Combining eigenvalues and variation of eigenvectors for order determination

Author(s): WEI LUO and BING LI

Source: *Biometrika*, Vol. 103, No. 4 (DECEMBER 2016), pp. 875-887
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: https://www.jstor.org/stable/26363492
Accessed: 25-07-2023 08:13 +00:00

Biometrika (2016), **103**, 4, *pp.* 875–887
Printed in Great Britain
doi: 10.1093/biomet/asw051

Combining eigenvalues and variation of eigenvectors for order determination

BY WEI LUO

Department of Statistics and Computer Information Systems, Baruch College, One Bernard Baruch Way, New York, New York 10010, U.S.A.
wei.luo@baruch.cuny.edu

AND BING LI

Department of Statistics, The Pennsylvania State University, 326 Thomas Building, University Park, Pennsylvania 16802, U.S.A.
bing@stat.psu.edu

SUMMARY

In applying statistical methods such as principal component analysis, canonical correlation analysis, and sufficient dimension reduction, we need to determine how many eigenvectors of a random matrix are important for estimation. This problem is known as order determination, and amounts to estimating the rank of a matrix. Previous order-determination procedures rely either on the decreasing pattern, or elbow, of the eigenvalues, or on the increasing pattern of the variability in the directions of the eigenvectors. In this paper we propose a new order-determination procedure by exploiting both patterns: when the eigenvalues of a random matrix are close together, their eigenvectors tend to vary greatly; when the eigenvalues are far apart, their variability tends to be small. The combination of both helps to pinpoint the rank of a matrix more precisely than the previous methods. We establish the consistency of the new order-determination procedure, and compare it with other such procedures by simulation and in an applied setting.

Some key words: Bootstrap; Canonical correlation analysis; Directional regression; Ladle estimator; Principal component analysis; Sliced inverse regression.

1. INTRODUCTION

Order determination is needed in many statistical methodologies. For example, underlying principal component analysis is the following statistical model (Jolliffe, 2002, p. 151)

$$X = Z + \epsilon, \tag{1}$$

where Z and ϵ are independent p-dimensional random vectors, $M = \mathrm{var}(Z)$ is a singular matrix with rank $d < p$, and $\mathrm{var}(\epsilon) = \sigma^2 I_p$ where I_p is the identity matrix. The principal components are the projections of X onto the first d eigenvectors of M. Here, the order-determination problem is to estimate d, the rank of M.

In canonical correlation analysis (Jolliffe, 2002, p. 222), we observe samples of two random vectors X and Y, and need to determine how many left- or right-singular vectors to retain. This

corresponds to determining the rank of the matrix

$$M = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1/2},$$

where $\Sigma_{XX} = \mathrm{var}(X)$, $\Sigma_{YY} = \mathrm{var}(Y)$, $\Sigma_{XY} = \mathrm{cov}(X, Y) = \Sigma_{YX}^{\mathrm{T}}$.

In independent component analysis (Hyvärinen et al., 2001), one often needs to determine how many components of a random vector have excess kurtosis, which can be regarded as a special case of the general order-determination problem. Specifically, let X be a p-dimensional random vector and let $Z = \mathrm{var}(X)^{-1/2}\{X - E(X)\}$ be its standardized version. The fourth-order blind identification estimator is based on the sample version of the matrix

$$M = [E\{(Z^{\mathrm{T}}Z)\,ZZ^{\mathrm{T}}\} - (p+2)I_p]^2.$$

In independent component analysis, the eigenvectors with nonzero eigenvalues of this matrix correspond to components with excess kurtosis, which are of interest; whereas components that have no excess kurtosis are regarded as noise.

Finally, this problem arises in sufficient dimension reduction (Li, 1991, 1992; Cook, 1994, 1998) where X is a p-dimensional random vector, Y is a random variable, and we are interested in finding a lower-dimensional linear predictor $\beta^{\mathrm{T}}X$ such that Y is independent of X given $\beta^{\mathrm{T}}X$; that is,

$$Y \perp\!\!\!\perp X \mid \beta^{\mathrm{T}}X. \tag{2}$$

Because this relation is invariant under the transformation $\beta \mapsto \beta A$ for any nonsingular matrix A, the identifiable parameter in (2) is $\mathrm{span}(\beta)$, the column space of β. The central subspace is defined as the intersection of all subspaces $\mathrm{span}(\beta)$ that satisfy (2). This subspace, denoted by $\mathcal{S}_{Y|X}$, is the target of sufficient dimension reduction (Cook, 1998; Cook & Li, 2002). As shown in Yin et al. (2008), under very mild conditions $\mathcal{S}_{Y|X}$ also satisfies (2). A typical sufficient dimension reduction estimator is in the form of a matrix-valued statistic $\hat{M}$, which converges to a fixed M with the property $\mathrm{span}(M) \subseteq \mathcal{S}_{Y|X}$ or even $\mathrm{span}(M) = \mathcal{S}_{Y|X}$. Thus, again, we face the problem of estimating the rank of M. Important examples of such estimators include sliced inverse regression (Li, 1991), the sliced average variance estimator (Cook & Weisberg, 1991), and directional regression (Li & Wang, 2007). A similar problem also arises in nonlinear sufficient dimension reduction (Wu, 2008; Lee et al., 2013).

Existing methods for order determination rely either on the magnitude of the eigenvalues of $\hat{M}$ or on the variability of the eigenvectors of $\hat{M}$. Sequential tests (Fujikoshi, 1977; Schott, 1994; Cook & Li, 2004; Bura & Yang, 2011) are based on the eigenvalues of $\hat{M}$. Information criteria (Bai & Ng, 2002; Gunderson & Muirhead, 1997; Zhu et al., 2006) are based on a monotone function of the eigenvalues plus a deterministic penalty term. Both types of procedures are uniquely determined by the scree plot of the kth eigenvalues $\hat{\lambda}_k$ of $\hat{M}$ versus k. They hinge on the fact that $\hat{\lambda}_k$ drops to near 0 at $k = d + 1$. Figure 1(a) is a typical scree plot, taken from Model 4 in §5, which has $d = 2$. It indeed shows a drop at $k = 3$.

As an alternative to the eigenvalue-based methods, Ye & Weiss (2003) proposed an eigenvector-based order-determination procedure for sufficient dimension reduction, which is also readily applicable to other problems. In applying this procedure, we first generate a set of bootstrap samples and compute the corresponding set of bootstrap estimates of $\mathcal{S}_{Y|X}$ in the form of eigenvectors. Then, for each $k < p$, we evaluate the variability of the bootstrap estimates of $\mathcal{S}_{Y|X}$ around the full sample estimate of $\mathcal{S}_{Y|X}$. We refer to this variability as the bootstrap variability of eigenvectors;

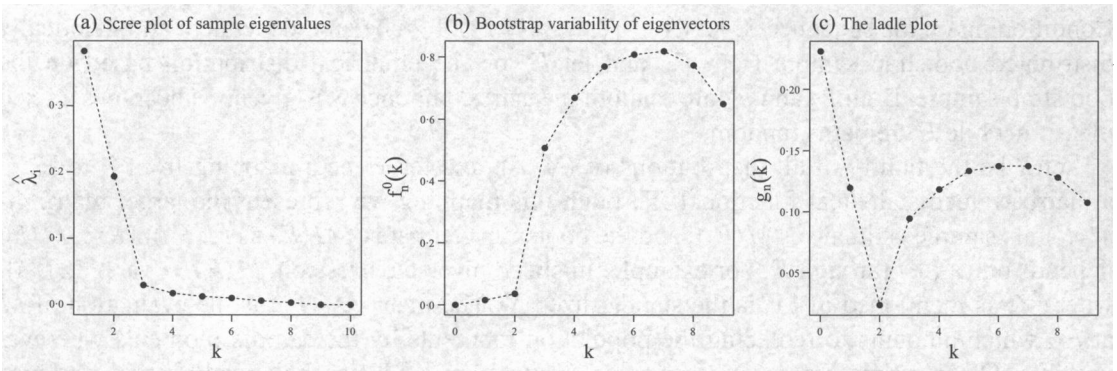

Fig. 1. Benefit of combining eigenvalues and variability of eigenvectors. The vertical axis in panel (a) represents the sample eigenvalues $\hat{\lambda}_k$; that in panel (b) represents the bootstrap variability of eigenvectors $f_n^0(k)$; that in panel (c) is a combination of both measures, $g_n(k)$. The quantities $f_n^0(k)$ and $g_n(k)$ are rigorously defined in §2.

see §2. Ye and Weiss speculated that when $k = d$, due to the consistency of $\hat{M}$, each of the k-dimensional bootstrap subspaces converges to the column space of M, and hence has small bootstrap variability. When $k > d$, each of the k-dimensional bootstrap subspaces estimates the column space of M and an arbitrary part of the null space of M, leading to large bootstrap variability. Based on this intuition, Ye and Weiss proposed to estimate d as the largest k before a jump in the bootstrap variability. Figure 1(b) shows the bootstrap variability of eigenvectors for the same example, which does show a jump at $k = 3$.

We propose a new estimator that combines both the eigenvalues and the bootstrap eigenvector variability of $\hat{M}$. It is based on the observation that when the eigenvalues of a random matrix are far apart, the bootstrap variability of the corresponding eigenvectors tends to be small, and when the eigenvalues are close together, this bootstrap variability tends to be large. By exploiting this special eigenvalue-eigenvector pattern, we develop a new order-determination method that is more sensitive and accurate.

This eigenvalue-eigenvector pattern is clearly visible in Figs. 1(a) and (b): for $k \leqslant d = 2$, the bootstrap variability of eigenvectors stays relatively flat but the magnitude of eigenvalues sharply decreases; whereas for $k > d$, the bootstrap variability of eigenvectors sharply increases while the magnitude of eigenvalues becomes relatively flat. As both functions provide useful information about the true rank d, a combination of them could lead to a sharper estimator of d. Indeed, Fig. 1(c) is based on a combination of both functions. We see that the curve is minimized at d, and its shape resembles a ladle. From our experience this shape appears in many examples, so we call such a plot the ladle plot and its minimizer the ladle estimator. In addition to proposing the ladle estimator as an order-determination method, we also propose the ladle plot as an alternative to the scree plot to assist order determination.

We relegate all the proofs, some lemmas, a corollary, some simulation results, and the R code to implement the method, to the online Supplementary Material.

2. Ladle estimator and ladle plot

Let $S = \{(X_1, Y_1), (X_2, Y_2), \ldots\}$ be a sequence of independent copies of (X, Y), where Y_i and Y can be ignored in an unsupervised learning such as principal component analysis. Let F be the distribution of (X, Y), and let F_n be the empirical distribution based on $(X_1, Y_1), \ldots, (X_n, Y_n)$.

Conditioning on the sequence S, let $(X^*_{1,n}, Y^*_{1,n}), \ldots, (X^*_{n,n}, Y^*_{n,n})$ be an independent and identically distributed bootstrap sample from F_n, and let F^*_n be the empirical distribution based on the bootstrap sample. Both F_n and F^*_n are random measures, but once S is given, F_n becomes a fixed measure while F^*_n remains random.

Let $\mathfrak{F}$ be the family of all distributions of (X, Y), and let M be a mapping from $\mathfrak{F}$ to $\mathbb{R}^{p\times p}$, a matrix-valued statistical functional. Through this mapping we represent the target matrix as $M(F)$, its sample estimate as $M(F_n)$, and the bootstrap estimate of $M(F)$ as $M(F^*_n)$. Here, $M(F)$ depends on (X, Y) through F. For example, in sliced inverse regression, $M(F) = \mathrm{var}\{E(Z|Y)\}$, where Z, as mentioned in §1, is the standardized X. The matrix $M(F_n)$ is the evaluation of M at F_n, which amounts to replacing the population moments by the sample moments wherever possible. Other estimators, such as principal component analysis, the sliced average variance estimator and directional regression, can also be expressed using statistical functionals. More details can be found in Li et al. (2005). For simplicity, we write $M(F)$ as M, $M(F_n)$ as $\hat{M}$, and $M(F^*_n)$ as M^*. Note that M is used both as the mapping $M : \mathfrak{F} \to \mathbb{R}^{p\times p}$ and as its evaluation at the true distribution F.

In applications such as sufficient dimension reduction and principal component analysis, the matrices M and $\hat{M}$ are symmetric, and our development focuses on this case. However, our method allows asymmetric matrices, because we can apply it to $\hat{M}\hat{M}^{\mathrm{T}}$ or $\hat{M}^{\mathrm{T}}\hat{M}$, which amounts to replacing the eigenvalues and eigenvectors of $\hat{M}$ by its squared singular values and singular vectors. With this modification, all the subsequent results remain valid for asymmetric $\hat{M}$.

Let d be the rank of M. Suppose the eigenvalues of M are arranged as

$$\lambda_1 \geqslant \cdots \geqslant \lambda_d > 0 = \lambda_{d+1} = \cdots = \lambda_p.$$

Let $v_1, \ldots, v_p$ be the corresponding eigenvectors. Then $Mv_i = \lambda_i v_i$ and $(v_1, \ldots, v_p)$ is an orthogonal matrix. Let $\mathcal{S}_1, \ldots, \mathcal{S}_\ell$ ($\ell \leqslant p$) be the eigenspaces of M corresponding to the distinct eigenvalues in descending order. For example, $\mathcal{S}_1$ corresponds to the largest eigenvalue, ..., $\mathcal{S}_{\ell-1}$ corresponds to the smallest nonzero eigenvalue, and $\mathcal{S}_\ell$ is the null space of M spanned by $\{v_{d+1}, \ldots v_p\}$. Similarly, we define $\{\hat{\lambda}_1, \ldots, \hat{\lambda}_p, \hat{v}_1, \ldots, \hat{v}_p\}$ and $\{\lambda^*_1, \ldots, \lambda^*_p, v^*_1, \ldots, v^*_p\}$ for $\hat{M}$ and M^*. For each $k < p$, let

$$B_k = (v_1, \ldots, v_k), \quad \hat{B}_k = (\hat{v}_1, \ldots, \hat{v}_k), \quad B^*_k = (v^*_1, \ldots, v^*_k).$$

Since B^*_k is repeatedly calculated for n bootstrap samples, we denote its realization at the ith bootstrap sample by $B^*_{k,i}$.

Define a function from $\{0, \ldots, p-1\}$ to $\mathbb{R}$ as

$$f^0_n(k) = \begin{cases} 0, & k = 0, \\ n^{-1}\sum_{i=1}^n \{1 - |\det(\hat{B}^{\mathrm{T}}_k B^*_{k,i})|\}, & k = 1, \ldots, p-1. \end{cases} \tag{3}$$

As in Ye & Weiss (2003), $1 - |\det(\hat{B}^{\mathrm{T}}_k B^*_{k,i})|$ is a number between 0 and 1 that measures the discrepancy between column spaces of $\hat{B}_k$ and $B^*_{k,i}$, with 1 representing the largest discrepancy. Therefore, $f^0_n(k)$ measures the variability of the bootstrap estimates $B^*_{k,1}, \ldots, B^*_{k,n}$ around the full-sample estimate $\hat{B}_k$. This is the bootstrap variability of eigenvectors we mentioned in §1. When $k = 0$, we define this variability to be 0. The range of f^0_n is also $[0, 1]$: it reaches 0 if each $B^*_{k,i}$ spans the same column space as $\hat{B}_k$, and 1 if each $B^*_{k,i}$ spans a space orthogonal to $\hat{B}_k$. We renormalize f^0_n to be

$$f_n : \{0, \ldots, p-1\} \to \mathbb{R}, \quad f_n(k) = f^0_n(k) \,/\, \{1 + \textstyle\sum_{i=0}^{p-1} f^0_n(i)\}, \tag{4}$$

where the constant 1 in the denominator is introduced to stabilize the performance of this criterion when $d = p - 1$.

Ye & Weiss (2003) speculated that f_n is large for $k > d$ and small for $k = d$. We now give a careful investigation of this intuition. First, consider the two scenarios:

(i) $\lambda_k > \lambda_{k+1}$, which implies that if $v_k \in \mathcal{S}_i$ then $v_{k+1} \in \mathcal{S}_{i+1}$. Then for every bootstrap sample we estimate the same k-dimensional subspace spanned by $\mathcal{S}_1, \ldots, \mathcal{S}_i$. The consistency of $\hat{M}$ then guarantees a small value of $f_n(k)$. In particular, this is true for $k = d$, because $\lambda_d > 0$ and $\lambda_{d+1} = 0$;

(ii) $\lambda_k = \lambda_{k+1}$, which implies that v_k and v_{k+1} reside in the same eigenspace, for example, $\mathcal{S}_i$. Then for each bootstrap sample we estimate the space spanned by $\mathcal{S}_1, \ldots, \mathcal{S}_{i-1}$ and an arbitrary proper subspace of $\mathcal{S}_i$. This arbitrariness causes large variation in the estimates, and consequently a large value of $f_n(k)$. In particular, this is the case for all $k > d$, because $\lambda_k = \lambda_{k+1} = 0$.

Because $\lambda_{d+1} = \cdots = \lambda_p = 0$, scenario (ii) applies to $k = d + 1, \ldots, p$. For $k = 1, \ldots, d$, either scenario may occur, depending on the set of equal values in $\{\lambda_1, \ldots, \lambda_d\}$. For example, if $\lambda_1 > \cdots > \lambda_d > 0$, then $f_n(1), \ldots, f_n(d)$ are small but $f_n(d + 1), \ldots f_n(p - 1)$ are large. If $\lambda_1 = \cdots = \lambda_d > 0$, then $f_n(1), \ldots, f_n(d - 1)$ are large, $f_n(d)$ is small, and $f_n(d + 1), \ldots, f_n(p - 1)$ are large. Hence f_n is always small at d and always has a jump at $d + 1$, regardless of how many of $\lambda_1, \ldots, \lambda_d$ are equal.

As M has rank d and $\hat{M}$ is consistent, $\hat{\lambda}_k$ drops to a small value at $k = d + 1$ from a large value at $k = d$. Thus, by carefully combining the bootstrap eigenvector variability $f_n(k)$ with the sample eigenvalue $\hat{\lambda}_k$, we can pinpoint the true rank d more precisely.

Parallel to f_n, we renormalize the sample eigenvalues and define the function

$$\phi_n : \{0, \ldots, p - 1\} \to \mathbb{R}, \quad \phi_n(k) = \hat{\lambda}_{k+1} / (1 + \textstyle\sum_{i=0}^{p-1} \hat{\lambda}_{i+1}), \tag{5}$$

where the constant 1 in the denominator is introduced to stabilize the performance of the criterion when $d = 0$. We shift the eigenvalues so that ϕ_n takes a small value at $k = d$ instead of at $k = d+1$. We define the objective function of our estimator as

$$g_n : \{0, \ldots, p - 1\} \to \mathbb{R}, \quad g_n(k) = f_n(k) + \phi_n(k), \tag{6}$$

which collects information from both the eigenvectors and the eigenvalues. In the spirit of information criteria, f_n can be viewed as the penalty term that increases the objective function after k reaches d. However, unlike with information criteria, this new penalty term is an intrinsic property of the random matrix $\hat{M}$. Moreover, since f_n is combined with ϕ_n in a scale-free manner, no tuning parameter is needed.

To sum up the above intuitions, below d the eigenvalue term ϕ_n is large; above d the eigenvector term f_n is large; and they are both small at d. Therefore, we expect g_n to reach its minimum approximately at d.

Note that $f_n^0(p)$ is identically 0 because both $\hat{B}_p$ and $B_{p,i}^*$ in (3) span $\mathbb{R}^p$. This causes the function $f_n^0(k)$ to bend downwards when k gets close to p, resulting in a shape similar to the handle of a ladle shown in Fig. 1(c). Moreover, as p increases, so does the denominator $1 + \sum_{i=0}^{p-1} f_n^0(i)$ of f_n, resulting in a smaller weight for the bootstrap component. Although these artifacts do not affect the consistency of the ladle estimator, in finite samples it makes sense to define and minimize g_n for k up to a certain integer $q < p - 1$, rather than go all the way to $k = p - 1$. As a rule of thumb, in most applications it is quite justifiable to assume $d \leqslant \lfloor p/\log(p) \rfloor$, and minimize $g_n(k)$

over $k \in \{0, \ldots, \lfloor p/\log(p) \rfloor\}$, where $\lfloor a \rfloor$ stands for the greatest integer less than or equal to a. Correspondingly, the range of the sums in the denominators in (4) and (5) should be changed to $\{0, \ldots, \lfloor p/\log(p) \rfloor\}$, which leads to

$$g_n(k) = f_n(k) + \phi_n(k) \equiv \frac{f_n^0(k)}{1 + \sum_{i=0}^{\lfloor p/\log(p) \rfloor} f_n^0(i)} + \frac{\hat{\lambda}_{k+1}}{1 + \sum_{i=0}^{\lfloor p/\log(p) \rfloor} \hat{\lambda}_{i+1}}. \tag{7}$$

Let $\mathcal{D}(f)$ denote the domain of a function f. We formally define the order-determination estimator as follows.

DEFINITION 1. *The ladle estimator of the rank d is*

$$\hat{d} = \arg\min \{g_n(k) : k \in \mathcal{D}(g_n)\}, \tag{8}$$

where g_n is defined by (6) if $p \leqslant 10$ and by (7) if $p > 10$.

The transition point $p = 10$ is chosen empirically: it works well in most of the examples we considered. The systematic choice of this transition point deserves further research. The scatterplot of $\{k, g_n(k)\}$, $k \in \mathcal{D}(g_n)$, is called the ladle plot. This plot can be used to assist order determination, as an alternative to the scree plot: we look for the minimizer, rather than the elbow, of the curve.

In the above developments, the bootstrap sample size is taken to be the sample size and the bootstrap sample is generated with replacement; see also Ye & Weiss (2003). This regime is not necessary for the asymptotic development, but the proofs in this paper are tailored for this regime, and would have to be modified slightly should we use other regimes. The bootstrap sample size affects the performance in principle, but we did not experience substantial changes in accuracy with bootstrap sample sizes $n/2$ or $2n$. Sampling with replacement is an important assumption for our asymptotic analysis. Specifically, the self-similarity condition, Assumption 2 in §3 depends on the fact that the relation between F_n^* and F_n resembles the relation between F_n and F, and sampling with replacement is key to this.

3. THEORETICAL CHARACTERIZATION OF THE EIGENVALUE-EIGENVECTOR PATTERN

We first introduce some regularity assumptions that will be used throughout the article.

Assumption 1. There is a random matrix $H(X, Y)$ with mean zero and finite second moment such that $\hat{M} = M + E_n H(X, Y) + o_P(n^{-1/2})$.

Assumption 2. The bootstrap estimator M^* satisfies

$$n^{1/2}\{\text{vech}(M^*) - \text{vech}(\hat{M})\} \to N(0, \text{var}_F[\text{vech}\{H(X, Y)\}]) \tag{9}$$

where vech$(\cdot)$ is the vectorization of the upper triangular part of a matrix and $\text{var}_F[\text{vech}\{H(X, Y)\}]$ is positive definite. The arrow $\to$ in (9) should be understood as convergence in distribution almost surely. That is, for almost every sequence S defined in the first paragraph of §2, the left-hand side converges in distribution to the right-hand side, where convergence in distribution is in terms of the conditional probability given S. See, for example, Bickel & Freedman (1981).

Assumption 3. For any sequence of nonnegative random variables $\{Z_n : n = 1, 2, \ldots\}$ involved hereafter, if $Z_n = O_P(c_n)$ for some sequence $\{c_n : n \in \mathbb{N}\}$ with $c_n > 0$, then $E(c_n^{-1} Z_n)$ exists for each n and $E(c_n^{-1} Z_n) = O(1)$.

Assumption 1 is quite mild: it is satisfied if the statistical functional M is Fréchet differentiable. See, for example, Bickel et al. (1993, p. 19), and Fernholz (1983). Because $n^{1/2}(\hat{M} - M)$ also converges in distribution to the right-hand side of (9), Assumption 2 amounts to asserting that the asymptotic behaviour of $n^{1/2}(M^* - \hat{M})$ mimics that of $n^{1/2}(\hat{M} - M)$. The validity of this self-similarity is discussed, for example, in Bickel & Freedman (1981), Parr (1985), Liu et al. (1989) and Gill (1989). We have given further justification, information, and intuition about this assumption in the Supplementary Material. Assumption 3 has been commonly used in the literature. These assumptions should not much restrict the applicability of our estimator.

Since $O_P(1)$ does not preclude convergence to zero, we need a more specific notation for a sequence of random variables bounded away from zero.

DEFINITION 2. *A sequence of random variables $\{Z_n\}$ is bounded below from zero in probability and written as $Z_n = O_P^+(1)$, if for any positive sequence $\{\epsilon_n\}$ such that $\epsilon_n = o(1)$,*

$$\lim_{n \to \infty} \mathrm{pr}(Z_n > \epsilon_n) = 1.$$

Furthermore, if $Z_n / c_n = O_P^+(1)$ for a positive sequence $\{c_n\}$, then we write $Z_n = O_P^+(c_n)$.

Roughly, this concept is the asymptotic analogue of a random variable Z taking positive values with probability 1. With this definition, the following theorem rigorously characterizes the eigenvalue-eigenvector pattern.

THEOREM 1. *Let $c_n = [\log\{\log(n)\}]^{-2}$. If Assumptions 1, 2 and 3 hold, and $M \in \mathbb{R}^{p \times p}$ is a positive semi definite matrix of rank $d \in \{0, \ldots, p-1\}$, then for any $k = 1, \ldots, p-1$, the following relation holds for almost every sequence $S = \{(X_n, Y_n) : n = 1, 2, \ldots\}$:*

$$f_n(k) = \begin{cases} O_P(n^{-1}), & \lambda_k > \lambda_{k+1}, \\ O_P^+(c_n), & \lambda_k = \lambda_{k+1}, \end{cases}$$

where the probability in O_P and O_P^+ is the conditional probability given S.

Although $\{c_n\}$ converges to zero, the convergence rate is very slow. For example, $\{\log(\log 10^4)\}^{-2} \approx 0{\cdot}2$. Hence, in practice $O_P^+(c_n)$ can be nearly treated as $O_P^+(1)$.

From this theorem, almost surely in the probability space of S, the bootstrap variability of eigenvectors is negligible whenever two consecutive eigenvalues are different, and is nonnegligible whenever they are equal.

4. CONSISTENCY OF THE LADLE ESTIMATOR

As mentioned in Assumption 2, we must be more nuanced when discussing convergence in the bootstrap context. Since ϕ_n is a function of the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, it is nonrandom given a sequence S. Bootstrap resampling introduces an extra layer of randomness, so $\hat{d}$, the minimizer of g_n, also has two layers of randomness. Reflecting this composite probability structure, the

consistency in the following theorem means that the event that the conditional probability of $\hat{d} = d$ given S converges to 1 happens almost surely in the probability space of S.

THEOREM 2. *Under Assumptions 1, 2 and 3, for any positive semidefinite matrix $M \in \mathbb{R}^{p\times p}$ of rank $d \in \{0, \ldots, p-1\}$, the ladle estimator (8) satisfies*

$$\mathrm{pr}\left\{\lim_{n\to\infty} \mathrm{pr}(\hat{d} = d \mid S) = 1\right\} = 1.$$

In the special case where all the nonzero eigenvalues of M are equal, Theorem 1 suggests a ladle shape for f_n itself. By applying the proof of Theorem 2, one can easily show that the minimizer of f_n is also consistent in this case. Nevertheless, it is still beneficial to incorporate ϕ_n, because it amplifies the downward trend of the curve $f_n(k)$ before k reaches d.

5. SIMULATION STUDIES

In this section, we compare the performances of the ladle estimator and some existing methods for order determination in a variety of settings: principal component analysis, canonical correlation analysis, independent component analysis, sufficient dimension reduction, and nonlinear sufficient dimension reduction. We consider one or two models for each setting, with each model giving rise to a matrix M whose rank is to be determined. From each model, we simulate n independent observations and determine the rank of M by different methods. This is repeated 1000 times and the percentages of correct estimation are reported in Tables 1 and 2. Four types of methods are included in the comparison: sequential tests, information criteria, the Ye–Weiss bootstrap estimator, and the ladle estimator. Sequential tests and information criteria may have different meanings under different settings, but they are based on the same general principles.

For the bootstrap estimator, Ye & Weiss (2003) proposed to use either the jump or the minimizer of $f_n(k)$ to estimate d, but did not give a criterion to decide which one to use, nor did they recommend a threshold to detect a jump. We experimented with both the jump and the minimizer, and found that the former performs better in our models, and therefore use it for the comparison. Specifically, let

$$\tau(\delta) = \delta \max\{f_n(k) : k \in \mathcal{D}(f_n)\},$$

where f_n is defined in (4) if $p \leqslant 10$ and in (7) otherwise, and $0 \leqslant \delta \leqslant 1$. We estimate d as the largest k such that $f_n(k) \leqslant \tau(\delta)$ if this happens to some k, and as the minimizer of $f_n(k)$ if $f_n(k) > \tau(\delta)$ for all k. We use $\delta = 0{\cdot}2, 0{\cdot}4, 0{\cdot}6$, to ensure that the comparison is not biased against the bootstrap estimator.

We sometimes need to distinguish between the indices for an observation in a sample and for a component in a random vector. For a generic random vector X, let $X_{(j)}$ denote the jth component of X; for a sample of observations on X, let X_i denote the ith observation and $X_{i(j)}$ denote the jth component of X_i. We now describe the models in detail. For all the sequential tests below, the significance levels are set at 0·05.

(i) Principal component analysis. Following (1), let M be the diagonal matrix whose $(1,1)$th entry is 2 and $(2,2)$th and $(3,3)$th entries are 1, and all the other entries are 0. Then M has

Table 1. *Comparison of order-determination methods at $p = 10$. The two sequential tests for SIR and DR are in the order described in the text. Entries in Columns 5–10 are percentages of correct order estimates in* 1000 *runs*

Setting	Model	n	d	ST	IC	YW1	YW2	YW3	Ladle
PCA	1	100	3	–	85	98	73	23	99
CCA	2	100	2	92	97	99	89	36	96
ICA	3	500	2	–	–	75	90	63	90
SDR-SIR	4	300	2	(78, 75)	0	48	95	86	91
SDR-DR	5	300	2	(79, 96)	4	87	89	27	98
NSDR	6	500	1	–	–	79	99	95	99

PCA, principal component analysis; CCA, canonical correlation analysis; ICA, independent component analysis; SDR, sufficient dimension reduction; SIR, sliced inverse regression; DR, directional regression; NSDR, nonlinear sufficient dimension reduction; ST, sequential test; IC, information criterion; YW1, YW2 and YW3, the bootstrap estimators with $\delta = 0{\cdot}2, 0{\cdot}4, 0{\cdot}6$; Ladle, the ladle estimator.

Table 2. *Comparison of order-determination methods at $p = 40$*

Method	Model	n	d	ST	IC	YW1	YW2	YW3	Ladle
PCA	1	200	3	–	99	99	99	48	99
CCA	2	200	2	91	100	99	100	96	100
ICA	3	1000	2	–	–	12	50	85	72
SDR-SIR	4	600	2	(0, 0)	0	20	97	97	90
SDR-DR	5	600	2	–	38	4	91	98	98
NSDR	6	100	1	–	–	98	97	58	96

The abbreviations and other specifications are in the legend of Table 1.

rank $d = 3$, with two of the nonzero eigenvalues equal to each other. We generate X from

$$\text{Model 1: } X \sim N(0, \Sigma_X), \quad \Sigma_X = M + 0{\cdot}54^2 I_p.$$

Let $\hat{\Sigma}_X$ be the sample variance matrix of X based on n independent observations from X, and let $\hat{M} = \hat{\Sigma}_X - \hat{\lambda}_p I_p$. Then it can be shown that $\hat{M} - M = E_n H(X) + o_P(n^{-1/2})$ for some random matrix $H(X) \in \mathbb{R}^{p \times p}$ with $E\{H(X)\} = 0$ and finite variance. We then apply the ladle estimator, the bootstrap estimator, and an information criterion in Bai & Ng (2002) called PC_{p1} to estimate d.

(ii) Canonical correlation analysis. We generate X from $N(0, I_p)$ and ε from $N(0, 0{\cdot}5^2 I_p)$ independently, and generate Y from

$$\text{Model 2: } Y_{(1)} = X_{(1)} + X_{(2)} + \varepsilon_{(1)},\ Y_{(2)} = X_{(3)} + \varepsilon_{(2)},\ Y_{(i)} = 2\,\varepsilon_{(i)} \text{ for } i = 3, \ldots, p.$$

Let V_X, V_Y and V_{XY} be the variance matrix of X, the variance matrix of Y and the covariance matrix of X and Y, respectively. Then the correlation matrix $M = V_X^{-1/2} V_{XY} V_Y^{-1} V_{XY}^{\mathrm{T}} V_X^{-1/2}$ has rank $d = 2$, with nonzero eigenvalues $8/9$ and $4/5$. We then construct $\hat{M}$ by replacing V_X, V_Y and V_{XY} with their sample estimates. Besides the ladle estimator and the bootstrap estimator, we apply the sequential test in Fujikoshi (1977) called $S_F(T_d^2)$ and the information criterion in Gunderson & Muirhead (1997) called $\hat{K}_{MC}$ to estimate the rank of M. See Caliński et al. (2006) for more information about $S_F(T_d^2)$.

(iii) Independent component analysis. We assume that X is generated from

$$\text{Model 3: } X = AU,$$

where $A \in \mathbb{R}^{p\times p}$. The random vector $U \in \mathbb{R}^p$ has independent components, some of which are Gaussian, which are regarded as noise, and some are non-Gaussian, which are regarded as signals. The goals of independent component analysis are (i) to identify the number of non-Gaussian components and (ii) to isolate these non-Gaussian components. As mentioned in §1, a commonly used method is the fourth-order blinded identification (Hyvärinen et al., 2001, Ch. 11), which is based on $\Lambda = E\{(Z^{\mathrm{T}}Z)ZZ^{\mathrm{T}}\} - (p+2)I_p$ where Z is the standardized X. The number of non-Gaussian components is the rank d of Λ, or equivalently the rank of $M = \Lambda^2$. We derive the sample versions $\hat{\Lambda}$ and $\hat{M}$ by replacing Z with $\hat{Z} = \hat{\Sigma}_X^{-1/2}\{X - \bar{X}\}$ and replacing the expectation in Λ with the sample mean. We choose A to be the matrix whose diagonal entries are 1 and off-diagonal entries are 0·5. We generate $U_{(1)}$ and $U_{(2)}$ from the exponential distribution with mean 1, and $U_{(3)}, \ldots, U_{(p)}$ from $N(0,1)$. Thus $d = 2$ and the two nonzero eigenvalues of M are equal. We apply the ladle estimator and the bootstrap estimator to estimate d.

(iv) Sufficient dimension reduction. We generate X from $N(0, I_p)$ and ε from $N(0, 0{\cdot}5^2)$ independently; we generate Y from

$$\text{Model 4: } Y = X_{(1)} / \{0{\cdot}5 + (1{\cdot}5 + X_{(2)})^2\} + \varepsilon,$$

$$\text{Model 5: } Y = X_{(1)}^2 + X_{(2)} + \varepsilon,$$

where Model 4 was used in Li (1991). We then apply the sliced inverse regression to Model 4 and directional regression to Model 5. In each case, the candidate matrix has rank $d = 2$ and the two nonzero eigenvalues are unequal. Besides the ladle estimator and the bootstrap estimator, we also apply the sequential tests proposed by Bura & Yang (2011), including the weighted chi-square test and the Wald-type chi-square test, and the Bayesian information criterion proposed by Zhu et al. (2006). For the latter, we follow the authors' suggestion to fix the number of slices in the dimension reduction methods at $H = n/20$, and use the tuning parameter $C_n = H\{0{\cdot}5\log(n) + 0{\cdot}1n^{1/3}\}/(2n)$. From a simulation study not presented here, the sequential tests are more sensitive to the number of slices H than the other order-determination procedures. We fixed H at 10 for sliced inverse regression and at 3 for directional regression, which are favourable to the sequential tests.

(v) Nonlinear sufficient dimension reduction. The idea of nonlinear sufficient dimension reduction is, intuitively, to first map the predictor X to a higher-dimensional feature space, perform dimension reduction there, and then map the result back to the original space (Wu, 2008; Yeh et al., 2009; Li et al., 2011; Lee et al., 2013). We generate (X, ε) in the same way as in the case of sufficient dimension reduction, and generate Y from

$$\text{Model 6: } Y = \sin\{(X_{(1)}^2 + X_{(2)}^2)/3\} + 0{\cdot}6\,\varepsilon.$$

In sufficient dimension reduction, the central dimension reduction subspace is of dimension 2, spanned by $(1,0,0,\ldots,0)^{\mathrm{T}}$ and $(0,1,0,\ldots,0)^{\mathrm{T}}$, but in nonlinear sufficient dimension reduction, the central dimension reduction σ-field (Lee et al., 2013) is generated by the nonlinear function $X_{(1)}^2 + X_{(2)}^2$. We use the feature mapping

$$\Phi = \varphi(X) = \{1, X_{(k)}, X_{(i)}X_{(j)},\ i \leqslant j,\ i,j,k = 1,\ldots,p\},$$

which is a $\{1+p+p(p+1)/2\}$-dimensional vector. The theory of Lee et al. (2013) implies that, in this case, the central σ-field is complete and sufficient, and hence can be recovered by the range space of $M = \mathrm{var}\{E(\Phi|Y)\}$, which is of rank $d = 1$.

We divide the sample of responses $Y_1, \ldots, Y_n$ evenly into m slices, say $J_1, \ldots, J_m$, and discretize Y_j as $\tilde{Y}_j$, which takes value k if $Y_j \in J_k$. Let $\Phi_j = \varphi(X_j)$, and let

$$E_n(\Phi|\tilde{Y} = k) = (m/n)\textstyle\sum_{Y_j \in J_k} \Phi_j, \quad E_n\Phi = n^{-1}\sum_{j=1}^{n} \Phi_j,$$

$$\hat{M} = \mathrm{var}_n\{E_n(\Phi|\tilde{Y})\} = E_n[\{E_n(\Phi|\tilde{Y}) - E_n\Phi\}\{E_n(\Phi|\tilde{Y}) - E_n\Phi\}^{\mathrm{T}}].$$

We then apply the ladle estimator and the bootstrap estimator to $\hat{M}$ to estimate d.

We first set $p = 10$ for all the models. Because different statistical problems often require different sample sizes to have reasonable accuracy, for example, higher sample moments require larger sample sizes to converge, we let $n = 100$ in Models 1 and 2, $n = 300$ in Models 4 and 5, and $n = 500$ in Models 3 and 6. The percentages of correct order determination from 1000 simulated samples are reported in Table 1.

Table 1 shows that the ladle estimator is clearly best overall. The sequential tests are generally consistent, but fail to reach the nominal significance level. The bootstrap estimator performs well when $\delta = 0{\cdot}4$, but the optimal choice of δ varies with the model.

Next, we increase the dimension to $p = 40$ and the sample size to $n = 200$ in Models 1 and 2, and $n = 600$ in Models 4 and 5, and $n = 1000$ in Models 3 and 6. At this dimension, the sample eigenvalues deviate substantially from their asymptotic distributions for $k > d$. This causes the sequential tests for sliced inverse regression to perform poorly, as shown in Table 2. However, the procedures that incorporate eigenvector variations, such as the bootstrap estimator and the ladle estimator, remain accurate. The sequential tests for the directional regression at the current dimension are computationally very expensive and are omitted.

6. Wine cultivar identification

We now apply the ladle estimator in conjunction with directional regression to the wine cultivar data in Forina et al. (1988), which consist of 178 wine samples from three different cultivars. For each sample, the name of the cultivar and 13 covariates are recorded; alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, oroanthocyanins, color intensity, hue, OD280/OD315 of diluted wines and proline. The goal is to identify the cultivars based on the covariates. We use directional regression to find a low-dimensional predictor and identify the cultivars in the reduced space. The order-determination methods would then tell us the appropriate dimension of the reduced predictor. To satisfy the requirements by directional regression, we take the logarithms of malic acid, color intensity and proline, take the reciprocal of magnesium, and then standardize each component of the modified predictor. For details about these requirements, see Li & Wang (2007).

Figure 2(a) shows the scree plot, Fig. 2(b) the bootstrap variability plot, and Fig. 2(c) the ladle plot. Figure 2(d) is the scatterplot of the first two components of the reduced predictor found by directional regression, $(\hat{\beta}_1^{\mathrm{T}}X, \hat{\beta}_2^{\mathrm{T}}X)$, standardized to have mean zero and identity variance matrix. The bootstrap estimator estimates d to be 2 for $\delta = 0{\cdot}2, 0{\cdot}4$ and $0{\cdot}6$, respectively; the ladle estimator also yields $\hat{d} = 2$.

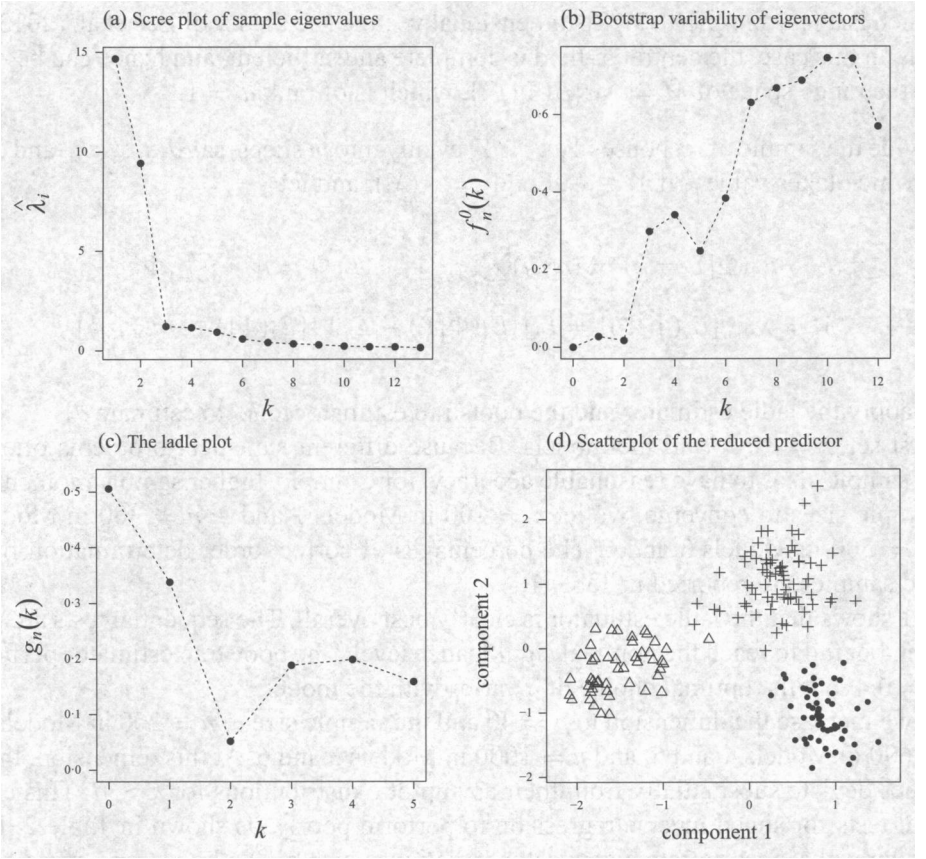

Fig. 2. Wine cultivar analysis. (a) the scree plot of sample eigenvalues; (b) the bootstrap variability of eigenvectors; (c) the ladle plot; (d) the scatterplot of the reduced predictor indexed by cultivars: ($\bullet$, $+$, $\triangle$) represent data from cultivar 1, 2, 3 correspondingly.

To see that $\hat{d} = 2$ is a reasonable estimate, note that the first two components of the reduced predictor from directional regression give nearly perfect separation of the three cultivars, suggesting that two directions are sufficient to identify the cultivars; that is, the dimension d of the central subspace is less than or equal to 2. Both axes in the scatterplot offer significant separations of the three groups, suggesting that the dimension d is at least 2. Thus $\hat{d} = 2$ is a good choice.

ACKNOWLEDGEMENT

We are grateful to reviews for their insightful comments and suggestions, which have helped us greatly in improving an earlier manuscript. This research was supported in part by National Science Foundation grants awarded to Bing Li and Naomi Altman at the Pennsylvania State University.

REFERENCES

BAI, J. & NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191–221.
BICKEL, P. & FREEDMAN, D. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9**, 1196–217.

BICKEL, P., KLAASSEN, C., RITOV, Y. & WELLNER, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: John Hopkins University Press.

BURA, E. & YANG, J. (2011). Dimension estimation in sufficient dimension reduction: a unifying approach. *J. Mult. Anal.* **102**, 130–42.

CALIŃSKI, T., KRZYŚKO, M. & WOLYŃSKI, W. (2006). A comparison of some tests for determining the number of nonzero canonical correlations. *Commun. Statist. B* **35**, 727–49.

COOK, R. D. (1994). Using dimension reduction subspaces to identify important imputs in models of physical systems. *In 1994 Proceedings of the Section on Physical and Engineering Sciences: American Statistical Association, Alexandria, VA.*, 18–25.

COOK, R. D. (1998). *Regression Graphics*. New York: Wiley.

COOK, R. D. & LI, B. (2002). Dimension reduction for conditional mean in regression. *Ann. Statist.* **30**, 455–74.

COOK, R. D. & LI, B. (2004). Determining the dimension of iterative Hessian transformation. *Ann. Statist.* **32**, 2501–31.

COOK, R. D. & WEISBERG, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *J. Am. Statist. Assoc.* **86**, 316–42.

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

FORINA, M., LEARDI, R., ARMANINO, C. & LANTERI, S. (1988). *PARVUS - An extendable package of programs for data exploration, classification and correlation*. Amsterdam: Elsevier.

FUJIKOSHI, Y. (1977). Asymptotic expansion for the distributions of some multivariate tests. *In: Krishnaiah, P. R., ed. Multivariate Analysis. Vol. IV.* Amsterdam: North-Holland, pp. 55–71.

GILL, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method, part 1. *Scand. J. Statist.* **16**, 97–128.

GUNDERSON, B. & MUIRHEAD, R. (1997). On estimating the dimensionality in canonical correlation analysis. *J. Mult. Anal.* **62**, 121–36.

HYVÄRINEN, A., KARHUNEN, J. & OJA, E. (2001). *Independent Component Analysis*. New York: Wiley.

JOLLIFFE, I. T. (2002). *Principal Component Analysis, Second Edition*. New York: Springer.

LEE, K.-Y., LI, B. & CHIAROMONTE, F. (2013). A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. *Ann. Statist.* **41**, 221–49.

LI, B., ARTEMIOU, A. & LI, L. (2011). Principal support vector machines for linear and nonlinear sufficient dimension reduction. *Ann. Statist.* **39**, 3182–210.

LI, B. & WANG, S. (2007). On directional regression for dimension reduction. *J. Am. Statist. Assoc.* **35**, 2143–72.

LI, B., ZHA, H. & CHIAROMONTE, F. (2005). Contour regression: a general approach to dimension reduction. *Ann. Statist.* **33**, 1580–616.

LI, K. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *J. Am. Statist. Assoc.* **87**, 1025–39.

LI, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Am. Statist. Assoc.* **86**, 316–42.

LIU, R. Y., SINGH, K. & LO, S. H. (1989). On a representation related to the bootstrap. *Sankhya* **51**, 168–77.

PARR, W. (1985). The bootstrap: some large sample theory and connections with robustness. *Statist. Prob. Lett.* **3**, 97–100.

SCHOTT, J. R. (1994). Determining the demensionality in sliced inverse regression. *J. Am. Statist. Assoc.* **89**, 141–8.

WU, H. M. (2008). Kernel sliced inverse regression with applications on classification. *J. Comp. Graph. Statist.* **17**, 590–610.

YE, Z. & WEISS, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *J. Am. Statist. Assoc.* **98**, 968–79.

YEH, Y.-R., HUANG, S.-Y. & LEE, Y.-Y. (2009). Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Trans. Knowledge Data Eng.* **21**, 1590–603.

YIN, X., LI, B. & COOK, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *J. Mult. Anal.* **99**, 1733–57.

ZHU, L., MIAO, B. & PENG, H. (2006). On sliced inverse regression with high-dimensional covariates. *J. Am. Statist. Assoc.* **101**, 630–42.

[*Received November* 2013. *Revised August* 2016]