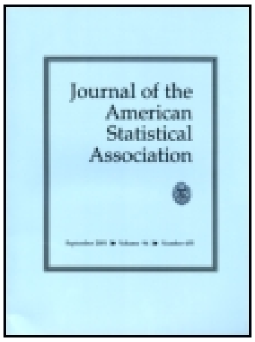

Journal of the American Statistical Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

On Directional Regression for Dimension Reduction

Bing Li & Shaoli Wang

To cite this article: Bing Li & Shaoli Wang (2007) On Directional Regression for Dimension Reduction, Journal of the American Statistical Association, 102:479, 997-1008, DOI: 10.1198/016214507000000536

To link to this article: http://dx.doi.org/10.1198/016214507000000536

On Directional Regression for Dimension Reduction

Bing LI and Shaoli WANG

We introduce directional regression (DR) as a method for dimension reduction. Like contour regression, DR is derived from empirical directions, but achieves higher accuracy and requires substantially less computation. DR naturally synthesizes the dimension reduction estimators based on conditional moments, such as sliced inverse regression and sliced average variance estimation, and in doing so combines the advantages of these methods. Under mild conditions, it provides exhaustive and $\sqrt{n}$-consistent estimate of the dimension reduction space. We develop the asymptotic distribution of the DR estimator, and from that a sequential test procedure to determine the dimension of the central space. We compare the performance of DR with that of existing methods by simulation and find strong evidence of its advantage over a wide range of models. Finally, we apply DR to analyze a data set concerning the identification of hand-written digits.

KEY WORDS: Contour regression; Exhaustive estimation; Efficiency; Sliced inverse regression; Sliced average variance estimation.

1. INTRODUCTION

Dimension reduction for regression, as pioneered by such authors as Li and Duan (1989), Duan and Li (1991), Li (1991, 1992), Cook and Weisberg (1991), and Cook (1994, 1996), is aimed at reducing the dimension of a vector-valued predictor X, while preserving its regression relation with a real-valued response Y. Research into dimension reduction has gained considerable momentum in recent years due to the rapidly increasing data volume and dimension, which demand preprocessing techniques to reduce their scope.

Let X be a p-dimensional random vector representing the predictor, and let Y be a random variable representing the response. Dimension reduction seeks a set of linear combinations of X, say $\beta^T X$, where β is a $p \times q$ matrix with $q \leq p$, such that Y depends on X only through these linear combinations. Mathematically, this can be formulated as $Y \perp\!\!\!\perp X | \beta^T X$; that is, Y and X are independent conditioning on $\beta^T X$. If β satisfies this relation, then its column space is called a dimension reduction space. Under mild assumptions, the intersection of dimension reduction spaces is itself a dimension reduction space (Cook 1998, sec. 6.3); in this case we call the intersection of all dimension reduction spaces the central space, written as $\mathcal{S}_{Y|X}$. One of the main objectives of dimension reduction is the statistical inference of the central space.

Sliced inverse regression (SIR; see Li 1991) and sliced average variance estimator (SAVE; see Cook and Weisberg 1991), which are based on the first two conditional moments $E(X|Y)$ and $E(XX^T|Y)$, are among the most commonly used dimension reduction estimators. They have well-known limitations, however. In particular, SIR is known to fail when the response surface is symmetric about the origin, whereas SAVE is not very efficient in estimating monotone trends for small to moderate sample sizes. Other classical estimators, such as the principal Hessian directions (PHD; see Li 1992) and SIRII (see Li 1991) also have similar difficulties. To overcome these difficulties, several authors have proposed combining various forms of first two conditional moments by convex combination, such as the convex combinations of SIR and SIRII, of SIR and PHD, and of SIR and SAVE (see Li 1991; Gannoun and Saracco 2003; Ye and Weiss 2003; Zhu, Ohtaki, and Li 2005). Furthermore, Ye and Weiss (2003) proposed a bootstrap method that can be used to select the coefficient in the convex combination methods.

In this article we introduce a natural and simple principle for dimension reduction, called *directional regression* (DR), that synthesizes the dimension reduction methods based on first two conditional moments and achieves substantial improvement in accuracy. The idea is derived from empirical directions. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample of (X, Y). Li, Zha, and Chiaromonte (2005) introduced empirical directions as the set of vectors $\{X_i - X_j : 1 \leq i < j \leq n\}$ and argued that they contain all of the directional information about predictor X that is available in the sample. Based on this consideration, they introduced simple contour regression (SCR) and general contour regression (GCR), which estimate the central space by extracting the empirical directions along which Y varies the least. Contour regressions mitigate the difficulties of the classical methods considerably.

DR makes more efficient use of the empirical directions by regressing them directly on the responses in the L_2 sense. This results in a $\sqrt{n}$-consistent and exhaustive estimator of the central space. The advantages of DR are its high accuracy, short computing time, and natural combination of first two inverse conditional moments. We provide strong evidence for these advantages by comparing its accuracy and computing time with those of a wide range of existing dimension-reduction estimators, and by performing a theoretical analysis of its structure.

The notion of empirical directions is related to several previous works in nonparametric multivariate analysis, which used similar reasoning to introduce interdirections and multivariate signs. For any pair of points in the sample, say X_i and X_j, the *interdirection* between them is defined (roughly) as the number of hyperplanes, passing through the origin and any $p-1$ out of the remaining $n-2$ points, that intersect the line segment connecting X_i and X_j. Here the origin is the hypothesized center of the distribution of X. The $\binom{n}{2}$ interdirections can be used to estimate the angles between any two points, leading to a multivariate and affine-invariant generalization of the sign test (see, e.g., Randles 1989; Hettmansperger and Oja 1994; Oja 1999; Hallin and Paindaveine 2002, 2005). Although similarly motivated, empirical directions and interdirections have different definitions and emphases. Technically, the former are vectors that are invariant under translation, whereas the latter are scalars

Bing Li is Professor, Department of Statistics, Pennsylvania State University, University Park, PA 16802 (E-mail: *bing@stat.psu.edu*). Shaoli Wang is Postdoctoral Researcher, Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520. This work was supported in part by National Science Foundation grants DMS-0204662 and DMS-0405681. The authors thank two anonymous referees, an associate editor, and the joint editors for their many useful suggestions that helped us improve both the presentation and the substance of this article.

Journal of the American Statistical Association
September 2007, Vol. 102, No. 479, Theory and Methods
DOI 10.1198/016214507000000536

defined in reference to a location. Moreover, empirical directions are used in conjunction with a response Y, to determine the directions in X that affect Y, whereas the interdirections are used to characterize the distribution of X, with emphasis on testing a multivariate location parameter.

In Section 2 we introduce DR at the population level, derive its various expressions in relation to SIR and SAVE, and demonstrate its consistency. In Section 3 we derive sufficient conditions for DR to be exhaustive, that is, to cover the entire central space. In Section 4 we develop the estimation procedure and an asymptotic sequential test for order determination. We compare the performance of the DR estimator with that of the existing estimators by simulation in Section 5. In Section 6 we apply the DR estimator to analyze a data set concerning the identification of hand-written digits. All proofs are relegated to the Appendix.

2. DIRECTIONAL REGRESSION

Let $\mu = E(X)$, $\Sigma = \mathrm{var}(X)$, and Z be the standardized predictor $\Sigma^{-1/2}(X - \mu)$. We work with the standardized predictor Z in Sections 2 and 3 and transform the result into the X-scale in Section 4.

Let $(\widetilde{Z}, \widetilde{Y})$ be an independent copy of (Z, Y), and let

$$A(Y, \widetilde{Y}) = E[(Z - \widetilde{Z})(Z - \widetilde{Z})^T | Y, \widetilde{Y}]. \qquad (1)$$

This is regressing the direction $Z - \widetilde{Z}$, as represented by $(Z - \widetilde{Z})(Z - \widetilde{Z})^T$, onto the space of functions of $(Y, \widetilde{Y})$ that are squared integrable with respect to $L_2(P_Y)$. Intuitively, those directions $Z - \widetilde{Z}$ that are aligned with $\mathcal{S}_{Y|Z}$ are significantly affected by Y, but those aligned with $\mathcal{S}_{Y|Z}^{\perp}$ are not. It is this change of behavior of $Z - \widetilde{Z}$ in accordance with its alignment with $\mathcal{S}_{Y|Z}$ that allows regression (1) to identify $\mathcal{S}_{Y|Z}$.

Let P be the projection onto $\mathcal{S}_{Y|Z}$ with respect to the inner product $\langle a, b\rangle = a^T b$, and let $Q = I_p - P$. The next theorem asserts that under appropriate conditions, the column space of $2I_p - A(Y, \widetilde{Y})$ is contained in $\mathcal{S}_{Y|Z}$ for any given $(Y, \widetilde{Y}) = (y, \tilde{y})$.

Theorem 1. Suppose, for any $v \in \mathbb{R}^p$ and $v \perp \mathcal{S}_{Y|Z}$, that (a) $E(v^T Z | PZ)$ is a linear function of Z, and (b) $\mathrm{var}(v^T Z | PZ)$ is a nonrandom number. Then, for any $(Y, \widetilde{Y})$, the column space of $2I_p - A(Y, \widetilde{Y})$ is contained in $\mathcal{S}_{Y|Z}$.

Conditions such as those assumed in Theorem 1 are well known; a stronger form of them is often used as sufficient conditions for the consistency of SAVE and PHD. They are not as restrictive as they seem. Let $\eta_1, \ldots, \eta_k$ be iid random vectors uniformly distributed over the unit sphere in $\mathbb{R}^p$, let $\eta = (\eta_1, \ldots, \eta_k)$, and assume that $\eta \perp\!\!\!\perp (Z, Y)$. Using the results of Diaconis and Freedman (1984) and Hall and Li (1993), it can be shown that as $p \to \infty$, the (random) characteristic function of the conditional distribution of $\eta^T Z | \eta$ converges in probability to a normal characteristic function. Intuitively, this means that as $p \to \infty$, the proportion of the η's for which $\eta^T Z | \eta$ converges to normal tends to 1. (This is called "weak convergence in probability" in Diaconis and Freedman 1984.) In our context, $E(v^T Z | PZ)$ and $\mathrm{var}(v^T Z | PZ)$ can be written equivalently as $E(v^T Z | \beta^T Z)$ and $E(v^T Z | \beta^T X)$, where β is a $p \times q$ matrix with columns that form an orthonormal bases of $\mathcal{S}_{Y|Z}$. Then the matrix (v, β) plays the role of η and due to the asymptotic normality just described, conditions (a) and (b) of Theorem 1 holds approximately when p is large. Note that this is different than saying that Z itself tends to multivariate normal as $p \to \infty$. In fact, this need not be true (see Portnoy 1986). It is the asymptotic normality of the linear combinations of Z that is needed for the conditions of Theorem 1 to hold approximately. (For further discussions on this subject, see Li and Yin 2006.)

We can also satisfy these conditions approximately by transforming or reweighting the original predictor (see, e.g., Cook and Nachtsheim 1994).

Theorem 1 suggests the object

$$G = E[2I_p - A(Y, \widetilde{Y})]^2 \qquad (2)$$

as (the population version of) the estimate of $\mathcal{S}_{Y|Z}$. We can disentangle the (Z, Y) and $(\widetilde{Z}, \widetilde{Y})$ in G to express it as a nonlinear functional of conditional moments of Z given Y, whose estimation requires only $O(n)$ operations, compared with $O(n^2)$ operations required by SCR and $O(n^3)$ operations required by GCR. As we show in Sections 5 and 6, this saving of computing time is practically important for moderate to large data sets. This disentangled expression also simplifies the asymptotic development considerably because, unlike contour regressions, it no longer involves U-statistics. The next theorem gives the reexpression of G.

Theorem 2. The matrix G can be reexpressed as

$$\begin{aligned} G = {} & 2E[E^2(ZZ^T|Y)] + 2E^2\big[E(Z|Y)E(Z^T|Y)\big] \\ & + 2E\big[E(Z^T|Y)E(Z|Y)\big]E\big[E(Z|Y)E(Z^T|Y)\big] \\ & - 2I_p. \end{aligned} \qquad (3)$$

We denote the column space of G by $\mathrm{span}(G)$. The following two expressions of G also are informative. Because $E(ZZ^T) = I_p$, we can rewrite (3) as

$$\begin{aligned} G = {} & 2E[E^2(ZZ^T - I_p|Y)] + 2E^2\big[E(Z|Y)E(Z^T|Y)\big] \\ & + 2E\big[E(Z^T|Y)E(Z|Y)\big]E\big[E(Z|Y)E(Z^T|Y)\big]. \end{aligned} \qquad (4)$$

Furthermore, because $E(ZZ^T|Y) = \mathrm{var}(Z|Y) + E(Z|Y)E(Z^T|Y)$, we can rewrite the foregoing as

$$\begin{aligned} G = {} & 2E[\mathrm{var}(Z|Y) - I_p]^2 \\ & + 2E\big[(\mathrm{var}(Z|Y) - I_p)E(Z|Y)E(Z^T|Y)\big] \\ & + 2E\big[E(Z|Y)E(Z^T|Y)(\mathrm{var}(Z|Y) - I_p)\big] \\ & + 2E\big[E(Z|Y)E(Z^T|Y)\big]^2 \\ & + 2E^2\big[E(Z|Y)E(Z^T|Y)\big] \\ & + 2E\big[E(Z^T|Y)E(Z|Y)\big]E\big[E(Z|Y)E(Z^T|Y)\big]. \end{aligned} \qquad (5)$$

This last expression of G is akin to a general class of dimension-reduction estimators considered by Ye and Weiss (2003), who demonstrated, under (stronger forms of) (a) and (b) of Theorem 1, that the column space of any matrix of the form

$$\sum_{i+j\ge 1} E\big(\alpha_{ij}(Y)\big[E(Z|Y)E(Z^T|Y)\big]^i[E(ZZ^T|Y) - I_p]^j\big),$$

where $\alpha_{ij}(Y)$ are real-valued function of Y, is contained in $\mathcal{S}_{Y|Z}$. Strictly speaking, however, G does not belong to the Ye–Weiss class; for example, the fifth term in (5) cannot be accommodated by the foregoing form. But if we extend the Ye–Weiss class slightly, it will accommodate G. Let $\mathcal{C}$ denote the Ye–Weiss class, and let

$$\mathcal{C}^* = \left\{ \sum_{i=1}^{\ell} C_{i1} \cdots C_{ik_i} : C_{ij} \in \mathcal{C}, \right.$$

$$\left. i = 1, \ldots, \ell, j = 1, \ldots, k_\ell, \ell, k_\ell = 1, 2, \ldots \right\}.$$

Then G becomes a member of $\mathcal{C}^*$.

As with SIR and SAVE, in practice we use the sample estimate of the discretized version of G to estimate $\mathcal{S}_{Y|Z}$. Let Ω_Y be the sample space of Y and let $\{J_1, \ldots, J_m\}$ be a partition of Ω_Y. Let $p_k = P(Y \in J_k)$. Then G is discretized as

$$\begin{aligned} F = 2 \sum E^2(ZZ^T - I_p | Y \in J_k) p_k \\ + 2\Big[\sum E(Z|Y \in J_k) E(Z^T|Y \in J_k) p_k\Big]^2 \\ + 2 \sum E(Z^T|Y \in J_k) E(Z|Y \in J_k) p_k \\ \times \sum E(Z|Y \in J_k) E(Z^T|Y \in J_k) p_k. \end{aligned} \tag{6}$$

Of course, F depends on the partition $(J_1, \ldots, J_m)$, but for simplicity we omit this dependence from the notation. Theorem 1 also implies $\text{span}(F) \subset \mathcal{S}_{Y|Z}$ because, if we let $Y^* = \sum k I(Y \in J_k)$ be the discretized Y, then Theorem 1 applies to Y^*.

3. CONDITIONS FOR EXHAUSTIVE ESTIMATION

Typically, a dimension-reduction method estimates a subspace $\mathcal{S}$ of $\mathcal{S}_{Y|Z}$. If it happens that $\mathcal{S} = \mathcal{S}_{Y|Z}$, then we say the method is exhaustive. Many known dimension-reduction methods, including ordinary least square (OLS), SIR, and PHD, are not exhaustive. Currently, the only known sufficient condition for SAVE to be exhaustive is that conditional distribution of Z given Y is multivariate normal, which is quite restrictive (see Cook and Lee 1999). Li et al. (2005) established the exhaustiveness of contour regressions under fairly mild assumptions. We now present two equivalent sufficient conditions for $\mathcal{S}_{\text{DR}}$ to be exhaustive, which we deem quite mild.

Theorem 3. Under assumptions (a) and (b) of Theorem 1, the following statements are equivalent:

a. For any $v \in \mathcal{S}_{Y|Z}$, $v \neq 0$, the random variable $E[(v^T(Z - \widetilde{Z}))^2 | Y, \widetilde{Y}]$ is nondegenerate; that is, it is not equal almost surely to a constant.

b. For any $v \in \mathcal{S}_{Y|Z}$, $v \neq 0$, at least one of the random variables $E(v^T Z|Y)$ and $E((v^T Z)^2|Y)$ is nondegenerate.

Moreover, they imply that $\mathcal{S}_{\text{DR}} = \mathcal{S}_{Y|Z}$.

Statement a is a very mild requirement; if we let P_v be the orthogonal projection onto $\text{span}(v)$, then statement a amounts to requiring $E[\|P_v(Z - \widetilde{Z})\|^2 | Y, \widetilde{Y}]$ to be nondegenerate for all $v \in \mathcal{S}_{Y|Z}$, $v \neq 0$. This is reasonable because, by the definition of the central space $\mathcal{S}_{Y|Z}$, Y is genuinely dependent on $v^T Z$. Thus statement a can be roughly interpreted as

> The squared length of the v component of the increment
> $Z - \widetilde{Z}$ should vary with $(Y, \widetilde{Y})$ whenever Y is genuinely
> dependent on the v component of Z.

Theorem 3 also sheds new light onto the exhaustiveness of SAVE, which is based on the sample estimate of the matrix $E[\text{var}(Z|Y) - I_p]^2$. We denote the column space of this matrix by $\mathcal{S}_{\text{SAVE}}$.

Theorem 4. Suppose that the moments involved in the definition of SAVE and DR are finite. Then $\mathcal{S}_{\text{DR}} = \mathcal{S}_{\text{SAVE}}$.

Li et al. (2005) speculated that the inaccuracy of SAVE for some response shapes is due to its lack of efficiency rather than its lack of comprehensiveness. This is confirmed by Theorem 4; the spaces spanned by (population versions of) SAVE and DR are the same, and thus, by Theorem 3, are exhaustive under mild conditions. The gain in accuracy by DR, as we repeatedly demonstrate in our simulation studies, must then come from its improved efficiency—most likely from its coordination of the first two conditional moments through the efficient arrangement of the empirical directions. From (5), we see that G includes the interaction between the first and second conditional moments (the second and third terms) and three different arrangements of the first conditional moments (the last three terms). These terms reflect the difference between arranging the predictor X according to Y and arranging the empirical directions according to Y and $\widetilde{Y}$.

Note that Theorem 4 holds under virtually no conditions; in particular, it does not require assumptions (a) and (b) of Theorem 1. In this connection, it is helpful to note that SAVE (and hence DR) is more comprehensive than SIR in the following sense. Let $\mathcal{S}_{\text{SIR}}$ be the column space of the SIR matrix $\text{var}(E(Z|Y))$. Then it is known that $\mathcal{S}_{\text{SIR}} \subset \mathcal{S}_{\text{SAVE}}$, under no conditions other than the existence of the moments involved. This result is from Ye and Weiss (2003); the sample version of this relation was shown earlier by Cook and Critchley (2000).

4. STATISTICAL INFERENCE

4.1 Estimation

In this section we develop the sample estimator of $\mathcal{S}_{\text{DR}}$, which equals $\mathcal{S}_{Y|Z}$ under the conditions of Theorem 3. We first treat the dimension q of $\mathcal{S}_{Y|Z}$ as known, and then tackle its determination in Section 4.2. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be an iid sample of (X, Y). Let

$$\widehat{Z}_i = \widehat{\Sigma}^{-1/2}(X_i - \hat{\mu}),$$

where

$$\hat{\mu} = E_n(X) \qquad \text{and} \qquad \widehat{\Sigma} = E_n[(X - \hat{\mu})(X - \hat{\mu})^T],$$

where $E_n(\cdot)$ indicates the sample average $n^{-1}\sum_{i=1}^n(\cdot)$. The original definition (2) of G suggests the following sample estimator of $\mathcal{S}_{Y|Z}$:

$$\binom{m}{2}^{-1} \sum_{k<\ell} [2I_p - \widehat{A}(J_k, J_\ell)]^2,$$

where

$$\widehat{A}(J_k, J_\ell) = \frac{\sum_{i<j} (\widehat{Z}_i - \widehat{Z}_j)(\widehat{Z}_i - \widehat{Z}_j)^T I(Y_i \in J_k, Y_j \in J_\ell)}{\sum_{i<j} I(Y_i \in J_k, Y_j \in J_\ell)}.$$

This expression directly reflects the idea of regressing the empirical directions onto $\{(Y_i, Y_j): i < j\}$ as explained in Section 2, but it requires $O(n^2)$ operations. For computational efficiency, it is better to use a sample estimator of expressions such as (6). Let

$$\begin{aligned}\widehat{F} &= 2\sum E_n^2(\widehat{Z}\widehat{Z}^T - I_p|Y \in J_k)\hat{p}_k \\ &\quad + 2\Big[\sum E_n(\widehat{Z}|Y \in J_k)E_n(\widehat{Z}^T|Y \in J_k)\hat{p}_k\Big]^2 \\ &\quad + 2\sum E_n(\widehat{Z}^T|Y \in J_k)E_n(\widehat{Z}|Y \in J_k)\hat{p}_k \\ &\quad \times \sum E_n(\widehat{Z}|Y \in J_k)E_n(\widehat{Z}^T|Y \in J_k)\hat{p}_k, \qquad (7)\end{aligned}$$

where the summation is over $k = 1, \dots, m$, $\hat{p}_k = E_n(Y \in J_k)$, and the notation such as $E_n(\widehat{Z}|Y \in J_k)$ stands for sample conditional moments, defined by

$$E_n(\widehat{Z}|Y \in J_k) = \frac{E_n[\widehat{Z}I(Y \in J_k)]}{E_n I(Y \in J_k)} = \frac{\sum_{i=1}^n \widehat{Z}_i I(Y_i \in J_k)}{\sum_{i=1}^n I(Y_i \in J_k)}.$$

Let $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_p$ be the eigenvalues of $\widehat{F}$ and let $\hat{v}_1, \dots, \hat{v}_p$ be the corresponding eigenvectors. We use the first q of these vectors (namely $\hat{v}_1, \dots, \hat{v}_q$) to estimate $\mathcal{S}_{Y|Z}$. We then use $\widehat{\Sigma}^{-1/2}\hat{v}_1, \dots, \widehat{\Sigma}^{-1/2}\hat{v}_q$ to estimate $\mathcal{S}_{Y|X}$. This last transformation (by $\widehat{\Sigma}^{-1/2}$) is derived from an invariance property of a dimension reduction space, as described by Cook (1998; prop. 6.3).

DR also applies when Y is discrete, in which case each value of Y can be used as a natural slice. As with other dimension reduction methods, such as SIR and SAVE, here we do not require that the length of J_i decreases with n. This is different from a typical kernel estimate in which the kernel size is required to decrease with n. It is this aspect that allows us to maintain $\sqrt{n}$ convergence rate regardless of the dimension of X, avoiding the curse of dimensionality.

4.2 Sequential Test for Order Determination

We now tackle the estimation of q by a sequential test. Sequential tests have been frequently used in the dimension reduction for order determination (see Li 1991, 1992; Cook 1998; Cook and Li 2004; Zhu and Fang 1996; Schott 1994).

By Theorem 3, for sufficiently fine partition $\{J_1, \dots, J_m\}$, $q = \text{rank}(F)$. Thus we estimate q by testing the following sequence of hypotheses:

$$H_0^{(\ell)}: \text{rank}(F) = \ell, \quad \text{for } \ell = 0, 1, \dots, p-1. \qquad (8)$$

The rank of F is estimated as the first ℓ at which the hypothesis is accepted; that is, $\hat{q} = \text{argmin}\{\ell: H_0^{(\ell)} \text{ is accepted}\}$ if $H_0^{(\ell)}$ is accepted for some $0 \le \ell \le p-1$, and $\hat{q} = p$ otherwise.

Evidently, it is equivalent to test (8) for any $p \times p$ matrix F_1 with the same rank as F. We choose an F_1 such that its sample estimator has a simplified asymptotic expansion. In particular, the choice will allow us to ignore the factors, such as $\widehat{\Sigma}^{-1}$ and $\widehat{\Sigma}^{-1/2}$, that do not affect the rank of interest but would add many terms in the asymptotic expansion. Let $A_0(Y, \widetilde{Y}) = E[(X - \widetilde{X})(X - \widetilde{X})^T|Y, \widetilde{Y}]$.

Proposition 1. Let $G_1 = E[2\Sigma - A_0(Y, \widetilde{Y})]^2$. Then G and G_1 have the same rank.

The proof is fairly simple and is omitted. Let F_1 be the discretized version of G_1, defined in analogy with the discretization of G. Because the foregoing proposition also applies to the discretized response Y^*, it implies that F and F_1 also have the same rank. Thus hypothesis (8) is equivalent to

$$H_0^{(\ell)}: \text{rank}(F_1) = \ell, \quad \text{for } \ell = 0, 1, \dots, p-1. \qquad (9)$$

We first express F_1 and its estimate $\widehat{F}_1$ in quadratic forms. Without loss of generality, assume that $E(X) = \mu = 0$. Let

$$\begin{aligned} U_k &= E(X|Y \in J_k) \qquad \text{and} \\ V_k &= E(XX^T|Y \in J_k) - \Sigma. \end{aligned} \qquad (10)$$

Following the proofs of Theorem 2 and (4), we can write F_1 as

$$\begin{aligned}F_1 &= 2\sum V_k V_k p_k + 2\Big(\sum U_k U_k^T p_k\Big)^2 \\ &\qquad + 2\Big(\sum U_k^T U_k p_k\Big)\Big(\sum U_k U_k^T p_k\Big).\end{aligned}$$

Thus $F_1 = HH^T$, where $H = (H_{11}, \dots, H_{1m}; H_2; H_{31}, \dots, H_{3m})$ and

$$\begin{aligned} H_{1k} &= \sqrt{2p_k}V_k, \\ H_2 &= \sqrt{2}\Big(\sum p_\ell U_\ell U_\ell^T\Big), \qquad \text{and} \\ H_{3k} &= \sqrt{2p_k}\Big(\sum U_\ell^T U_\ell p_\ell\Big)^{1/2} U_k. \end{aligned} \qquad (11)$$

Correspondingly, we define the sample estimate $\widehat{F}_1$ of F_1 as follows. Let

$$\widehat{U}_k = E_n(\widehat{X}|Y \in J_k) \qquad \text{and} \qquad \widehat{V}_k = E_n(\widehat{X}\widehat{X}^T - \widehat{\Sigma}|Y \in J_k),$$

where $\widehat{X} = X - \hat{\mu}$, and the sample conditional moments are as defined in Section 4. Then $\widehat{F}_1 = \widehat{H}\widehat{H}^T$, where $\widehat{H}$ is defined by substituting $\widehat{U}_k$, $\widehat{V}_k$, and $\hat{p}_k$ for U_k, V_k, and p_k in H.

Let $\hat{\lambda}_1 \ge \cdots \ge \hat{\lambda}_p$ be the singular values of $\widehat{H}$. We use

$$T_\ell = n \sum_{i=\ell+1}^{p} \hat{\lambda}_\ell^2$$

as the test statistic for hypothesis (9): Reject the hypothesis for large values of T_ℓ. To derive the asymptotic distribution of T_ℓ under $H_0^{(\ell)}$, let $mp + m + p = h$ and suppose that H, whose rank is ℓ under $H_0^{(\ell)}$, have the SVD

$$H = (\Gamma_1 \quad \Gamma_0)\begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \Psi_1^T \\ \Psi_0^T \end{pmatrix}, \qquad (12)$$

where (Γ_1, Γ_0) is a $p \times p$ orthogonal matrix in which Γ_1 and Γ_0 have dimensions $p \times \ell$ and $p \times (p - \ell)$, (Ψ_1, Ψ_0) is an $h \times h$ orthogonal matrix, in which Ψ_1 and Ψ_0 have dimensions $h \times \ell$ and $h \times (h - \ell)$, and D is an $\ell \times \ell$ diagonal matrix of positive diagonal elements. Following Eaton and Tyler (1994), under $H_0^{(\ell)}$, T_ℓ has the same asymptotic distribution as

$$\text{vec}^T[\sqrt{n}\Gamma_0^T(\widehat{H} - H)\Psi_0]\,\text{vec}[\sqrt{n}\Gamma_0^T(\widehat{H} - H)\Psi_0].$$

Here, for a matrix A with columns $a_1, \dots, a_p$, $\mathrm{vec}(A)$ denotes the vector $(a_1^T, \dots, a_p^T)^T$. Thus we need only derive the asymptotic distribution of $\sqrt{n}\,\Gamma_0^T(\widehat{H} - H)\Psi_0$.

The asymptotic expansions will have the following basic form. Let F_n be the empirical measure based on the iid sample $(X_1, Y_1), \dots, (X_n, Y_n)$. Let $\mathcal{F}$ be a convex set of distributions that includes empirical distributions and the true distribution F. Let S be a real- or matrix-valued functional on $\mathcal{F}$. Then, under regularity conditions,

$$S(F_n) = S(F) + E_n S^*(F) + O_p(n^{-1}), \tag{13}$$

where $S(F)$ is nonrandom, $S^*(F)$ is a random array with entries having finite variance, and $ES^*(F) = 0$. The random object $S^*(F)$ is called the *influence function* (see, e.g., Hampel 1974; Fernholz 1983; Serfling 1980, sec. 6.2; Bickel, Klaassen, Ritov, and Wellner 1993, p. 19). By the Lindeberg–Levy central limit theorem, $E_n S^*(F) = O_p(n^{-1/2})$. We first give the expansions of $\hat{p}_k$, $\widehat{U}_k$, and $\widehat{V}_k$.

Lemma 1. Suppose that $\mu = 0$, and that the entries of $E(X|Y)$ and $E(XX^T|Y)$ have finite second moments. Then the expansions of $\hat{p}_k$, $\widehat{U}_k$, and $\widehat{V}_k$ have the form (13) with p_k, U_k, or V_k in place of $S(F)$, p_k^*, U_k^*, or V_k^* in place of $S^*(F)$, where

$$p_k^* = I(Y \in J_k) - p_k, \qquad U_k^* = (X - U_k)I(Y \in J_k)/p_k - X,$$

and

$$V_k^* = (XX^T - \Sigma - V_k)I(Y \in J_k)/p_k \\ - U_k X^T - XU_k^T - XX^T + \Sigma.$$

Using these expansions, we now present the expansion of $\widehat{H}$.

Theorem 5. Under the assumptions of Lemma 1,

$$\widehat{H} = H + E_n(H^*) + O_p(n^{-1}),$$

where $H^* = (H_{11}^*, \dots, H_{1m}^*; H_2^*; H_{31}^*, \dots, H_{3m}^*)$, in which

$$H_{1k}^* = (2p_k)^{-1/2} p_k^* V_k + (2p_k)^{1/2} V_k^*,$$

$$H_2^* = \sqrt{2}\sum(p_\ell^* U_\ell U_\ell^T + p_\ell U_\ell^* U_\ell^T + p_\ell U_\ell U_\ell^{*T}),$$

and

$$H_{3k}^* = 2^{-3/4}[\mathrm{tr}(H_2)]^{-1/2}\mathrm{tr}(H_2^*) p_k^{1/2} U_k \\ + 2^{-3/4}[\mathrm{tr}(H_2)]^{1/2} p_k^{-1/2} p_k^* U_k + 2^{1/4}[\mathrm{tr}(H_2)]^{1/2} p_k^{1/2}.$$

The theorem implies that

$$\sqrt{n}\,\mathrm{vec}[\Gamma_0^T(\widehat{H} - H)\Psi_0] \xrightarrow{\mathcal{L}} \mathrm{N}(0, \Lambda), \\ \text{where } \Lambda = \mathrm{var}[\mathrm{vec}(\Gamma_0^T H^* \Psi_0)]. \tag{14}$$

Thus T_ℓ converges in distribution to

$$\sum_{i=1}^{(p-\ell)(mp+p+m-\ell)} \omega_i K_i,$$

where the ω_i's are the eigenvalues of Λ and the K_i's are iid χ_1^2 random variables.

Based on Theorem 5, the asymptotic distribution of T_ℓ can be approximated as follows. First, compute $\widehat{\Gamma}_0$, the $p \times (p - \ell)$ matrix whose columns are the eigenvectors of $\widehat{H}\widehat{H}^T$ corresponding to its $p - \ell$ smallest eigenvalues, and $\widehat{\Psi}_0$, the $h \times (h - \ell)$ matrix whose column vectors are the eigenvectors of $\widehat{H}^T\widehat{H}$ corresponding to its $h - \ell$ smallest eigenvalues. Next, note that H^* is a function of U_k, V_k, Σ, p_k, and (X, Y). Therefore, we write H^* as

$$H^*(X, Y, \theta), \\ \text{where } \theta = (U_1, \dots, U_m, V_1, \dots, V_m, p_1, \dots, p_m, \Sigma).$$

We approximate the asymptotic variance Λ by

$$\widehat{\Lambda} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{vec}\big(\widehat{\Gamma}_0^T H^*(\widehat{X}_i, Y_i, \hat{\theta})\widehat{\Psi}_0\big) \\ \times \mathrm{vec}^T\big(\widehat{\Gamma}_0^T H^*(\widehat{X}_i, Y_i, \hat{\theta})\widehat{\Psi}_0\big),$$

where $\widehat{X}_i = X_i - \hat{\mu}$ and $\hat{\theta}$ is obtained by substituting $\widehat{U}_k$, $\widehat{V}_k$, $\hat{p}_k$, and $\widehat{\Sigma}$ for U_k, V_k, p_k, and Σ in θ. It is easy to see that $\widehat{\Lambda}$ is a $\sqrt{n}$-consistent estimator of Λ. Let $\hat{\omega}_1, \dots, \hat{\omega}_{mp+m+p-\ell}$ be the eigenvalues of $\widehat{\Lambda}$. Then the distribution of T_ℓ is approximately (with $\sqrt{n}$-convergence rate) that of the random variable

$$\sum_{i=1}^{(p-\ell)(mp+p+m-\ell)} \hat{\omega}_i K_i, \tag{15}$$

where the K_i's are iid χ_1^2.

Algorithms are available in the literature for calculating the distribution of linear combination of independent χ_1^2's, such as the foregoing (see Field 1993; Bentler and Xie 2000; Cook and Setodji 2003). Alternatively, any desired quantile of (15) can be computed by simulation, which is the approach that we take. Denote $(p - \ell)(mp + p + m - \ell)$ by s. Let $W = (\hat{\omega}_1, \dots, \hat{\omega}_s)^T$ and K be an $N \times s$ matrix of iid χ_1^2 realizations. Then KW is an N-dimensional vector of iid realizations of (15). The proportion in which they are larger than T_ℓ is used as the approximate p value. Usually, an approximation for N as small as $500 \sim 1{,}000$ works quite well for our purposes.

5. SIMULATION STUDIES FOR COMPARISON AND ORDER DETERMINATION

In this section we compare the performance and computing time of DR with that of other dimension-reduction methods. In particular, we compare it with two groups of estimators. The first group includes SIR, PHD, SAVE, SIRII (Li 1991), SCR, and GCR; the second group includes the convex combinations of (SIR, SAVE), (SIR, PHD), and (SIR, SIRII), with the coefficients determined by bootstrap, as proposed by Ye and Weiss (2003). To be comprehensive, we select not only models that are known to be difficult for the classical methods, but also those that favor these methods. We also use simulation to investigate the performance of the sequential test introduced in Section 4.2 in determining the dimension of $\mathcal{S}_{Y|X}$.

5.1 Comparison With the First Group

The predictor X is generated from $\mathrm{N}(0, I_p)$, where p is taken to be 6 or 20. Let β_1 and β_2 be p-dimensional vectors with their first six components being $(1, 1, 1, 0, 0, 0)$ and $(1, 0, 0, 0, 1, 3)$. When $p = 20$, the subsequent components of β_1 and β_2 are taken to be 0. The response Y is generated from each of the following four models:

$$\text{Model I:} \quad Y = .4(\beta_1^T X)^2 + 3\sin(\beta_2^T X/4) + \sigma\epsilon;$$

$$\text{Model II:} \quad Y = 3\sin(\beta_1^T X/4) + 3\sin(\beta_2^T X/4) + \sigma\epsilon;$$

$$\text{Model III:} \quad Y = .4(\beta_1^T X)^2 + |\beta_2^T X|^{1/2} + \sigma\epsilon;$$

and

$$\text{Model IV:} \quad Y = 3\sin(\beta_2^T X/4) + [1 + (\beta_1^T X)^2]\sigma\epsilon,$$

where $\epsilon \sim \mathrm{N}(0, 1)$ and $\sigma = .2$.

In model I, the first component is symmetric about 0, so it cannot be estimated by SIR. The second component is roughly monotone within the observed domain of X, flattening out at each end; thus it is favorable to SIR. In model II, both the β_1 and the β_2 components are roughly monotone and thus are favorable to SIR. In model III, both components are symmetric about the origin and are favorable to PHD, SAVE, and SIRII. In models I–III, the structural directions β_1 and β_2 both appear in the mean function. In model IV, the direction β_1 appears in the variance component. The coefficients (e.g., .4 and 3 in model I) are chosen so that the two components are of comparable magnitude.

For each sample, the central space span(β_1, β_2) is estimated by the seven methods in the first group. The errors between the true and estimated central spaces are measured by the following distance: If $\mathcal{S}_1$ and $\mathcal{S}_2$ are two subspaces of $\mathbb{R}^p$ and P_1 and P_2 are the orthogonal projections onto them, then $\mathrm{dist}(\mathcal{S}_1, \mathcal{S}_2) = \|P_1 - P_2\|^2$, where $\|\cdot\|$ is the Euclidean matrix norm (see Li et al. 2005).

The simulation includes comparing all combinations of the four models, seven methods, and the two configurations $(n, p) = (100, 6)$ and $(500, 20)$. For all combinations except the combination of GCR and $(n, p) = (500, 20)$, we used $M = 1{,}000$ simulated random samples. We used $M = 200$ for GCR with $(n, p) = (500, 20)$ (indicated by $*$ in Table 1) because of the considerable computing time required in this case. The means of the distances from the M samples are presented in Table 1.

Because M is quite large, the standard errors of the mean distances are very small (all in the range of $.01 \sim .03$), and are not presented here. For $n = 100$, we divide X into 5 slices, each containing 20 observations. For $n = 500$, we divide X into 10 slices, each containing 50 observations. For SCR and GCR, the proportion of contour vectors is taken to be 10%, and for GCR, the tube radius is taken to be 1.5 for the case where $(n, p) = (100, 6)$ and 3.3 for the case where $(n, p) = (500, 20)$. (For more information on these tuning constants see Li et al. 2005.)

Table 1. Comparison of DR with methods in the first group

(n, p)	Model	SIR	PHD	SAVE	SIRII	SCR	GCR	DR
100, 6	I	1.504	1.433	.594	2.082	.574	.303	.355
	II	1.405	2.038	1.455	2.953	1.690	1.627	1.313
	III	2.434	.816	.540	.574	1.283	.688	.486
	IV	1.625	1.904	1.540	2.907	.959	.793	1.560
500, 20	I	1.740	1.883	1.054	2.177	.315	.290*	.252
	II	1.542	2.662	1.785	3.600	1.939	1.927*	1.523
	III	3.418	.899	.466	.468	1.482	.909*	.445
	IV	1.900	2.703	1.763	3.358	.710	.829*	1.662

*Computed from $M = 200$ samples.

As can be seen from Table 1, DR is the clear winner. DR performs better in 43 of the 48 comparisons, and in most cases the improvement is very substantial. It is also significant that DR actually outperforms SIR, PHD, SAVE, and SIRII even for models II and III, for which the classical methods work the best. SCR and GCR perform very well for model I and outstandingly for IV—a feature also noticed by Li et al. (2005). Nevertheless, except for model IV, their performance is on the whole dominated by DR.

As already mentioned, SCR and GCR require more computing time than DR. Roughly, SIR, PHD, SAVE, SIRII, and DR require similar computing time, but SCR [because it involves selecting contour directions from $\binom{n}{2}$ empirical directions] substantially increases the computing time, and GCR (because it involves calculating a conditional variance along each empirical direction) further increases the computing time by an order of magnitude. Table 2 gives the CPU times (in seconds) required by SCR, GCR, and DR for model I with $p = 20$ and n in the range of $500 \sim 10{,}000$. (For GCR, with $n > 2{,}000$, the computing times are longer than 1 hour and are not recorded.) The calculations are performed using FORTRAN on IBM ThinkPad Model T40 (1.7 GHz).

The savings in computing time is even more important when using higher-level languages, such as S–PLUS or R.

5.2 Comparison With Bootstrap Convex Combinations

We first give a brief description of the convex combination methods. Let

$$A_1 = \mathrm{var}[E(Z|Y)] \quad (\mathrm{SIR}),$$

$$A_2 = E(eZZ^T) \quad (\mathrm{PHD}),$$

$$A_3 = E[I_p - \mathrm{var}(Z|Y)]^2 \quad (\mathrm{SAVE}), \qquad \text{and}$$

$$A_4 = E\big[\mathrm{var}(Z|Y) - E(\mathrm{var}(Z|Y))\big]^2 \quad (\mathrm{SIRII}),$$

where in A_2, e is the residual for OLS; that is, $e = Y - E(Y) - E(Z^T Y)Z$. Convex combination estimators include the following variations:

$$A_{12}(\alpha) = \alpha A_1 + (1 - \alpha)A_2 \quad (\mathrm{SIR} + \mathrm{PHD}),$$

$$A_{13}(\alpha) = \alpha A_1 + (1 - \alpha)A_3 \quad (\mathrm{SIR} + \mathrm{SAVE}),$$

and

$$A_{14}(\alpha) = \alpha A_1^2 + (1 - \alpha)A_4 \quad (\mathrm{SIR}\alpha),$$

Table 2. Comparison of computing times (in seconds)

n	SCR	GCR	DR
500	2.2	35	.053
1,000	5.9	319	.081
2,000	24	4,563	.146
3,000	49		.203
5,000	137		.341
10,000	545		.699

Table 3. Comparison with bootstrapped convex combinations

Model	B–SIRα	B–(SIR + PHD)	B–(SIR + SAVE)	DR
I	$.343 \pm .020$	$.434 \pm .033$	$.529 \pm .043$	$.377 \pm .024$
II	$1.422 \pm .043$	$1.342 \pm .040$	$1.345 \pm .040$	$1.314 \pm .038$
III	$.698 \pm .042$	$.515 \pm .032$	$.503 \pm .032$	$.514 \pm .032$
IV	$1.294 \pm .036$	$1.574 \pm .035$	$1.513 \pm .036$	$1.588 \pm .033$

where $0 \le \alpha \le 1$. For estimation, we first standardize X to $\widehat{Z}$ and estimate $\mathcal{S}_{Y|Z}$ by the eigenvectors of the sample estimates of the A_{12}, A_{13}, and A_{14} corresponding to their large eigenvalues, and then transform the eigenvectors to the X-scale to estimate $\mathcal{S}_{Y|X}$, as described in Section 4. These estimators were recommended by Li (1991), Gannoun and Saracco (2003), Ye and Weiss (2003), and Zhu et al. (2005), to mitigate the difficulties of the classical methods.

Ye and Weiss (2003) proposed using the bootstrap to optimize α in the convex combinations, which we call the bootstrap convex combinations (BCCs). We now compare DR with these estimators. Because the BCC methods are computationally intensive, we take $M = 200$ and $(n, p) = (100, 6)$. We choose the optimal α by minimizing the q^2 of Ye and Weiss (2003) over the grid $\{0, .1, \ldots, .9, 1\}$, with number of bootstrap samples $n_B = 400$, as taken in that article. The mean distances from the central space and their standard errors are presented in Table 3, with B-$(\cdots)$ indicating the bootstrapped convex combinations.

Comparing Tables 1 and 3 shows that the performance of BCC is a significant improvement over that of the classical methods. However, DR remains highly competitive compared with these computationally intensive methods; DR performs better in 7 of the 12 combinations in Table 3, and in the remaining cases the differences are small relative to the magnitude of the mean distances.

In the meantime, DR requires much less computing time than the BCC methods. Let n_B be the number of bootstrap samples and let n_α be the number of α values over which q^2 is maximized. Because the computing time required by DR is similar to that required by the convex combination methods for a single α and a single bootstrap sample, BCC increases the computing time by approximately $n_B n_\alpha$-fold. Thus, if one takes $n_B \approx n$ to maintain a reasonable convergence rate for $\hat{\alpha}$, then the increase is nn_α-fold. Table 4 compares the computing times for DR and the BCC methods when applied to model I for a single sample, with $p = 20$, n in the range of $500 \sim 10{,}000$, $n_B = n$, and $n_\alpha = 11$, again using FORTRAN on IBM ThinkPad-T40. The computing times for the BCC methods for $n > 1{,}000$ are estimated by the times used for a single bootstrap sample and a single α multiplied by $n_\alpha n$.

Table 4. Comparison of computing times (in seconds)

n	B–SIRα	B–(SIR + PHD)	B–(SIR + SAVE)	DR
500	203	183	195	.053
1,000	940	880	894	.081
2,000	3,653	3,520	3,756	.146
3,000	8,048	8,027	8,317	.203
5,000	22,804	22,279	22,804	.341
10,000	93,008	91,320	93,902	.699

In practice, the sample size n is easily in the range of several thousands (see Sec. 6), in which cases the bootstrap methods can be less feasible for a single sample (the data set). In fact, even for $(n, p) = (500, 20)$, it is already infeasible to perform a full-scale simulation study; running the three BCC methods on $M = 1{,}000$ random samples would require approximately 27 days. In contrast, similar calculations for DR required only 53 seconds.

Beyond saving computing time while maintaining high accuracy, DR seems to combine the first two inverse conditional moments more naturally than the convex combination methods. For example, in SIR + PHD, the matrix A_2 is not necessarily positive semidefinite, whereas the matrix A_1 is always positive semidefinite. Thus SIR + PHD itself need not be positive semidefinite, and there seems no compelling reason to restrict α to $[0, 1]$. Moreover, in SIR + SAVE, A_1 and A_3 have different powers in Z, and it also seems plausible to use $\alpha A_1^2 + (1 - \alpha)A_3$, which brings SIR to the same scale as SAVE. Furthermore, a convex combination has no place for the interaction between first and second inverse conditional moments. In contrast, DR combines the first two inverse conditional moments, as well as their interactions, naturally through their contributions to the regression of empirical directions.

5.3 Simulation Study for Order Determination

We now evaluate by simulation the performance of the sequential test procedure described in Section 4.2. We use model I in Section 5.1 with $p = 6$ and two sample sizes, $n = 150$ and 200. We take the number of slices as $m = 25$.

For each sample size, 50 samples are generated from model I, and for each sample, the p values for hypothesis tests (9) for $\ell = 0, 1, \ldots, 5$ are computed using the simulation method described in Section 4.2, with $N = 500$. At the significance level $\alpha = .1$, and for $n = 150$, the sequential procedure gives the correct estimate of $\hat{q} = 2$ for 41 of the 50 samples. Thus the percentage of correct order determination is 82%. For $n = 200$, the proportion of correct order determination is $42/50 = 84\%$.

Figure 1 presents 12 boxplots for the p values of the 50 samples corresponding to each combination of $n = 150, 200$ and $\ell = 0, \ldots, 5$. The dotted lines indicate the significance level $\alpha = .1$. We see that the p values do behave as expected; for both sample sizes, there is a significant gap between $\ell = 0, 1$ and $\ell = 2, \ldots, 5$, indicating the correct order is 2. For $\ell = 0, 1$, the p values are small and more concentrated, but for $\ell = 2, \ldots, 5$, they are much larger and more scattered.

6. IDENTIFICATION OF HAND–WRITTEN DIGITS

To further investigate the performance of DR and demonstrate its use in real-life situations, we now apply it to a data set concerning the identification of hand-written digits $\{0, 1, \ldots, 9\}$. In the study, 44 subjects are each asked to write 250 random digits. Each digit yields a 16-dimensional feature vector, consisting of 8 pairs of randomly sampled 2-dimensional locations on the digit. The 44 subjects are divided into two groups of size 30 and 14, in which the first formed the training set (of sample size $n = 7{,}494$) and the second formed the test set (of sample size $n' = 3{,}498$). The

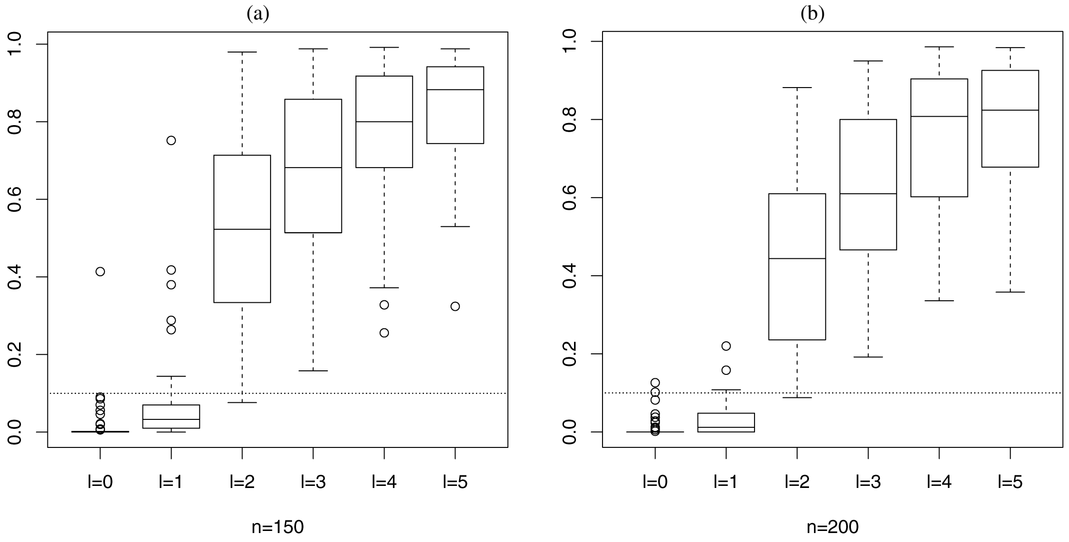

Figure 1. Boxplots of p values for sequential tests $H_0^{(\ell)}$: rank$(F_1) = 0$, for $\ell = 0, 1, \ldots, 5$. (a) $n = 150$; (b) $n = 200$. The dotted lines indicate the $\alpha = .1$ significance level.

data set is available in the UCI machine-learning repository at *ftp://ftp.ics.uci.edu/pub/machine-learning-databases/pendigits/* (see also Zhu and Hastie 2003).

We focus on the dimension reduction of the 16-dimensional feature vectors for the training set, which serves as a preparatory step for developing an efficient classifier. For clarity, we first consider the digits 0 and 9. The reduced data set comprises 1,499 cases, of which 780 are identified as 0 and 719 are identified as 9. We applied SIR, SAVE, SIRII, SCR, and DR to this data set. Figure 2 presents the perspective plots of the first three predictors estimated by SAVE, SIRII, SCR, and DR, as well as a dot plot for the SIR predictor, with the cases for digit 9 represented by “+” and the cases for digit 0 represented by “◦”. Because the response is binary, the SIR matrix has rank 1, and produces only one predictor. Although the observed values of this predictor can in principle be represented by dots plotted on a one-dimensional axis, this would be hard to see with 1,499 cases. For clearer visual effect, we plot the values of the SIR predictor on two separate axes, the upper one indicating digit 9 and the lower one indicating digit 0.

Figure 2 shows that DR performs much better than SAVE and SIRII. Although the latter succeeded in separating the two groups by variation, they failed to separate them by location. A similar result for SAVE was also reported by Zhu and Hastie (2003). In comparison, DR separates the two groups by both location and variation: projected on DR1-direction the two groups are separated by location, whereas projected on the (DR2, DR3)-plane the digit-9 group has much larger variation than the digit-0 group. SCR performs very similarly to DR for this data set, but takes longer to compute.

We also see that, although SIR does provide good locational separation, it cannot provide information about the differentiated variations such as that revealed by DR and SCR. The additional separation in variation is clearly useful for classification. For example, if there is an observation with projection on the SIR direction (which is roughly the same as the DR1 direction) in the middle of the two groups, but projection on the (DR2, DR3) plane outside the range of the (much smaller) variation of the digit-0 group, then it is difficult to classify by SIR but easier to classify by DR (as belonging to the digit-9 group).

To make this point clearer, we now apply SIR and DR to the data set that involves three digits: {0, 6, 9}. Because of the similar shape of these digits, they are among the more difficult to classify. The number of digit-6 cases is 720; thus the sample size is $n = 2{,}219$. With three slices, the SIR matrix has rank 2, and therefore can estimate at most two vectors in the central space. Figure 3 presents the two-dimensional plot of (SIR1, SIR2), the three-dimensional plot of (DR1, DR2, DR3), and the two-dimensional plots of (DR1, DR2) (top view), (DR1, DR3) (front view), and (DR2, DR3) (side view). Similar to the previous comparison, SIR only provides locational separation of the three groups, but DR provides, in addition, a sharp distinction in variation. The plot of (DR1, DR3) shows a substantial difference in the covariance between DR1 and DR3 for the digit-9 group from the covariance between DR1 and DR3 for the other two groups. A similar difference is apparent in the plot of (DR2, DR3). These differences in covariances provide valuable information for classification. We also see that two-dimensional plots of (DR1, DR2) and (SIR1, SIR2) are very similar, indicating that DR and SIR provide similar separation of locations. The different features of the three groups are comprehensively demonstrated by the three-dimensional plot of (DR1, DR2, DR3).

The computing time saved by DR is already important for data sets of this size, as can be seen in Tables 2 and 4, which cover the present sizes with a similar p. For example, for the reduced data set involving digits {0, 6, 9}, DR requires a small fraction of a second, SCR requires about 30 seconds and GCR and the BCC methods require about 1 hour or longer. For analyzing the full data set, DR would require no more than a few seconds; whereas GCR and BCC would require 1 day or longer.

7. CONCLUDING REMARKS

DR substantially improves the accuracy of contour regressions and decreases computing time. For a wide range of

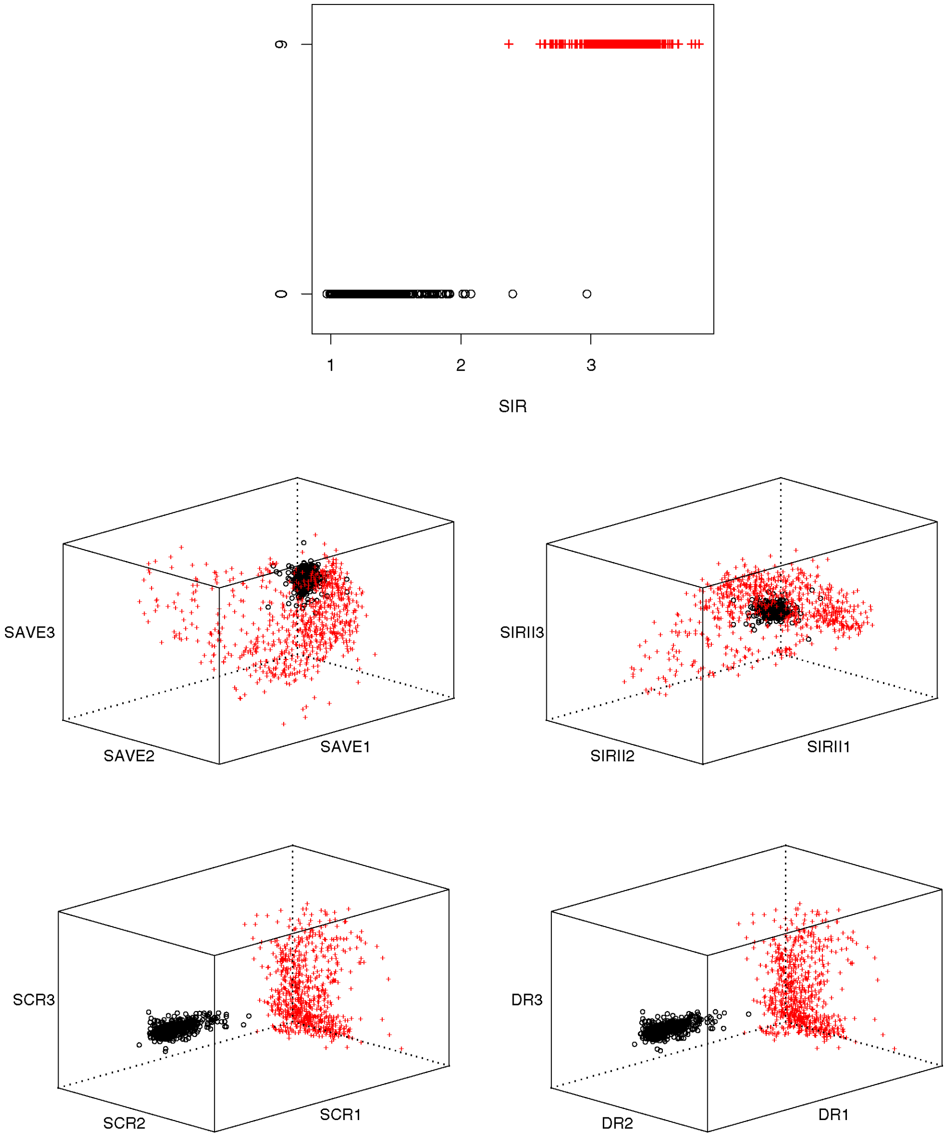

Figure 2. Perspective plots for the first three predictors estimated by SIR, SAVE, SIRII, SCR, and DR (+, digit 9; ○, digit 0).

models, it is clearly more accurate than classical dimension-reduction methods such as SIR, SAVE, SIRII, and PHD and is highly competitive with the BCC methods, with sharply reduced computing time. For commonly seen data sizes, the savings in computing time is of practical importance, especially when compared with GCR and the BCC methods. Its simple asymptotic structure allows us to derive its asymptotic distribution and sequential test for order determination, which are not yet available for contour regressions and the BCC methods. It is $\sqrt{n}$-consistent and exhaustive under very mild assumptions—milder than those required by SCR. It should be noted that DR, like SAVE, SIRII, PHD, and SCR, is a second-moment–based method. Thus, for the situations in which the regression surfaces have highly fluctuating shapes (such as trigonometric function with high frequency), such methods as GCR may perform better. But in wide applications, the first and second inverse moments, combined adequately, will give practical and satisfactory results, and DR provides a natural, economic, and efficient combination of such inverse moments.

APPENDIX: PROOFS

Proof of Theorem 1

By lemma 2.1 of Li et al. (2005) and Cook (1998, prop. 4.6), $(Z, Y) \perp\!\!\!\perp (\widetilde{Z}, \widetilde{Y})$ implies that $Z \perp\!\!\!\perp \widetilde{Z}|(Y, \widetilde{Y})$, $Z \perp\!\!\!\perp \widetilde{Y}|Y$, and $\widetilde{Z} \perp\!\!\!\perp Y|\widetilde{Y}$. Thus $A(Y, \widetilde{Y})$ can be expanded as

$$E(ZZ^T|Y) - E(Z|Y)E(\widetilde{Z}^T|\widetilde{Y}) - E(\widetilde{Z}|\widetilde{Y})E(Z^T|Y) + E(\widetilde{Z}\widetilde{Z}^T|\widetilde{Y}). \quad \text{(A.1)}$$

It suffices to show that $\mathcal{S}_{Y|Z}^{\perp} \subset \{\text{span}(I_p - A(Y, \widetilde{Y}))\}^{\perp}$. Let $v \in \mathcal{S}_{Y|Z}^{\perp}$. By assumption (a), $E(v^T Z|PZ) = \alpha^T PZ$ for some $\alpha \in \mathbb{R}^p$. Multiply both sides by $\alpha^T PZ$ and then take unconditional expectation to obtain $\alpha^T P\alpha = v^T P\alpha = 0$. Thus $E(v^T Z|PZ) = 0$. By assumption (b), $E[(v^T Z)^2|PZ] = c + E^2(v^T Z|PZ) = c$, where c is a constant. Take unconditional expectations on both sides to obtain $c = v^T v$. Thus $E[(v^T Z)^2|PZ] = v^T v$. Because $Y \perp\!\!\!\perp Z|PZ$, we have

$$E(v^T Z|Y) = E\big[E(v^T Z|PZ)|Y\big] = 0,$$

$$E[(v^T Z)^2|Y] = E\big\{E[(v^T Z)^2|PZ]|Y\big\} = v^T v.$$

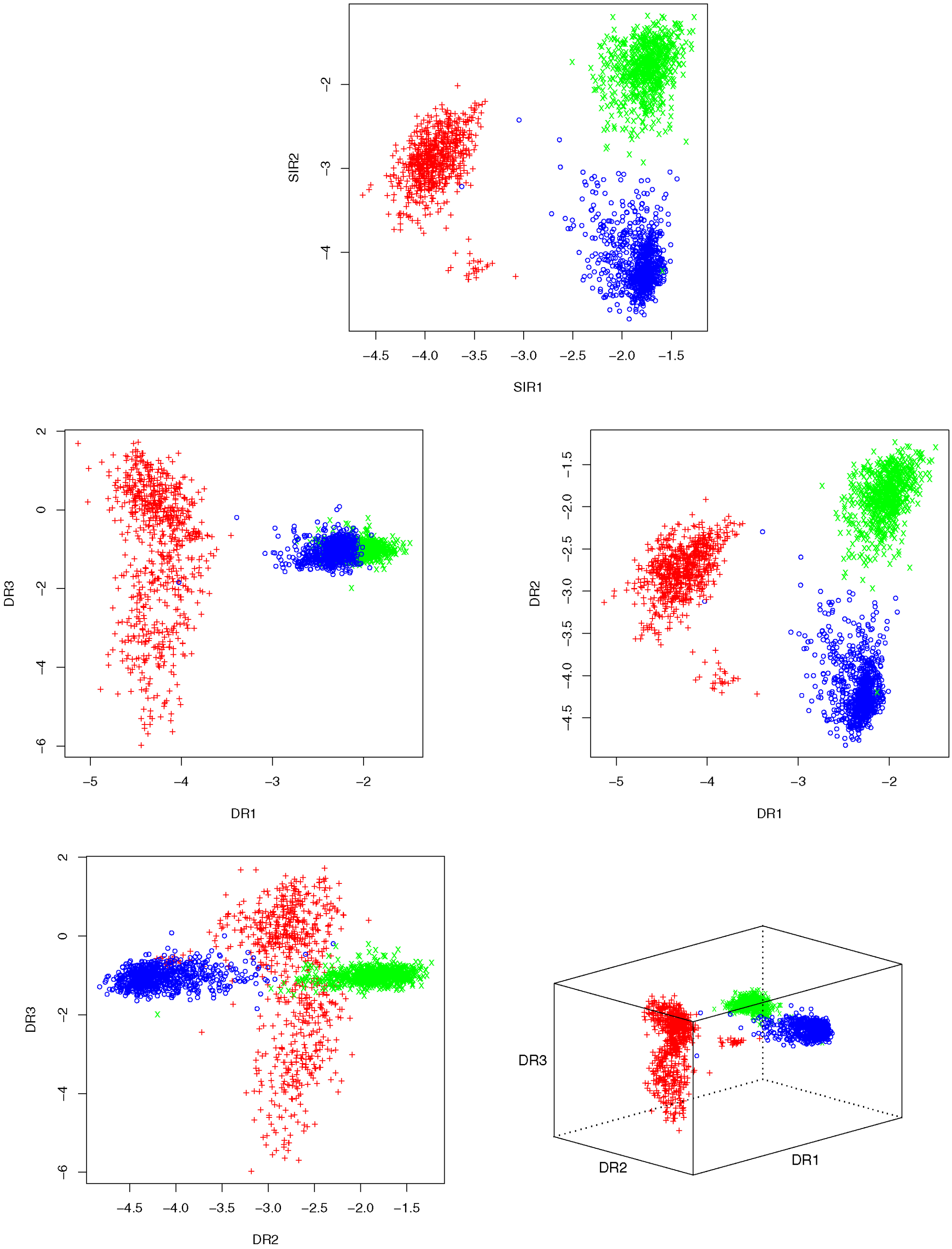

Figure 3. Comparison of SIR and DR when three digits are involved (+, digit 9; ∘, digit 0; ×, digit 6).

Substitute these into (A.1), bearing in mind that (Z, Y) and $(\widetilde{Z}, \widetilde{Y})$ have the same distribution, to obtain $v^T A(Y, \widetilde{Y}) v = 2v^T v$, implying that $v^T[2I_p - A(Y, \widetilde{Y})]v = 0$.

Proof of Theorem 2

Because $E(Z) = 0$ and $\mathrm{var}(Z) = I_p$, we have, by (A.1), $EA(Y, \widetilde{Y}) = 2I_p$. Thus

$$G = 4I_p - 4EA(Y, \widetilde{Y}) + EA^2(Y, \widetilde{Y}) = -4I_p + EA^2(Y, \widetilde{Y}). \qquad \text{(A.2)}$$

Write the first two terms in (A.1) as $B(Y, \widetilde{Y})$. Then $A(Y, \widetilde{Y}) = B(Y, \widetilde{Y}) + B(\widetilde{Y}, Y)$ and, because $Y, \widetilde{Y}$ are iid, $B(Y, \widetilde{Y})$ and $B(\widetilde{Y}, Y)$ have the same distribution. Thus

$$EA^2(Y, \widetilde{Y}) = 2EB^2(Y, \widetilde{Y}) + 2E[B(Y, \widetilde{Y})B(\widetilde{Y}, Y)]. \qquad \text{(A.3)}$$

In (A.3), the expectation $EB^2(Y, \widetilde{Y})$ can be decomposed into the four terms,

$$\begin{aligned} &E[E^2(ZZ^T|Y)] - E\big[E(ZZ^T|Y)E(Z|Y)E(\widetilde{Z}^T|\widetilde{Y})\big] \\ &\quad - E\big[E(Z|Y)E(\widetilde{Z}^T|\widetilde{Y})E(ZZ^T|Y)\big] \\ &\quad + E\big[E(Z|Y)E(\widetilde{Z}^T|\widetilde{Y})E(Z|Y)E(\widetilde{Z}^T|\widetilde{Y})\big]. \end{aligned}$$

The second and third terms are 0 because $Y \perp\!\!\!\perp \widetilde{Y}$ and $E(\widetilde{Z}) = 0$. Moreover, because (Z, Y) and $(\widetilde{Z}, \widetilde{Y})$ have the same distribution, the fourth term can be rewritten as $E^2[E(Z|Y)E(Z^T|Y)]$. Thus

$$EB^2(Y,\widetilde{Y}) = E[E^2(ZZ^T|Y)] + E^2\big[E(Z|Y)E(Z^T|Y)\big]. \qquad (A.4)$$

Use the same argument to show that

$$\begin{aligned} &E[B(Y,\widetilde{Y})B(\widetilde{Y},Y)] \\ &\quad = I_p + E\big[E(Z^T|Y)E(Z|Y)\big]E\big[E(Z|Y)E(Z^T|Y)\big]. \qquad (A.5)\end{aligned}$$

Now combine (A.2)–(A.5) to complete the proof.

Proof of Theorem 3

We first note that if $\operatorname{span}(G) \subset \mathcal{S}_{Y|Z}$, $G = G^T$ and $G \ge 0$, then

$$\operatorname{span}(G) = \mathcal{S}_{Y|Z} \quad \text{iff } v^T G v > 0 \text{ for all } v \in \mathcal{S}_{Y|Z}, v \neq 0. \qquad (A.6)$$

To see this, suppose that $\operatorname{span}(G) \subsetneqq \mathcal{S}_{Y|Z}$. Then $v^T G v = 0$ for any $v \neq 0$, $v \in \mathcal{S}_{Y|Z} \ominus \operatorname{span}(G)$. Conversely, for $\operatorname{span}(G) = \mathcal{S}_{Y|Z}$, $v \in \mathcal{S}_{Y|Z}$, $v \neq 0$, we have $v \in \operatorname{span}(G)$, and hence $v^T G v > 0$. Note that $\operatorname{span}(G) \subset \mathcal{S}_{Y|Z}$ is guaranteed by assumptions (a) and (b) of Theorem 1, and $G = G^T$ and $G \ge 0$ follow from the definition of G.

We first show that statement a implies that $\mathcal{S}_{\mathrm{DR}} = \mathcal{S}_{Y|Z}$. Let $v \in \mathcal{S}_{Y|Z}$ and $v \neq 0$. By (A.6), it suffices to show that $v^T G v > 0$. Without loss of generality, assume that $\|v\| = 1$. Write $A(Y,\widetilde{Y}) - 2I_p$ as $C(Y,\widetilde{Y})$. Then

$$\begin{aligned} v^T G v = v^T E\big[C(Y,\widetilde{Y})(I_p - vv^T)C(Y,\widetilde{Y})\big]v^T \\ + E\big[(v^T C(Y,\widetilde{Y})v)^2\big]. \qquad (A.7)\end{aligned}$$

Because $I_p - vv^T \ge 0$, the first term on the right is nonnegative. By statement a, $v^T A(Y,\widetilde{Y})v$ is nondegenerate; thus $v^T C(Y,\widetilde{Y})v$ is nondegenerate. Then, by Jensen's inequality,

$$E\big[(v^T C(Y,\widetilde{Y})v)^2\big] > \big[E(v^T C(Y,\widetilde{Y})v)\big]^2 = 0, \qquad (A.8)$$

where the equality holds because $EC(Y,\widetilde{Y}) = 0$.

We now show that a $\Leftrightarrow$ b. Let $\Phi_1(Y) = E(v^T Z|Y)$ and $\Phi_2(Y) = E[(v^T Z)^2|Y]$. Then

$$E\big[(v^T(Z - \widetilde{Z}))^2|Y,\widetilde{Y}\big] = \Phi_2(Y) - 2\Phi_1(Y)\Phi_1(\widetilde{Y}) + \Phi_2(\widetilde{Y}).$$

Because $E[\Phi_2(Y)] = 1$ and $E[\Phi_1(Y)] = 0$, we have

$$\begin{aligned} &\operatorname{var}\big[\Phi_2(Y) - 2\Phi_1(Y)\Phi_1(\widetilde{Y}) + \Phi_2(\widetilde{Y})\big] \\ &\qquad = E\big[\Phi_2(Y) - 1 - 2\Phi_1(Y)\Phi_1(\widetilde{Y}) + \Phi_2(\widetilde{Y}) - 1\big]^2.\end{aligned}$$

It is easy to see that the three random variables $\{\Phi_2(Y) - 1, \Phi_1(Y)\Phi_1(\widetilde{Y}), \Phi_2(\widetilde{Y}) - 1\}$ are pairwise uncorrelated. Thus the foregoing reduces to

$$E[\Phi_2(Y) - 1]^2 + 4E[\Phi_1^2(Y)]E[\Phi_1^2(\widetilde{Y})] + E[\Phi_2(\widetilde{Y}) - 1]^2,$$

which is 0 if and only if both $\operatorname{var}(\Phi_2(Y))$ and $\operatorname{var}(\Phi_1(Y))$ are 0.

Proof of Theorem 4

Write the right side of (5) as $G_1 + \cdots + G_6$. Let v be a vector orthogonal to $\mathcal{S}_{\mathrm{SAVE}}$. Because $\mathcal{S}_{\mathrm{SIR}} \subset \mathcal{S}_{\mathrm{SAVE}}$, we have $E(Z^T|Y)v = 0$ and $[I_p - \operatorname{var}(Z|Y)]v = 0$ almost surely. Therefore, $G_i v = 0$ for $i = 1, \ldots, 6$, which implies that $v \perp \mathcal{S}_{\mathrm{DR}}$, and thus $\mathcal{S}_{\mathrm{DR}} \subset \mathcal{S}_{\mathrm{SAVE}}$.

Now let v be a vector orthogonal to $\mathcal{S}_{\mathrm{DR}}$. Then, by (4),

$$v^T E[E^2(ZZ^T - I_p|Y)]v = 0$$

and

$$v^T E\big[E(Z^T|Y)E(Z|Y)\big]E\big[E(Z|Y)E(Z^T|Y)\big]v = 0.$$

The second equality implies that $E(Z^T|Y)v = 0$ almost surely. The first equality can be reexpressed as

$$\begin{aligned} 0 &= v^T E[\operatorname{var}(Z|Y) - I_p]^2 v \\ &\quad + v^T E\big[(\operatorname{var}(Z|Y) - I_p)E(Z|Y)E(Z^T|Y)\big]v \\ &\quad + v^T E\big[E(Z|Y)E(Z^T|Y)(\operatorname{var}(Z|Y) - I_p)\big]v \\ &\quad + v^T E\big[E(Z|Y)E(Z^T|Y)\big]^2 v.\end{aligned}$$

That $E(Z^T|Y)v = 0$ almost surely implies that the second, third, and fourth terms are 0. Thus the first term must also be 0, implying that $v \perp \mathcal{S}_{\mathrm{SAVE}}$.

The derivation of asymptotic expansions hinges on calculation of the influence function in (13), which is in fact the Frechet derivative of the mapping $S(\cdot)$. (Strictly speaking, the influence function is the representation of the Frechet derivative, but there is no need to make this distinction here. Our terminology identifies Frechet derivative with its representation.) We use $S^*(F)$ to indicate the Frechet derivative; for example, $E^* g(X,F)$ denotes the Frechet derivative of $\int g(X,F)\,dF$. One Frechet derivative is particularly useful; if $g(x)$ does not depend on F and $Eg^2(X) < \infty$, then

$$E^* g(X) = g(X) - E(g(X)). \qquad (A.9)$$

We will use this repeatedly in the sequel. We also note that, even though $E(X)$ is assumed to be 0, its Frechet derivative is not 0. Thus, when taking the Frechet derivative, it is helpful to write down $E(X)$ to remind ourselves of this fact.

Proof of Lemma 1

The expression for p_k^* follows directly from (A.9). Write $I(Y \in J_k)$ as R_k. In this notation,

$$ER_k = p_k, \qquad E(XR_k)/p_k = U_k, \qquad E(XX^T R_k)/p_k - \Sigma = V_k. \qquad (A.10)$$

Differentiating the second equation yields

$$U_k^* = E^*[(X - EX)R_k]/ER_k + E[(X - EX)R_k](1/ER_k)^*. \qquad (A.11)$$

In the foregoing,

$$\begin{aligned} &(1/ER_k)^* = -(1/E^2 R_k)E^* R_k = -(R_k - p_k)/p_k^2, \\ &E^*[(X - EX)R_k] \\ &\quad = E^*(XR_k) - E^*(X)ER_k - E(X)E^*(R_k) \\ &\quad = XR_k - E(XR_k) - Xp_k. \end{aligned} \qquad (A.12)$$

Substitute these into (A.11) and apply the relations in (A.10) to verify the formula for U_k^*.

To verify the formula for V_k^*, rewrite V_k in (A.10) as

$$V_k = E\big[(X - EX)(X - EX)^T R_k\big]/ER_k - E[(X - EX)(X - EX)].$$

Differentiating both sides of the equation yields

$$\begin{aligned} V_k^* &= E^*\big[(X - EX)(X - EX)^T R_k\big]/ER_k \\ &\quad + E\big[(X - EX)(X - EX)^T R_k\big](1/ER_k)^* \\ &\quad - E^*[(X - EX)(X - EX)]. \end{aligned} \qquad (A.13)$$

In the foregoing, $(1/ER_k)^*$ is given by (A.12) and

$$\begin{aligned} &E^*\big[(X - EX)(X - EX)^T R_k\big] \\ &\quad = E^*(XX^T R_k) - E(XR_k)E^*(X^T) - E^*(X)E(X^T R_k) \\ &\quad = XX^T R_k - E(XX^T R_k) - E(XR_k)X^T - XE(X^T R_k), \\ &E^*\big[(X - EX)(X - EX)^T\big] = XX^T - \Sigma. \end{aligned}$$

Substitute these into (A.13) and evoke the relations in (A.10) to complete the proof.

Proof of Theorem 5

Differentiate H_{1k} as defined in (11) to obtain

$$H_{1k}^* = (1/\sqrt{2p_k})p_k^* V_k + \sqrt{2p_k}V_k^*.$$

The Frechet derivative H_2^* can be derived by straightforward differentiation. Finally, H_3 can be expressed in terms of H_2 as

$$\begin{aligned} H_{3k} &= \sqrt{2}(\sqrt{2})^{-1/2}\Big(\sqrt{2}\sum U_\ell^T U_\ell p_\ell\Big)^{1/2}\sqrt{p_k}U_k \\ &= 2^{1/4}[\mathrm{tr}(H_2)]^{1/2}\sqrt{p_k}U_k. \end{aligned}$$

Now (Frechet) differentiate the right side to obtain H_{3k}^*.

[Received October 2005. Revised January 2007.]

REFERENCES

Bentler, P. M., and Xie, J. (2000), "Corrections to Test Statistics in Principal Hessian Direction," *Statistics and Probability Letters*, 47, 381–389.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore: Johns Hopkins University Press.

Cook, R. D. (1994), "Using Dimension Reduction Subspaces to Identify Important Inputs in Models of Physical Systems," in *Proceedings of the Physical and Engineering Sciences Section*, American Statistical Association, pp. 18–25.

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

——— (1998), *Regression Graphics*, New York: Wiley.

Cook, R. D., and Critchley, F. (2000), "Identifying Regression Outliers and Mixtures Graphically," *Journal of the American Statistical Association*, 95, 781–794.

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

Cook, R. D., and Li, B. (2004), "Determining the Dimension of Iterative Hessian Transformation," *The Annals of Statistics*, 32, 2501–2531.

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

Cook, R. D., and Setodji, C. M. (2003), "A Model-Free Test for Reduced Rank in Multivariate Regression," *Journal of the American Statistical Association*, 98, 340–351.

Cook, R. D., and Weisberg, S. (1991), Discussion of "Sliced Inverse Regression for Dimension Reduction," by K. C. Li, *Journal of the American Statistical Association*, 86, 316–342.

Diaconis, P., and Freedman, D. (1984), "Asymptotics of Graphical Projection Pursuit," *The Annals of Statistics*, 12, 793–815.

Duan, N., and Li, K. C. (1991), "Slicing Regression: A Link-Free Regression Method," *The Annals of Statistics*, 19, 505–530.

Eaton, M. L., and Tyler, D. (1994), "The Asymptotic Distribution of Singular Values With Application to Canonical Correlations and Correspondence Analysis," *Journal of Multivariate Analysis*, 50, 238–264.

Fernholz, L. T. (1983), *von Mises Calculus for Statistical Functionals*, New York: Springer.

Field, C. (1993), "Tail Areas of Linear Combinations of Chi-Squares and Noncentral Chi-Squares," *Journal of Statistical Computation and Simulation*, 45, 243–248.

Gannoun, A., and Saracco, J. (2003), "An Asymptotic Theory for SIRα Method," *Statistica Sinica*, 13, 297–310.

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

Hallin, M., and Paindaveine, D. (2002), "Optimal Tests for Multivariate Location Based on Interdirections and Pseudo-Mahalanobis Ranks," *The Annals of Statistics*, 30, 1103–1133.

——— (2005), "Affine-Invariant Aligned Rank Tests for the Multivariate General Linear Model With VARMA Errors," *Journal of Multivariate Analysis*, 93, 122–163.

Hampel, F. R. (1974), "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383–397.

Hettmansperger, T. P., and Oja, H. (1994), "Affine-Invariant Multivariate Multisample Sign Tests," *Journal of the Royal Statistical Society*, Ser. B, 56, 235–249.

Li, B., and Yin, X. (2006), "On Surrogate Dimension Reduction for Measurement Error Regression: An Invariance Law," *The Annals of Statistics*, to appear.

Li, B., Zha, H., and Chiaromonte, C. (2005), "Contour Regression: A General Approach to Dimension Reduction," *The Annals of Statistics*, 33, 1580–1616.

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

——— (1992), "On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma," *Journal of the American Statistical Association*, 87, 1025–1039.

Li, K. C., and Duan, N. (1989), "Regression Analysis Under Link Violation," *The Annals of Statistics*, 17, 1009–1052.

Oja, H. (1999), "Affine Invariant Multivariate Sign and Rank Tests and Corresponding Estimates: A Review," *Scandinavian Journal of Statistics*, 26, 319–343.

Portnoy, S. (1986), "On the Central Limit Theorem in $\mathbb{R}^p$ When p Approaches ∞," *Probability Theory and Related Fields*, 73, 571–583.

Randles, R. H. (1989), "A Distribution-Free Multivariate Sign Test Based on Interdirections," *Journal of the American Statistical Association*, 84, 1045–1050.

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

Serfling, R. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley.

Ye, Z., and Weiss, R. E. (2003), "Using the Bootstrap to Select One of a New Class of Dimension Reduction Methods," *Journal of the American Statistical Association*, 98, 968–979.

Zhu, L.-X., and Fang, K.-T. (1996), "Asymptotics for Kernel Estimate of Sliced Inverse Regression," *The Annals of Statistics*, 24, 1053–1068.

Zhu, L.-X., Ohtaki, M., and Li, Y. (2005), "On Hybrid Methods of Inverse Regression–Based Algorithms," *Computational Statistics and Data Analysis*, 51, 2621–2635.

Zhu, M., and Hastie, T. J. (2003), "Feature Extraction for Nonparametric Discriminant Analysis," *Journal of Computational and Graphical Statistics*, 12, 101–120.