Biometrika (2015), **102**, 2, *pp.* 409–420
Printed in Great Britain
doi: 10.1093/biomet/asv004
Advance Access publication 20 April 2015

A validated information criterion to determine the structural dimension in dimension reduction models

BY YANYUAN MA

Department of Statistics, University of South Carolina, Columbia, South Carolina 29208, U.S.A.
yanyuan.ma@stat.sc.edu

AND XINYU ZHANG

International School of Economics and Management, Capital University of Economics and Business, Beijing 100070, China
xinyu@amss.ac.cn

SUMMARY

A crucial component of performing sufficient dimension reduction is to determine the structural dimension of the reduction model. We propose a novel information criterion-based method for this purpose, a special feature of which is that when examining the goodness-of-fit of the current model, one needs to perform model evaluation by using an enlarged candidate model. Although the procedure does not require estimation under the enlarged model of dimension $k+1$, the decision as to how well the current model of dimension k fits relies on the validation provided by the enlarged model; thus we call this procedure the validated information criterion, VIC(k). Our method is different from existing information criterion-based model selection methods; it breaks free from dependence on the connection between dimension reduction models and their corresponding matrix eigenstructures, which relies heavily on a linearity condition that we no longer assume. We prove consistency of the proposed method, and its finite-sample performance is demonstrated numerically.

Some key words: Dimension reduction; Estimating equation; Information criterion; Linearity condition; Model selection; Penalization; Structural dimension.

1. INTRODUCTION

Consider a p-dimensional covariate vector X and a univariate response variable Y. If the regression relation between Y and X is unspecified, the curse of dimensionality arises for large or even moderate p, and impairs the performance of familiar nonparametric methods. Sufficient dimension reduction alleviates this problem by assuming that Y is linked to X through d linear combinations of X, denoted by $\beta^{\mathrm{T}}X$. Here $\beta \in \mathbb{R}^{p\times d}$ and d is typically much smaller than p. For example, in the central space framework, it is assumed that the distribution of Y conditional on X, $f_{Y|X}(y, x)$, is a function of y and $\beta^{\mathrm{T}}x$ only. In the central mean space framework $E(Y \mid X = x)$ is assumed to be a function of $\beta^{\mathrm{T}}x$ only, and in the central variance space framework $\mathrm{var}(Y|x)$ is assumed to be a function of $\beta^{\mathrm{T}}x$ only. Cook (1998) established that except for some degenerate cases, the smallest dimension reduction space exists and is unique, and

the corresponding smallest d is defined as the structural dimension of the sufficient dimension reduction space.

An important aspect of dimension reduction is to determine the structural dimension d of the sufficient dimension reduction space. The most popular approach converts this to a problem of deciding how many eigenvalues of a matrix are nonzero, but it is applicable only to inverse regression-based estimation procedures. Methods that have been proposed to determine the number of nonzero eigenvalues of a matrix include sequential testing (Li, 1991, 1992; Schott, 1994; Velilla, 1998; Bura & Cook, 2001; Cook & Yin, 2001; Cook & Li, 2002, 2004; Cook & Ni, 2005; Bura & Yang, 2011), bootstrapping (Ye & Weiss, 2003; Zhu & Zeng, 2006; Zeng, 2008), the Bayesian information criterion, BIC (Zhu et al., 2006; Zhu & Zhu, 2007; Luo et al., 2009), and sparse eigendecomposition (Zhu et al., 2010). However, the equivalence of dimension reduction and matrix eigendecomposition relies critically on the linearity condition and/or constant variance condition (Ma & Zhu, 2012).

Another approach is leave-one-out crossvalidation (Xia et al., 2002), whereby one chooses a dimension d that minimizes the leave-one-out average prediction error. This approach inevitably requires estimation of the unspecified regression function, which can be the conditional density, conditional mean or conditional variance function, depending on the specific problem. Thus, the method is applicable only to nonparametric estimators in dimension reduction.

In the semiparametric estimation framework, where nonparametric estimation of the regression function is avoided and the link to matrix eigendecomposition is also given up due to relaxation of the conditions on the covariates X, the only existing method to determine the structural dimension is the bootstrap (Ma & Zhu, 2012). However, its theoretical properties have not been studied thoroughly and its validity has not yet been established.

The goal of this paper is to develop a method of selecting the structural dimension d without requiring special structures on the covariate vector X and nonparametric estimation procedures other than those needed for the original estimation of the reduction space. Following Ma & Zhu (2012, 2013), once a fixed parameterization of the dimension reduction space is adopted, all the root-n-consistent estimators of β can be obtained from an estimating equation. Hence it is very tempting to add penalty terms to the estimating equations to shrink some components of β to zero. However, this does not work because of the special structure imposed on β as a result of the identifiability requirement. For example, in the widely used orthogonal parameterization, all the columns of β have unit length, so it is impossible to shrink all the elements of any column of β to zero, and hence there is no way to modify the working dimension of β through penalization. Likewise, if we use the parameterization of Ma & Zhu (2013), which fixes the upper block of β to be the identity, then the working dimension is also fixed even if all the remaining elements of a column are shrunk to zero. Thus, the parameter structure is rigid, and the parameterization under a larger model automatically excludes the possibility of a smaller model. Unless the functional dependence is examined, one cannot evaluate whether a certain column of β is needed in the model, certainly not by examining the elements in the estimated $\hat{\beta}$ only.

Methods based on classical criteria such as AIC and BIC are also out of the question, since we do not want to estimate the likelihood. A natural modification of these methods involves replacing the loglikelihood with a pseudo-loglikelihood, constructed from the weighted square of the estimating equation. An inherent difficulty with this approach is that it is unclear whether the value of such a quadratic form really reflects the goodness-of-fit of the underlying model. To illustrate this point, let us examine the simple case where $Y = X_1 + X_2^2 + \epsilon$ and the estimating equation is $n^{-1/2}\sum_{i=1}^{n}(Y_i - \beta_1 X_{1i} - \beta_2 X_{2i}^2)(X_{1i}, X_{2i}^2)^{\mathrm{T}} = 0$. Obviously, the ordinary least-squares estimator $(\hat{\beta}_1, \hat{\beta}_2)$ converges to the true parameter value $\beta_1 = \beta_2 = 1$ as $n \to \infty$,

and the quadratic form

$$n^{-1}\left\{\sum_{i=1}^{n}(Y_i-\hat\beta_1X_{1i}-\hat\beta_2X_{2i}^2)(X_{1i},X_{2i}^2)\right\}\left\{\sum_{i=1}^{n}(Y_i-\hat\beta_1X_{1i}-\hat\beta_2X_{2i}^2)\begin{pmatrix}X_{1i}\\X_{2i}^2\end{pmatrix}\right\}$$

is always zero regardless of the sample size. Now if we fit an insufficient model $Y=X_1\beta_1+\epsilon$ and estimate β_1 from $\hat\beta_1=(\sum_{i=1}^n X_{1i}Y_i)/(\sum_{i=1}^n X_{1i}^2)$, then $\hat\beta_1$ converges to $\beta_1=E(X_1Y)/E(X_1^2)$ as $n\to\infty$, while the corresponding quadratic form $n^{-1}\{\sum_{i=1}^n(Y_i-\hat\beta_1X_{1i})X_{1i}\}^2$ remains zero for all n. Thus, regardless of which of the two models is used, the quadratic forms have the same performance. In other words, the quadratic form of the estimating equations does not reflect how well a model fits the data.

The difficulty with the penalization method arises from the fact that for sufficient dimension reduction, a model under a smaller structural dimension is not a special case of a general model under a larger structural dimension that corresponds to having some zero parameters. Whether or not an additional dimension is needed is influenced by the functional dependence of Y on the additional direction. The effect of this on the penalization method is that dimension determination cannot be transformed to evaluation of the number of nonzero coefficients. The difficulty with using a criterion-based method results from the fact that the estimating equation is adaptive to the working structural dimension and hence does not reflect the model's goodness-of-fit. Therefore, a new method for determining the structural dimension d is needed.

2. Validated information criterion

Let p be the number of covariates and d the true structural dimension. Let S_k be the estimated dimension reduction space with working structural dimension k, and let S_{+1} be a nonstochastic expansion of S_k to a subspace of dimension $k+1$, so that $S_k\subset S_{+1}$. In general, $S_{+1}\neq S_{k+1}$. When $k+1>d$, both S_{+1} and S_{k+1} contain S_d, so the additional directions are redundant. However, when $k+1\leqslant d$, all the directions in S_{k+1} and S_{+1} are essential, and the two spaces will have different properties. We aim to bring out this difference between $k+1\leqslant d$ and $k+1>d$, and hence use it to identify d. To avoid unnecessary ambiguity, we use exclusively the parameterization of β such that the upper block of β is the identity matrix.

Consider the working structural dimension k, and write $\beta_{(k)}$ for the corresponding $p\times k$ parameter matrix. Let $O=(X,Y)$. Following Ma & Zhu (2012), all the estimating functions in dimension reduction models consist of

$$f(O,\beta_{(k)})=\mathrm{vec}\big(\big[a(x)-E\big\{a(x)\mid\beta_{(k)}^{\mathrm T}x\big\}\big]\big[g(Y,\beta_{(k)}^{\mathrm T}x)-E\big\{g(Y,\beta_{(k)}^{\mathrm T}x)\mid\beta_{(k)}^{\mathrm T}x\big\}\big]\big)\qquad(1)$$

or linear combinations of $f(O,\beta_{(k)})$ corresponding to different a and g. Here vec stands for vectorization.

The practical implementation of (1) involves estimating $\beta_{(k)}$ via

$$\hat\beta_{(k)}=\underset{\beta_{(k)}}{\arg\min}\ n^{-1}\left\{\sum_{i=1}^{n}\hat f(O_i,\beta_{(k)})\right\}^{\mathrm T}W\left\{\sum_{i=1}^{n}\hat f(O_i,\beta_{(k)})\right\},\qquad(2)$$

where $\hat f(O,\beta_{(k)})=\mathrm{vec}([a(x)-\hat E\{a(x)\mid\beta_{(k)}^{\mathrm T}x\}][g(Y,\beta_{(k)}^{\mathrm T}x)-\hat E\{g(Y,\beta_{(k)}^{\mathrm T}x)\mid\beta_{(k)}^{\mathrm T}x\}])$, with the $\hat E(\cdot\mid\beta_{(k)}^{\mathrm T}x)$ being nonparametric kernel estimates of the corresponding quantities, and W is a weight matrix. If one aims to minimize estimation variability, the optimal choice of W is the inverse of $\mathrm{var}\{f(O,\beta_{(k)})\}$. In the following derivation, we assume that a decision on a suitable

W has been made and we treat W as fixed. For simplicity of presentation, we consider only $f(O, \beta_{(k)})$ given in (1); treatment of a linear combination of several such $f(O, \beta_{(k)})$, of the form $\sum_{j=1}^{m} \tau_j \mathrm{vec}([a_j(x) - E\{a_j(x) \mid \beta_{(k)}^{\mathrm{T}} x\}][g_j(Y, \beta_{(k)}^{\mathrm{T}} x) - E\{g_j(Y, \beta_{(k)}^{\mathrm{T}} x) \mid \beta_{(k)}^{\mathrm{T}} x\}])$ at fixed τ_j, is essentially identical. Let

$$\beta_{(k)}^0 = \arg\min_{\beta_{(k)}} E\{f(O, \beta_{(k)})\}^{\mathrm{T}} W E\{f(O, \beta_{(k)})\}, \tag{3}$$

which is the true parameter value when $k = d$ and can be thought of as an optimal parameter value when $k \neq d$.

We now perform subspace expansion via v. For any k that satisfies $0 < k < p$, write

$$\beta_{(k)}^0 = \begin{pmatrix} I_k \\ \beta_{(k)}^{0,\mathrm{U}} \\ \beta_{(k)}^{0,\mathrm{L}} \end{pmatrix}, \quad \hat{\beta}_{(k)} = \begin{pmatrix} I_k \\ \hat{\beta}_{(k)}^{\mathrm{U}} \\ \hat{\beta}_{(k)}^{\mathrm{L}} \end{pmatrix},$$

where $\beta_{(k)}^{0,\mathrm{U}}$ and $\hat{\beta}_{(k)}^{\mathrm{U}}$ are $1 \times k$ vectors while $\beta_{(k)}^{0,\mathrm{L}}$ and $\hat{\beta}_{(k)}^{\mathrm{L}}$ are $(p - 1 - k) \times k$ matrices. For any $(p - 1 - k) \times 1$ vector v, let

$$\gamma_{(k)}^0(v) = \begin{pmatrix} I_k & 0_{k\times 1} \\ 0_{1\times k} & 1 \\ \beta_{(k)}^{0,\mathrm{L}} - v\beta_{(k)}^{0,\mathrm{U}} & v \end{pmatrix}, \quad \tilde{\gamma}_{(k)}(v) = \begin{pmatrix} I_k & 0_{k\times 1} \\ 0_{1\times d} & 1 \\ \hat{\beta}_{(k)}^{\mathrm{L}} - v\hat{\beta}_{(k)}^{\mathrm{U}} & v \end{pmatrix}.$$

It is easy to verify that for any v,

$$\gamma_{(k)}^0(v) \begin{pmatrix} I_k \\ \beta_{(k)}^{0,\mathrm{U}} \end{pmatrix} = \beta_{(k)}^0, \quad \tilde{\gamma}_{(k)}(v) \begin{pmatrix} I_k \\ \hat{\beta}_{(k)}^{\mathrm{U}} \end{pmatrix} = \hat{\beta}_{(k)},$$

so the space spanned by the columns of $\gamma_{(k)}^0(v)$ or $\tilde{\gamma}_{(k)}(v)$ contains that spanned by the columns of $\beta_{(k)}^0$ or $\hat{\beta}_{(k)}$, respectively. In other words, we can view $\gamma_{(k)}^0(v)$ or $\tilde{\gamma}_{(k)}(v)$ as an arbitrary way of expanding $\beta_{(k)}^0$ or $\hat{\beta}_{(k)}$. The expansion is described by v.

We now consider the case where $k = d$. Ma & Zhu (2012) established that $\hat{\beta}_{(d)}$ is a root-n-consistent estimator of the true parameter value $\beta_{(d)}^0$. If we expand the model to $k + 1$, we get a redundant model. Thus, intuitively, we can pick an arbitrary v to form $\tilde{\gamma}_{(k)}(v)$ and we will still be able to obtain that

$$n^{-1} \sum_{i=1}^{n} f\{O_i, \tilde{\gamma}_{(k)}(v)\} \to E[f\{O, \gamma_{(k)}^0(v)\}] = 0$$

in probability as $n \to \infty$. Similar observations apply when $k > d$. Specifically, having obtained $\hat{\beta}_{(k)}$ from an estimating equation under the working structural dimension k, we can arbitrarily expand $\hat{\beta}_{(k)}$ to form $\tilde{\gamma}_{(k)}(v)$. Because the expanded model is redundant and the added direction has no effect, we would always have $n^{-1} \sum_{i=1}^{n} f\{O_i, \tilde{\gamma}_{(k)}(v)\} \to 0$ in probability for any v. The similarity between the observations concerning $k = d$ and $k > d$ indicates that of all the models corresponding to working structural dimensions k such that $k \geqslant d$, we should select the simplest one.

However, for a model with working structural dimension $k < d$, the situation is quite different: the dimension reduction assumption is not satisfied under the restrictive model. Now consider

$k+1$. Since $k+1 \leqslant d$, this assumption may or may not be satisfied under the expanded model, and even if $k+1=d$ and the assumption can be satisfied, it is satisfied only for a specific $\gamma^0_{(k)}(v_0)$. Thus, in general, $n^{-1}\sum_{i=1}^n f(\{O_i, \tilde{\gamma}_{(k)}(v)\}$ will not converge to zero when we insert an arbitrary v.

The contrast between the above two cases will allow us to determine d via a special information criterion. We will use the enlarged model of structural dimension $k+1$ to provide a validation of the goodness-of-fit of the current model with structural dimension k. Specifically, let

$$\text{VIC}(k) = \frac{1}{2}\left[\left\|n^{-1/2}\sum_{i=1}^n \hat{f}\{O_i, \tilde{\gamma}_{(k)}(v_1)\}\right\|^2 + \left\|n^{-1/2}\sum_{i=1}^n \hat{f}\{O_i, \tilde{\gamma}_{(k)}(v_2)\}\right\|^2\right] + pk\log(n). \tag{4}$$

To form the penalty term in (4), we use pk instead of the true number of parameters in β in the candidate model $(p-k)k$. This ensures that the penalty term increases with the model complexity indicated by k. In (4), $\tilde{\gamma}_{(k)}(v_1)$ and $\tilde{\gamma}_{(k)}(v_2)$ are two matrices from $\hat{\beta}_{(k)}$ corresponding to two different choices of the vector v. Because the difference in $\text{VIC}(k)$ between using different v vectors does not change the order of $\text{VIC}(k)$ for $k<d$ and does not affect the leading term of $\text{VIC}(k)$ when $k \geqslant d$, as will be shown in § 3, the method is not very sensitive to the choice of v_1 and v_2; in practice, we simply set v_1 to be the zero vector and v_2 to be the vector of ones. The computation of $\text{VIC}(k)$ is not much more difficult than the usual practice, in which all calculations are conducted under the fixed candidate model with working structural dimension k; this is because, although we need an enlarged model of working structural dimension $k+1$ to calculate the criterion, we never perform estimation in the enlarged model. According to the above analysis, as $n\to\infty$, when $k=d$ we have $\text{VIC}(k) = pd\log(n) + O_p(1)$; when $k>d$ we have $\text{VIC}(k) \geqslant pk\log(n)$, which is larger than $\text{VIC}(d)$ with probability approaching one. When $k<d$, $\text{VIC}(k) = cn + o_p(n) + pk\log(n)$ where c is a positive constant, and this is also larger than $\text{VIC}(d)$ with probability approaching one. Thus, $\text{VIC}(k)$ is minimized at $k=d$ and so, by minimizing $\text{VIC}(k)$, we can choose d consistently.

3. Selection consistency

To show consistency, we need some regularity conditions. For any matrix θ, let $\nu(\theta)$ denote the number of columns of θ. We assume that the following conditions hold for $\theta \in \{\beta^0_{(1)}, \ldots, \beta^0_{(d)}, \gamma^0_{(1)}, \ldots, \gamma^0_{(d)}\}$.

Condition 1. The univariate kernel function $K(\cdot)$ is Lipschitz, has compact support, and satisfies

$$\int K(u)\,\mathrm{d}u = 1, \quad \int u^i K(u)\,\mathrm{d}u = 0 \quad (1 \leqslant i \leqslant m_\theta - 1), \quad 0 \neq \int u^{m_\theta} K(u)\,\mathrm{d}u < \infty,$$

where m_θ is a positive integer such that $m_\theta > \nu(\theta)/2$. The $\nu(\theta)$-dimensional kernel function is a product of $\nu(\theta)$ univariate kernel functions; that is, $K_h(u) = h^{-\nu(\theta)}K(u/h) = \prod_{i=1}^{\nu(\theta)} K_h(u_i) = h^{-\nu(\theta)}\prod_{i=1}^{\nu(\theta)} K(u_i/h)$ for $u=(u_1,\ldots,u_{\nu(\theta)})^{\mathrm{T}}$.

Condition 2. Let $r_1(\theta^{\mathrm{T}}x) = E\{a(x) \mid \theta^{\mathrm{T}}x\}f(\theta^{\mathrm{T}}x)$ and $r_2(\theta^{\mathrm{T}}x) = E\{g(Y,\theta^{\mathrm{T}}x) \mid \theta^{\mathrm{T}}x\}f(\theta^{\mathrm{T}}x)$. The m_θth derivatives of $r_1(\theta^{\mathrm{T}}x)$, $r_2(\theta^{\mathrm{T}}x)$ and $f(\theta^{\mathrm{T}}x)$ are locally Lipschitz-continuous.

Condition 3. The probability density or mass functions of x and $\theta^{\mathrm{T}}x$, denoted by $f_x(x)$ and $f(\theta^{\mathrm{T}}x)$, respectively, are bounded and bounded away from zero; moreover, each entry in the matrices $E\{a(x)a(x)^{\mathrm{T}} \mid \theta^{\mathrm{T}}x\}$ and $E\{g(Y, \theta^{\mathrm{T}}x)g^{\mathrm{T}}(Y, \theta^{\mathrm{T}}x) \mid \theta^{\mathrm{T}}x\}$ is a locally Lipschitz-continuous and bounded function of $\theta^{\mathrm{T}}x$.

Condition 4. The bandwidth h is such that $h = O(n^{-\kappa_\theta})$, where κ_θ is a positive constant that satisfies $(4m_\theta)^{-1} < \kappa_\theta < \{2\nu(\theta)\}^{-1}$,

Condition 5. For $0 < k \leqslant d$, the equation $E\{f(O, \beta_{(k)})\} = 0$ has at most one solution.

Condition 1 comprises standard requirements on the multivariate kernel function in nonparametric kernel estimation. Conditions 2 and 3 contain the smoothness and boundedness requirements on the low-dimensional density functions and regression functions which allow mathematical operations such as differentiating, exchanging order of operations etc. to be performed. Condition 4 is the minimum requirement for ensuring the root-n convergence rate in estimating β. Condition 5 is a standard condition on the set of estimating equations to ensure that the estimation procedure does not degenerate.

THEOREM 1. *Let* $\hat{d} = \arg\min_k \mathrm{VIC}(k)$, *where* $\mathrm{VIC}(k)$ *is as defined in* (4). *Under Conditions* 1–5, *as* $n \to \infty$, $\mathrm{pr}(\hat{d} = d) \to 1$.

The proof of Theorem 1 is given in the Appendix. Condition 5 can be further relaxed to allow $E\{f(O, \beta)\} = 0$ to have finitely many, say r, solutions. When the solution is not unique, we set

$$\mathrm{VIC}(k) = \frac{1}{r+1} \sum_{j=1}^{r+1} \left\| n^{-1/2} \sum_{i=1}^{n} \hat{f}\{O_i, \tilde{\gamma}_{(k)}(v_j)\} \right\|^2 + pk \log(n),$$

which still allows us to select the structural dimension consistently.

We expand the directions in $\beta_{(k)}$ in two different directions, v_1 and v_2, instead of a single direction described by v, so that our method remains valid even if the zero-probability event $\gamma^0_{(k)}(v) = \beta^0_{k+1}$ occurs. Similarly, when $E\{f(O, \beta)\} = 0$ has r solutions, say $\beta^0_{(k),1}, \ldots, \beta^0_{(k),r}$, we use $r + 1$ different v_j values instead of r values to maintain the validity of our method even in the case where $\{\gamma^0_{(k)}(v_j) : j = 1, \ldots, r\}$ happens to be identical to $\{\beta^0_{(k),1}, \ldots \beta^0_{(k),r}\}$. If we can tolerate a zero-probability event, then we can let

$$\mathrm{VIC}(k) = \left\| n^{-1/2} \sum_{i=1}^{n} \hat{f}\{O_i, \tilde{\gamma}_{(k)}(v)\} \right\|^2 + pk \log(n).$$

Whether or not Condition 5 holds, this will still yield $\mathrm{VIC}(k) > \mathrm{VIC}(d)$ for $k \neq d$ with probability approaching 1 as $n \to \infty$, and usually suffices in practice.

4. NUMERICAL EXPERIMENTS

We performed Monte Carlo simulation studies to assess the finite-sample performance of the proposed validated information criterion for various estimators. Specifically, we examined the validated information criterion in combination with four different estimators, namely the semiparametric sliced inverse regression, semiparametric sliced average variance, semiparametric directional regression, and semiparametric principal Hessian direction estimators, proposed

Table 1. *Selection frequencies by* VIC(k)*,* multiplied by 100*, in simulations with* $p = 6$

	$k=1$	$k=2$	$k=3$	$k=4$	$k=1$	$k=2$	$k=3$	$k=4$
		$n=200$, $d=1$				$n=200$, $d=2$		
Semi-SIR	99·4	0·6	0·0	0·0	31·0	68·3	0·6	0·1
Semi-SAVE	94·1	5·2	0·6	0·1	9·9	76·0	12·1	2·0
Semi-DR	85·6	14·0	0·4	0·0	1·4	96·6	1·7	0·3
Semi-PHD	99·9	0·1	0·0	0·0	2·6	97·2	0·2	0·0
		$n=400$, $d=1$				$n=400$, $d=2$		
Semi-SIR	99·7	0·3	0·0	0·0	4·6	94·9	0·5	0·0
Semi-SAVE	94·4	5·5	0·1	0·0	1·7	91·7	5·7	0·9
Semi-DR	86·8	12·8	0·4	0·0	1·6	98·1	0·2	0·1
Semi-PHD	99·9	0·1	0·0	0·0	0·0	99·8	0·2	0·0

Semi-SIR, semiparametric sliced inverse regression; Semi-SAVE, semiparametric sliced average variance estimation; Semi-DR, semiparametric directional regression; Semi-PHD, semiparametric principal Hessian directions.

in Ma & Zhu (2012). The validated information criterion structural dimension selection in combination with the four estimation procedures was performed on data generated from the following four models: $Y = (\beta_1^{\mathrm{T}}X)/\{0{\cdot}5 + (\beta_2^{\mathrm{T}}X + 1{\cdot}5)^2\} + 0{\cdot}5\epsilon$, $Y = (\beta_1^{\mathrm{T}}X)^2 + 2|\beta_2^{\mathrm{T}}X| + 0{\cdot}6|\beta_2^{\mathrm{T}}X|\epsilon$, $Y = 0{\cdot}1|\beta_1^{\mathrm{T}}X| + 0{\cdot}9|\beta_2^{\mathrm{T}}X| + 0{\cdot}2\epsilon$ and $Y = (\beta_1^{\mathrm{T}}X)^2 + (\beta_2^{\mathrm{T}}X)^2 + 3\epsilon$, where ϵ is a standard normal random variable. Thus the structural dimension d is equal to 2. The covariate vector X has dimension $p = 6$, where the vector consisting of the first two components, $(X_1, X_2)^{\mathrm{T}}$, has a normal distribution with mean zero, variance 1 and covariance 0·5. We set $X_3 = |X_1 + X_2| + |X_1|\epsilon_1$ and $X_4 = (X_1 + X_2)^2 + |X_2|\epsilon_2$, where ϵ_1 and ϵ_2 are standard normal random variables. Finally, X_5 and X_6 are Bernoulli random variables with success probabilities $\exp(X_2)/\{1 + \exp(X_2)\}$ and $\Phi(X_2)$, respectively, where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. We take as true parameter values $\beta = \{(1, 0, 0{\cdot}5, 0{\cdot}3, 0{\cdot}6, 0{\cdot}7)^{\mathrm{T}}(0, 1, 0{\cdot}4, -0{\cdot}5, 0{\cdot}8, 0{\cdot}3)^{\mathrm{T}}\}$. Under these simulation settings, the signal-to-noise ratios of the above four models are approximately 5·7, 17·9, 14·9 and 2·3, respectively. We also experimented with the case where the true structural dimension d is 1, by letting $\beta_2 = \beta_1$ in the four models. This leads to signal-to-noise ratios of roughly 4·6, 32·3, 20·5 and 4·7 in the four models. From Table 1, the results of our method do not seem to show a systematic relationship with the signal-to-noise ratio. In each case, we set $v_1 = v_2 = 0$ and $v_1 = v_2 = 1$. We repeated the experiments 1000 times and report the selection frequencies in Table 1 for sample sizes $n = 200$ and 400.

Table 1 indicates selection consistency, where the correct selection rates are above 65% at $n = 200$ and improve to over 85% at $n = 400$ for the validated information criterion in combination with all four estimators. These rates are not inferior to the usual correct selection rates generally seen in criterion-based methods. Moreover, as the sample size grows, we see a general trend of improvement in terms of correct selection rates. This trend is especially clear at $d = 2$, possibly because at $d = 2$ the sample size $n = 200$ is relatively small for the asymptotic properties to become evident.

We proceed to perform space estimation after determining the structural dimension using the validated information criterion. The boxplots of the Euclidean distances between the estimated spaces and the true space are displayed in Fig. 1, along with the estimation results performed under the true structural dimension. Here, the Euclidean distance is defined as the Frobenius norm of the matrix $\hat{\beta}(\hat{\beta}^{\mathrm{T}}\hat{\beta})^{-1}\hat{\beta}^{\mathrm{T}} - \beta(\beta^{\mathrm{T}}\beta)^{-1}\beta^{\mathrm{T}}$. When the correct selection rate is high, the performance based on the estimated structural dimension is very close to that achieved by using the true dimension. Thus, overall, except for the semiparametric sliced inverse regression and

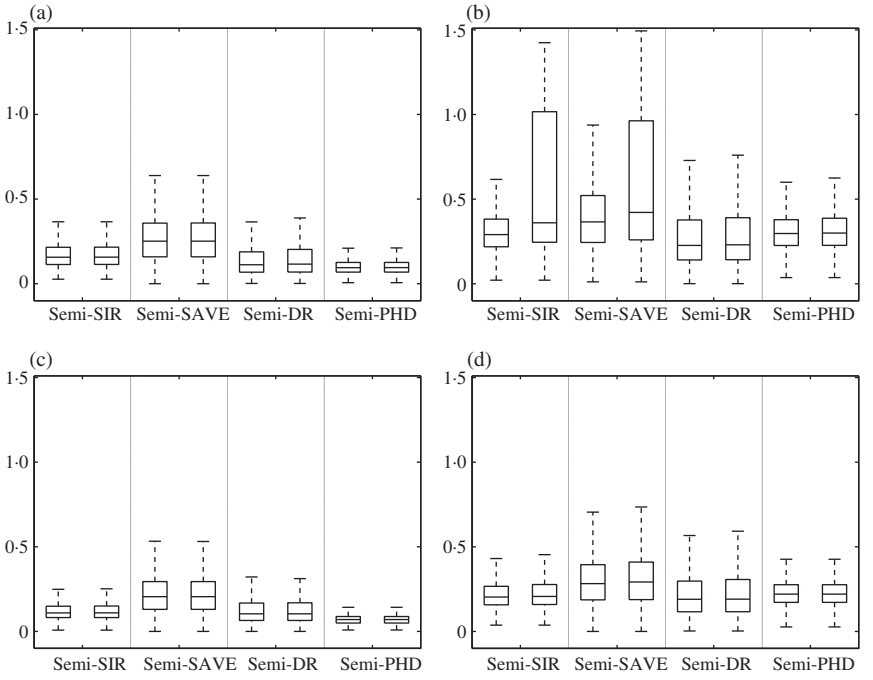

Fig. 1. Boxplots of the Frobenius norm ($p = 6$), for (a) $d = 1$, $n = 200$; (b) $d = 2$, $n = 200$; (c) $d = 1$, $n = 400$; (d) $d = 2$, $n = 400$. In each boxplot pair, the left plot is the result with the true d, and the right plot is the result with the selected d. Abbreviations have the same meanings as in Table 1.

Table 2. *Selection frequencies by* VIC(k)*,* multiplied by 100*, in simulations with* $p = 10$

	$k=1$	$k=2$	$k=3$	$k=4$	$k=1$	$k=2$	$k=3$	$k=4$
		$n=200, d=1$				$n=200, d=2$		
Semi-SIR	98·3	1·7	0·0	0·0	54·3	44·7	1·0	0·0
Semi-SAVE	91·5	5·4	3·1	0·0	20·9	57·8	21·0	0·3
Semi-DR	85·2	13·6	1·2	0·0	0·7	94·3	4·7	0·3
Semi-PHD	100·0	0·0	0·0	0·0	18·4	81·6	0·0	0·0
		$n=400, d=1$				$n=400, d=2$		
Semi-SIR	99·3	0·7	0·0	0·0	10·3	88·3	1·3	0·1
Semi-SAVE	92·7	5·7	1·6	0·0	6·1	71·2	22·5	0·2
Semi-DR	87·4	12·1	0·5	0·0	2·2	95·2	2·4	0·2
Semi-PHD	100·0	0·0	0·0	0·0	0·0	100·0	0·0	0·0

semiparametric sliced average variance estimators for $d = 2$ and $n = 200$, the different procedures yield very similar results.

We also extended the simulation to $p = 10$. We generated $(X_1, X_2, X_7, X_8, X_9, X_{10})^{\mathrm{T}}$ from a zero-mean multivariate normal distribution. The variance matrix Σ has (i, j) entry $0{\cdot}5^{|i-j|}$ for $i, j = 1, \ldots, 6$. We generated (X_3, X_4, X_5, X_6) in the same way as in the $p = 6$ case, and set $\beta = \{(1, 0, 0{\cdot}5, 0{\cdot}3, 0{\cdot}6, 0{\cdot}7, 0{\cdot}5, 0{\cdot}3, 0{\cdot}6, 0{\cdot}7)^{\mathrm{T}}(0, 1, 0{\cdot}4, -0{\cdot}5, 0{\cdot}8, 0{\cdot}3, 0{\cdot}4, -0{\cdot}5, 0{\cdot}8, 0{\cdot}3)^{\mathrm{T}}\}$. In this case, the signal-to-noise ratios of the four models are 8·1, 27·5, 25·1 and 5·7 for $d = 2$, and 4·5, 40·8, 31·2 and 10·5 for $d = 1$. The selection results are summarized in Table 2; it can be seen that results similar to those in Table 1 were obtained.

Table 3. *Analysis of the Fifth National Bank of Springfield data*

	VIC(k) values			
	$k=1$	$k=2$	$k=3$	$k=4$
Semi-SIR	145·8	169·1	212·0	417·3
Semi-SAVE	501·4	910·8	767·1	724·8
Semi-DR	131·7	277·3	155·8	139·6
Semi-PHD	105·1	249·3	216·8	200·6
	Estimates and standard errors			
	Semi-SIR	Semi-SAVE	Semi-DR	Semi-PHD
X_2	0·68 (0·04)	0·93 (0·12)	0·73 (0·18)	0·72 (0·06)
X_3	0·29 (0·05)	0·43 (0·10)	0·27 (0·18)	0·26 (0·07)
X_4	0·06 (0·05)	−0·07 (0·11)	0·04 (0·10)	−0·04 (0·05)
X_5	0·20 (0·05)	0·16 (0·07)	0·17 (0·10)	0·06 (0·09)
X_6	0·21 (0·04)	0·32 (0·06)	0·22 (0·10)	0·20 (0·04)

5. Real-data example

We applied the validated information criterion procedure to an employment dataset for the Fifth National Bank of Springfield (Albright et al., 1999), which contains a total of 207 observations. Previous analysis using the bootstrap concluded that $d = 1$ seems adequate for describing the relationship between salary and the covariates, which consist of current job level (X_1), years working at the bank (X_2), age (X_3), years working at other banks (X_4), gender (X_5) and whether the job is computer-related (X_6). Because salary is certainly related to job level, we denote this covariate by X_1 and take the corresponding coefficient to be $\beta_1 = 1$.

Using the validated information criterion in combination with the four semiparametric estimators, the structural dimension was determined to be one in all cases; see Table 3. This result confirms that the previous conclusion of Ma & Zhu (2012) is sensible and that, indeed, the effect of the six covariates on salary can be summarized by a single direction. We report in Table 3 the results of estimating this single direction, from which it can be seen that the four estimation results are similar.

Acknowledgement

Ma was supported by the U.S. National Science Foundation and the National Institute of Neurological Disorders and Stroke. Zhang, the corresponding author, was supported by the Natural Science Foundation of China and is also affiliated with the Academy of Mathematics and Systems Science, Chinese Academy of Sciences. Zhang's work was undertaken during his visit to Dr Raymond J. Carroll of Texas A&M University, whose support is greatly appreciated.

Appendix

Proof of Theorem 1

For any $p \times k$ matrix B with $p > k$, we use vecl(B) to denote the vector formed by the concatenation of the columns of the lower $(p-k) \times k$ submatrix of B. We first consider $k \leqslant d$. Using the definition of $\hat{f}(O, \tilde{\gamma}_{(k)})$ and the usual Taylor expansion, we have

$$n^{-1/2} \sum_{i=1}^{n} \hat{f}\{O_i, \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} = n^{-1/2} \sum_{i=1}^{n} \left[a(x_i) - \hat{E}\{a \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\}\right] \left[g\{Y_i, \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} - \hat{E}\{g \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\}\right]$$

$$= n^{-1/2} \sum_{i=1}^{n} \left[a(x_i) - E\{a \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\}\right] \left[g\{Y_i, \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} - E\{g \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\}\right]$$

$$
\begin{aligned}
&+ n^{-1/2} \sum_{i=1}^{n} \left[E\{a(x_i) \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} - \hat{E}\{a(x_i) \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} \right] \\
&\times \left(g(Y_i, \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i) - E\left[g\{Y_i, \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i \right] \right) \\
&+ n^{-1/2} \sum_{i=1}^{n} \left[a(x_i) - E\{a(x_i) \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} \right] \\
&\times \left(E\left[g\{Y_i, \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i \right] - \hat{E}\left[g\{Y_i, \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i \right] \right) \\
&+ n^{-1/2} \sum_{i=1}^{n} \left[E\{a(x_i) \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} - \hat{E}\{a(x_i) \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} \right] \\
&\times \left(E\left[g\{Y_i, \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i \right] - \hat{E}\left[g\{Y_i, \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i \right] \right) \\
&= n^{-1/2} \sum_{i=1}^{n} \left[a(x_i) - E\{a(x_i) \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} \right] \\
&\times \left(g\{Y_i, \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} - E[g\{Y_i, \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i\} \mid \tilde{\gamma}_{(k)}^{\mathrm{T}}(v)x_i] \right) + o_{\mathrm{p}}(1),
\end{aligned}
$$

where the last equality follows from Conditions 1–4 with $\theta = \gamma_{(k)}(v)$ and Lemmas 3 and 4 of Ma & Zhu (2012). Thus,

$$
n^{-1/2} \sum_{i=1}^{n} \hat{f}\{O_i, \tilde{\gamma}_{(k)}(v)\} = n^{-1/2} \sum_{i=1}^{n} f\{O_i, \tilde{\gamma}_{(k)}(v)\} + o_{\mathrm{p}}(1). \tag{A1}
$$

For any $p \times k$ matrix β with upper $k \times k$ submatrix being the identity matrix, we define

$$
A(\beta) = E\left\{ \frac{\partial f(O, \beta)}{\partial\, \mathrm{vecl}(\beta)^{\mathrm{T}}} \right\}, \quad \hat{A}(\beta) = n^{-1} \sum_{i=1}^{n} \frac{\partial \hat{f}(O_i, \beta)}{\partial\, \mathrm{vecl}(\beta)^{\mathrm{T}}}.
$$

From (2) and (3), we have

$$
\hat{A}^{\mathrm{T}}(\hat{\beta}_{(k)}) W \sum_{i=1}^{n} \hat{f}(O_i, \hat{\beta}_{(k)}) = 0, \quad A^{\mathrm{T}}(\beta_{(k)}^0) W E\left\{ f(O, \beta_{(k)}^0) \right\} = 0. \tag{A2}
$$

Under Conditions 1–4 with $\theta = \beta_{(k)}^0$, (A1) and the first equality of (A2) lead to

$$
\begin{aligned}
0 &= n^{-1/2} \left\{ A^{\mathrm{T}}(\beta_{(k)}^0) + o_{\mathrm{p}}(1) \right\} W \sum_{i=1}^{n} f(O_i, \hat{\beta}_{(k)}) + o_{\mathrm{p}}(1) \\
&= n^{-1/2} \left\{ A^{\mathrm{T}}(\beta_{(k)}^0) + o_{\mathrm{p}}(1) \right\} W \sum_{i=1}^{n} f(O_i, \beta_{(k)}^0) \\
&\quad + \left\{ A^{\mathrm{T}}(\beta_{(k)}^0) + o_{\mathrm{p}}(1) \right\} W \left\{ A(\beta_{(k)}^0) + o_{\mathrm{p}}(1) \right\} n^{1/2} \mathrm{vecl}(\hat{\beta}_{(k)} - \beta_{(k)}^0) + o_{\mathrm{p}}(1) \\
&= n^{-1/2} A^{\mathrm{T}}(\beta_{(k)}^0) W \sum_{i=1}^{n} f(O_i, \beta_{(k)}^0) + o_{\mathrm{p}}(1) n^{-1/2} \sum_{i=1}^{n} f(O_i, \beta_{(k)}^0) \\
&\quad + A^{\mathrm{T}}(\beta_{(k)}^0) W A(\beta_{(k)}^0) n^{1/2} \mathrm{vecl}(\hat{\beta}_{(k)} - \beta_{(k)}^0) + o_{\mathrm{p}}(1),
\end{aligned}
$$

which results in

$$n^{1/2}\mathrm{vecl}(\hat{\beta}_{(k)} - \beta^0_{(k)}) = -n^{-1/2}\sum_{i=1}^{n}\{A^{\mathrm{T}}(\beta^0_{(k)})WA(\beta^0_{(k)})\}^{-1}A^{\mathrm{T}}(\beta^0_{(k)})Wf(O_i, \beta^0_{(k)}) + o_{\mathrm{p}}(1)n^{-1/2}\sum_{i=1}^{n} f(O_i, \beta^0_{(k)}) + o_{\mathrm{p}}(1). \quad \text{(A3)}$$

Because of the definition of $\tilde{\gamma}_{(k)}(v)$ and $\gamma^0_{(k)}(v)$, there exists a $(p-1-k)(k+1)\times(p-k)k$ matrix $P_{(k)}(v)$ such that $\mathrm{vecl}\{\tilde{\gamma}_{(k)}(v) - \gamma^0_{(k)}(v)\} = P_{(k)}(v)\mathrm{vecl}(\hat{\beta}_{(k)} - \beta^0_{(k)})$. By Taylor expansion and Conditions 1–4 with $\theta = \gamma^0_{(k)}$, we further write

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n} f\{O_i, \tilde{\gamma}_{(k)}(v)\} &= n^{-1/2}\sum_{i=1}^{n} f\{O_i; \gamma^0_{(k)}(v)\} + n^{1/2}A\{\gamma^0_{(k)}(v)\}\,\mathrm{vecl}\{\tilde{\gamma}_{(k)}(v) - \gamma^0_{(k)}(v)\} + o_{\mathrm{p}}(1) \\ &= n^{-1/2}\sum_{i=1}^{n} f\{O_i; \gamma^0_{(k)}(v)\} + n^{1/2}A\{\gamma^0_{(k)}(v)\}P_{(k)}(v)\,\mathrm{vecl}(\hat{\beta}_{(k)} - \beta^0_{(k)}) + o_{\mathrm{p}}(1) \\ &= n^{-1/2}\sum_{i=1}^{n} f\{O_i; \gamma^0_{(k)}(v)\} - n^{-1/2}\sum_{i=1}^{n} M_{(k)}(v)f(O_i, \beta^0_{(k)}) \\ &\quad + o_{\mathrm{p}}(1)n^{-1/2}\sum_{i=1}^{n} f(O_i, \beta^0_{(k)}) + o_{\mathrm{p}}(1), \end{aligned} \quad \text{(A4)}$$

where $A\{\gamma^0_{(k)}(v)\}$ is defined analogously to $A(\beta^0_{(k)})$,

$$M_{(k)}(v) = A\{\gamma^0_{(k)}(v)\}P_{(k)}(v)\{A^{\mathrm{T}}(\beta^0_{(k)})WA(\beta^0_{(k)})\}^{-1}A^{\mathrm{T}}(\beta^0_{(k)})W,$$

and we have used (A3) to obtain the last equality in (A4). Using (A1) and (A4), we have

$$\begin{aligned} n^{-1/2}\sum_{i=1}^{n}\hat{f}\{O_i, \tilde{\gamma}_{(k)}(v)\} &= n^{-1/2}\sum_{i=1}^{n} f\{O_i; \gamma^0_{(k)}(v)\} - n^{-1/2}\sum_{i=1}^{n} M_{(k)}(v)f(O_i, \beta^0_{(k)}) \\ &\quad + o_{\mathrm{p}}(1)n^{-1/2}\sum_{i=1}^{n} f(O_i, \beta^0_{(k)}) + o_{\mathrm{p}}(1) \\ &= n^{-1/2}\sum_{i=1}^{n}\big(f\{O_i; \gamma^0_{(k)}(v)\} - E[f\{O; \gamma^0_{(k)}(v)\}] - M_{(k)}(v)f(O_i, \beta^0_{(k)})\big) \\ &\quad + o_{\mathrm{p}}(1)n^{-1/2}\sum_{i=1}^{n}\big[f(O_i, \beta^0_{(k)}) - E\{f(O, \beta^0_{(k)})\}\big] \\ &\quad + n^{1/2}E[f\{O; \gamma^0_{(k)}(v)\}] + o_{\mathrm{p}}(1)n^{1/2}E\{f(O, \beta^0_{(k)})\} + o_{\mathrm{p}}(1). \end{aligned} \quad \text{(A5)}$$

From the definition of $M_{(k)}(v)$ and (A2) and using the central limit theorem, the first term in (A5) converges to a normal variate with mean zero and variance $\mathrm{var}[f\{O, \gamma^0_{(k)}(v)\} - M_{(k)}(v)f(O, \beta^0_{(k)})]$; hence it is $O_{\mathrm{p}}(1)$. Similarly, the second term of (A5) is of order $o_{\mathrm{p}}(1)$.

We now examine the third and fourth terms in (A5). When $k = d$, to minimize (3), $\beta^0_{(k)}$ is the true parameter value and satisfies $E\{f(O, \beta^0_{(k)})\} = 0$; hence the fourth term equals 0. We also have

$$E\left\{\big[a(x) - E\{a(x) \mid \gamma^{0\mathrm{T}}_{(k)}(v)x\}\big]\big(g\{Y, \gamma^{0\mathrm{T}}_{(k)}(v)x\} - E[g\{Y, \gamma^{0\mathrm{T}}_{(k)}(v)x\} \mid \gamma^{0\mathrm{T}}_{(k)}(v)x]\big)\right\} = 0$$

or, equivalently, $E\{f(O, \gamma_{(k)}^0)\} = 0$. This is because $\beta_{(k)}^{0\mathrm{T}} x$ is a linear combination of $\gamma_{(k)}^{0\mathrm{T}}(v)x$, as we have pointed out, which gives

$$E[g\{Y, \gamma_{(k)}^{0\mathrm{T}}(v)x\} \mid x] = E[g\{Y, \gamma_{(k)}^{0\mathrm{T}}(v)x\} \mid \beta_{(d)}^{0\mathrm{T}} x, \gamma_{(k)}^{0\mathrm{T}}(v)x] = E[g\{Y, \gamma_{(k)}^{0\mathrm{T}}(v)x\} \mid \gamma_{(k)}^{0\mathrm{T}}(v)x].$$

This leads directly to the result that $\mathrm{VIC}(d) = pd\log(n) + O_{\mathrm{p}}(1)$. On the other hand, when $k < d$,

$$E\left\{\left[a(x) - E\{a(x) \mid \gamma_{(k)}^{0\mathrm{T}}(v)x\}\right]\left(g\{Y, \gamma_{(k)}^{0\mathrm{T}}(v)x\} - E[g\{Y, \gamma_{(k)}^{0\mathrm{T}}(v)x\} \mid \gamma_{(k)}^{0\mathrm{T}}(v)x]\right)\right\}$$

does not vanish unless $\gamma_{(k)}^0(v) = \beta_{(k+1)}^0$, by Condition 5. Thus, for $\gamma_{(k)}^0(v) \neq \beta_{(k+1)}^0$, we have that $E[f\{O, \gamma_{(k)}^0(v)\}] = c(v) \neq 0$; so the third term in (A5) is of order $n^{1/2}$. The fourth term is of order $o_{\mathrm{p}}(n^{1/2})$. This leads directly to the result that for $k < d$,

$$\mathrm{VIC}(k) = 2^{-1} n\{c(v_1)^{\mathrm{T}} c(v_1) + c(v_2)^{\mathrm{T}} c(v_2)\} + o_{\mathrm{p}}(n) + pk\log(n),$$

which is larger than $\mathrm{VIC}(d)$ with probability approaching 1 as $n \to \infty$, for any choices of v_1 and v_2 (with $v_1 \neq v_2$). Finally, for $k > d$, it is easy to see that $\mathrm{VIC}(k) \geqslant pk\log(n)$, which is larger than $\mathrm{VIC}(d)$ with probability approaching 1 as $n \to \infty$.

REFERENCES

ALBRIGHT, S. C., WINSTON, W. L. & ZAPPE, C. J. (1999). *Data Analysis and Decision Making with Microsoft Excel*. Pacific Grove, California: Duxbury.

BURA, E. & COOK, R. D. (2001). Extending sliced inverse regression: The weighted chi-squared test. *J. Am. Statist. Assoc.* **96**, 996–1003.

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

COOK, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through 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. & NI, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *J. Am. Statist. Assoc.* **100**, 410–28.

COOK, R. D. & YIN, X. (2001). Dimension reduction and visualization in discriminant analysis (with Discussion). *Aust. New Zeal. J. Statist.* **43**, 147–99.

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

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

LUO, R., WANG, H. & TSAI, C. L. (2009). Contour projected dimension reduction. *Ann. Statist.* **37**, 3743–78.

MA, Y. & ZHU, L. P. (2012). A semiparametric approach to dimension reduction. *J. Am. Statist. Assoc.* **107**, 168–79.

MA, Y. & ZHU, L. P. (2013). Efficiency loss and the linearity condition in dimension reduction. *Biometrika* **100**, 371–83.

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

VELILLA, S. (1998). Assessing the number of linear components in a general regression problem. *J. Am. Statist. Assoc.* **93**, 1088–98.

XIA, Y., TONG, H., LI, W. K. & ZHU, L. X. (2002). An adaptive estimation of dimension reduction space (with Discussion). *J. R. Statist. Soc.* B **64**, 363–410.

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.

ZHU, L. P., YU, Z. & ZHU, L. X. (2010). A sparse eigen-decomposition estimation in semi-parametric regression. *Comp. Statist. Data Anal.* **54**, 976–86.

ZHU, L. P. & ZHU, L. X. (2007). On kernel method for sliced average variance estimation. *J. Mult. Anal.* **98**, 970–91.

ZHU, L. X., MIAO, B. Q. & PENG, H. (2006). On sliced inverse regression with large dimensional covariates. *J. Am. Statist. Assoc.* **101**, 630–43.

ZHU, Y. & ZENG, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Am. Statist. Assoc.* **101**, 1638–51.

ZENG, P. (2008). Determining the dimension of the central subspace and central mean subspace. *Biometrika* **95**, 469–79.

[*Received February* 2014. *Revised December* 2014]