum_{h=1}^{H}\boldsymbol{\varepsilon}_{h,c}\mathbf{Q}_c'\boldsymbol{\varepsilon}_{h,c}^T\right\|^4\right)\le 3bH^2p^4c^8 = 3bn^2p^4c^6,$$

and then, for any $\eta>0$,

$$P(\|\mathbf{K}_2\|>\eta)\le\frac{E(\|\mathbf{K}_2\|^4)}{\eta^4}\le\frac{3bn^2p^4c^6}{n^4(c-1)^4\eta^4} = O(n^{-2}p^4/\eta^4).$$

Choosing $\eta = n^{-1/4}p(\log n)^{1/2}$, the Borel–Cantelli lemma implies the conclusion.

d. $\|\mathbf{K}_1-\boldsymbol{\Lambda}_p\| = O(n^{-1/2}p(\log n))$ a.s. as $n\to\infty$.

Note that $\boldsymbol{\Lambda}_p$ is the expectation of $\mathbf{K}_1$. This means that $\mathbf{K}_1-\boldsymbol{\Lambda}_p$ is a centered sample mean. Each of its elements is a sum of the independent identically distributed variables with mean 0 and finite second moment. Because the fourth moment of X is finite, invoking the result of theorem IX 20 of Petrov (1995, p. 278), every element of $\|\mathbf{K}_1-\boldsymbol{\Lambda}_p\|$ has order $O(n^{-1/2}\log n)$ a.s. Together with the definition of the Euclidean norm, the resulting convergence rate is $pn^{-1/2}\log n$. It can be immediately derived that $\|\mathbf{K}_1-\boldsymbol{\Lambda}_p\| = O(n^{-1/2}p\log n)$ a.s. as $n\to\infty$.

Together with steps a–d, and $r<1/4$, we complete the proof of the strong convergence of $\|\widehat{\boldsymbol{\Lambda}}_p-\boldsymbol{\Lambda}_p\|$.

As for the convergence of the term $\|(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}-\widehat{\boldsymbol{\Lambda}}_p)-(\boldsymbol{\Sigma}_{\mathbf{x}}-\boldsymbol{\Lambda}_p)\|$, we need only note that $\|\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}-\boldsymbol{\Sigma}_{\mathbf{x}}\|$ can have the same convergence rate as $\|\widehat{\boldsymbol{\Lambda}}_p-\boldsymbol{\Lambda}_p\|$. The proof is similar to step d, as follows. Consider every element of $\boldsymbol{\Sigma}_{\mathbf{x}}-\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}$. Without loss of generality, we consider the left-upper element on the diagonal. We know that the left-upper element, $\sigma_{1,1}$, of $\boldsymbol{\Sigma}_{\mathbf{x}}$ can be rewritten as the difference of the second moment and the square of the first moment of X, and similarly for $\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}$. As the lth moment of X is finite, applying the result of Petrov (1995) to the difference relating to the second-moment and first-moment terms, we can easily obtain the convergence rate $O(n^{-1/2}(\log n))$. Together with the definition of the Euclidean norm again, the resulting convergence rate is $O(pn^{-1/2}(\log n))$.

The Proof for Weak Convergence. We now turn to the proof of the weak convergence. From the lemmas, we have that in step a we use the rate for the convergence in probability to obtain $\|\mathbf{J}_2\| = O_P(n^{-1+1/l+r}p^{3/2})$, in step b we use the exactly the same argument to obtain $\|\mathbf{J}_3\| = O_p(pn^{-1+2r})$, in step c we can choose $\eta = n^{1/2}p$ with no use of the Borel–Cantelli lemma to achieve the rate $\|\mathbf{K}_2\| = O_p(pn^{-1/2})$, and in step d a standard argument for weak convergence is adopted to arrive at $\|\mathbf{K}_1-\boldsymbol{\Lambda}_p\| = (pn^{-1/2})$ (see Petrov 1995). When $p = o(n^{\min\{1/2,1-2/l-2r\}})$, the desired convergence rate can be achieved. The weak convergence of $\|(\widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}-\widehat{\boldsymbol{\Lambda}}_p)-(\boldsymbol{\Sigma}_{\mathbf{x}}-\boldsymbol{\Lambda}_p)\|$ can be proved similarly. The proof of Theorem 1 is finished.

A.4 Proof of Corollary 1

Recall the definition of λ_{pi} before Corollary 1. Note that $\lambda_{p1}\ge\lambda_{p2}\ge\cdots\ge\lambda_{pK}>\lambda_{p(K+1)}=\cdots=\lambda_{pp}=0$. It is clear that

$$\sum_{k=1}^{p}|\lambda_{pk}-\widehat{\lambda}_{pk}|^2\le\|\boldsymbol{\Lambda}_p-\widehat{\boldsymbol{\Lambda}}_p\|^2.$$

The convergence of $\widehat{\lambda}_{pk}$ can be derived from Theorem 1 immediately.

For the convergence of the estimate $P_{\widehat{\lambda}_{pi}}$ of the projection space $P_{\lambda_{pi}}$, because the nonzero eigenvalues are distinct, the projection spaces are just those of eigenvectors. For brevity, let $\mathbf{M}_p = \boldsymbol{\Sigma}_{\mathbf{x}}-\boldsymbol{\Lambda}_p$ and $\widetilde{\mathbf{M}}_{pn} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{x}}-\widehat{\boldsymbol{\Lambda}}_p$. Therefore, invoking Theorem 1, Corollary 1, and the representation of $P_{\widehat{\lambda}_{pi}}$ of Zhu and Ng (1995) for fixed p, the conclusion is reached immediately. We omit the details of the proof here.

A.5 Proof of Theorem 2

We deal with the strong convergence first. The proof for the weak convergence is similar. Throughout the proof, the argument is with

probability 1 unless stated otherwise. Let K be the true value of the dimension of $\mathbf{\Psi} = \mathbf{\Sigma}_{\mathbf{x}} - \mathbf{\Lambda}_p$. Note that

$$G(K) - G(k) = \log L_K - \log L_k - \frac{C_n(K-k)(2p-k-K+1)}{2}.$$

From Corollary 1, we have that for large n,

$$\widehat{\theta}_i > 1, \qquad i = 1, \ldots, K, \qquad \text{and} \qquad \min(\tau, K) = K,$$

where τ is the number of $\widehat{\theta}_i$ with $\widehat{\theta}_i > 1$. If $k < K$, then $\min(\tau, k) = k$. Thus, for large n,

$$\log L_K - \log L_k = -\frac{1}{2} n \sum_{i=k+1}^{K} (\log \widehat{\theta}_i + 1 - \widehat{\theta}_i) = \frac{1}{2} n W_n(K, k),$$

where

$$W_n(K, k) = -\sum_{i=k+1}^{K} (\log \widehat{\theta}_i + 1 - \widehat{\theta}_i).$$

We have that

$$\lim_{n\to\infty} W_n(K, k) = W(K, k) \equiv -\sum_{i=k+1}^{K} (\log \theta_i + 1 - \theta_i) > 0.$$

Hence we have that for large n,

$$\log L_K - \log L_k > \frac{1}{4} n W(K, k).$$

Note that $\lim_{n\to\infty} C_n / n^{1-s} = 0$ and $p = O(n^s)$. Then

$$\frac{C_n(K-k)(2p-k-K+1)}{n} \to 0. \tag{A.6}$$

Therefore, recalling the formula of $G(K) - G(k)$, for large n, we have that

$$G(K) - G(k) > 0. \tag{A.7}$$

If $k > K$, then we have, together with (10),

$$|\log L_K - \log L_k| \le n \sum_{i=K+1}^{p} |\log \widehat{\theta}_i + 1 - \widehat{\theta}_i|.$$

Invoking the Taylor expansion, we derive that

$$\begin{aligned} |\log L_K - \log L_k| &\le n \sum_{i=K+1}^{p} \frac{1}{2} (\widehat{\theta}_i - 1)^2 (1 + o(1)) \\ &\le n \|\mathbf{\Omega} - \widehat{\mathbf{\Omega}}\|^2 = O(n^{1-2t}) \quad \text{a.s.} \end{aligned}$$

Because $\lim_{n\to\infty} C_n / n^{1-2t} = \infty$, we can see that for large n,

$$G(K) - G(k) = O(n^{1-2t}) + \frac{C_n(k-K)(2p-k-K+1)}{2} > 0. \tag{A.8}$$

From (A.7) and (A.8), it follows that for large n,

$$\widehat{K} = K.$$

Thus the strong consistency of the estimate is proved.

[Received January 2004. Revised July 2005.]

REFERENCES

Antoniadis, A., Lambert-Lacroix, S., and Leblanc, F. (2003), "Effective Dimension Reduction Methods for Tumor Classification Using Gene Expression Data," *Bioinformatics*, 19, 563–570.

Bai, Z. D. (1999), "Methodologies in Spectral Analysis of Large-Dimensional Random Matrices, a Review," *Statistica Sinica*, 9, 611–677.

Bai, Z. D., Krishnaiah, R. D., and Zhao, L. C. (1989), "On Rates of Convergence of Efficient Detection Criteria in Signal Processing With White Noise," *IEEE Transactions on Information Theory*, 35, 380–388.

Bai, Z. D., and Saranadasa, H. (1996), "Effect of High Dimension by an Example of a Two-Sample Problem," *Statistica Sinica*, 6, 311–329.

Bickel, P. J., and Levina, E. (2004), "Some Theory for Fisher's Discriminant Function, 'Naive Bayes,' and Some Alternatives When There Are Many More Variables Than Observations," *Bernoulli*, 10, 989–1010.

Bura, E., and Cook, R. D. (2001a), "Extending Sliced Inverse Regression: The Weighted Chi-Squared Test," *Journal of the American Statistical Association*, 96, 996–1003.

——— (2001b), "Estimating the Structural Dimension of Regressions via Parametric Inverse Regression," *Journal of the Royal Statistical Society*, Ser. B, 63, 393–410.

Bura, E., and Pfeiffer, R. M. (2003), "Graphical Methods for Class Prediction Using Dimension Reduction Techniques on DNA Microarray Data," *Bioinformatics*, 19, 1252–1258.

Carroll, R. J., and Li, K. C. (1995), "Binary Regressors in Dimension Reduction Models: A New Look at Treatment Comparisons," *Statistica Sinica*, 5, 667–688.

Chiaromonte, F., and Cook, R. D. (1997), "On Foundations of Regression Graphics," Technical Report 616, University of Minnesota, School of Statistics.

Chiaromonte, F., Cook, R. D., and Li, B. (2002), "Sufficient Dimensions Reduction in Regressions With Categorical Predictors," *The Annals of Statistics*, 30, 475–497.

Chiaromonte, F., and Martinelli, J. (2002), "Dimension Reduction Strategies for Analyzing Global Gene Expression Data With a Response," *Mathematical Biosciences*, 176, 123–144.

Cook, R. D. (1994), "On the Interpretation of Regression Plots," *Journal of the American Statistical Association*, 89, 177–189.

——— (1996), "Graphics for Regressions With a Binary Response," *Journal of the American Statistical Association*, 91, 983–992.

——— (1998), *Regression Graphics: Ideas for Studying Regressions Through Graphics*, New York: Wiley.

——— (2000), "SAVE: A Method for Dimension Reduction and Graphics in Regression," *Communications in Statistics, Part A—Theory and Methods*, 29, 2109–2121.

Cook, R. D., and Lee, H. (1999), "Dimension Reduction in Binary Response Regression," *Journal of the American Statistical Association*, 94, 1187–1200.

Cook, R. D., and Li, B. (2002), "Dimension Reduction for Conditional Mean in Regression," *The Annals of Statistics*, 30, 455–474.

Cook, R. D., and Nachtsheim, C. J. (1994), "Reweighting to Achieve Elliptically Contoured Covariates in Regression," *Journal of the American Statistical Association*, 89, 592–599.

Cook, R. D., and Weisberg, S. (1991), Comment on "Sliced Inverse Regression for Dimension Reduction," by K. C. Li, *Journal of the American Statistical Association*, 86, 328–332.

Eubank, R. L. (1999), *Nonparametric Regression and Spline Smoothing* (2nd ed.), New York: Marcel Dekker.

Ferré, L. (1998), "Determining the Dimension in Sliced Inverse Regression and Related Methods," *Journal of the American Statistical Association*, 93, 132–140.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesorov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular Classification and Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537.

Greenshtain, E., and Ritov, Y. (2004), "Persistence in High-Dimensional Linear Predictor Selection and the Virtue of Overparametrization," *Bernoulli*, 10, 971–988.

Hall, P., and Li, C. K. (1993), "On Almost Linearity of Low-Dimensional Projections From High-Dimensional Data," *The Annals of Statistics*, 21, 867–889.

Hooper, J. (1959), "Simultaneous Equations and Canonical Correlation Theory," *Econometrica*, 27, 245–256.

Hsing, T., and Carroll, R. J. (1992), "An Asymptotic Theory for Sliced Inverse Regression," *The Annals of Statistics*, 20, 1040–1061.

Levina, E. (2002), "Statistical Issues in Texture Analysis," unpublished doctoral thesis, University of California Berkeley, California.

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction" (with discussions), *Journal of the American Statistical Association*, 86, 316–342.

——— (1992), "On Principal Hessian Directions for Data Visualization and Dimension Reduction," *Journal of the American Statistical Association*, 87, 1025–1040.

Li, Y. X., and Zhu, L. X. (2004), "When Are Sliced Average Variance Estimates Convergent?" technical report, University of Hong Kong, Dept. of Statistics and Actuarial Science.

Petrov, V. V. (1995), *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*, New York: Oxford University Press.

Pollard, D. (1984), *Convergence of Stochastic Processes*, New York: Springer-Verlag.

Schott, J. R. (1994), "Determining the Dimensionality in Sliced Inverse Regression," *Journal of the American Statistical Association*, 89, 141–148.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Mathematical Statistics*, 30, 461–464.

Velilla, S. (1998), "Assessing the Number of Linear Components in a General Regression Problem," *Journal of the American Statistical Association*, 93, 1088–1098.

Xia, Y. C., Tong, H., Li, W. K., and Zhu, L. X. (2002), "An Adaptive Estimation of Dimension Reduction Space," *Journal of the Royal Statistical Society*, Ser. B, 64, 363–410.

Yang, S. S. (1977), "General Distribution Theory of the Concomitants of Order Statistics," *The Annals of Statistics*, 5, 996–1002.

Yin, X. R., and Cook, R. D. (2002), "Dimension Reduction for the Conditional *k*th Moment in Regression," *Journal of the Royal Statistical Society*, Ser. B, 64, 159–175.

Zhao, L. C., Krishnaiah, P. R., and Bai, Z. D. (1986a), "On Detection of the Number of Signals in Presence of White Noise," *Journal of Multivariate Analysis* 20, 1–25.

——— (1986b), "On Detection of the Number of Signals When the Noise Covariance Matrix Is Arbitrary," *Journal of Multivariate Analysis* 20, 26–49.

Zhu, L. X., and Fang, K. T. (1996), "Asymptotics for Kernel Estimate of Sliced Inverse Regression, *The Annals of Statistics*, 24, 1055–1068.

Zhu, L. X., and Ng, K. W. (1995), "Asymptotics of Sliced Inverse Regression," *Statistica Sinica*, 5, 727–736.

Zhu, L. X., Ohtaki, M., and Li, Y. X. (2006), "Extensions of Sliced Inverse Regression–Based Algorithms," technical report, University of Hong Kong, Dept. of Statistics and Actuarial Science.

Zhu, L. X., Zhu, L. P., and Li, X. (2006), "Transformed Partial Least Squares for Multivariate Data," *Statistica Sinica*, accepted.