Kent Academic Repository

Kong, Efang and Xia, Yingcun (2007) *Variable selection for the single index model.* Biometrika, 94 (1). pp. 217-229. ISSN 0006-3444.

Variable selection for the single-index model

BY EFANG KONG AND YINGCUN XIA

Department of Statistics and Applied Probability, National University of Singapore, 117546, Singapore
g0201815@nus.edu.sg   staxyc@stat.nus.edu.sg

SUMMARY

We consider variable selection in the single-index model. We prove that the popular leave-m-out crossvalidation method has different behaviour in the single-index model from that in linear regression models or nonparametric regression models. A new consistent variable selection method, called separated crossvalidation, is proposed. Further analysis suggests that the method has better finite-sample performance and is computationally easier than leave-m-out crossvalidation. Separated crossvalidation, applied to the Swiss banknotes data and the ozone concentration data, leads to single-index models with selected variables that have better prediction capability than models based on all the covariates.

Some key words: Consistency; Crossvalidation; Nonparametric smoothing; Semiparametric model; Variable selection.

1. INTRODUCTION

Suppose Y is a response variable and $X = (x_1, \ldots, x_p)^\top$ are covariates. The single-index model is written as

$$Y = g(X^\top \theta^0) + \varepsilon, \tag{1}$$

where $E(\varepsilon|X) = 0$ almost surely, g is an unknown link function and θ^0 is an unknown unit vector with first nonzero component positive for identification purposes. Recent papers (Powell et al., 1989; Härdle & Stoker, 1989; Ichimura, 1993; Klein & Spady, 1993; Härdle et al., 1993; Horowitz & Härdle, 1996; Hristache et al., 2001; Xia et al., 2002) have considered the estimation of the index parameter and the nonparametric link function with focus on the root-n consistency of the former; efficiency issues have also been studied. Amongst the various methods of estimation, the most popular are the average derivative estimation method proposed by Härdle & Stoker (1989) and the method of Härdle et al. (1993).

All the studies mentioned above assume that all regressors X contain useful information to predict the response variable. If irrelevant regressors are included, which is very likely in high-dimensional environments (Naik & Tsai, 2000), the precision of parameter estimation as well as the accuracy of forecasting will suffer (Altham, 1984). Therefore, it makes sense to exclude irrelevant covariates from single-index models. Naik & Tsai (2001) considered variable selection using sliced inverse regression in which the predictor X is required to be continuous and elliptically symmetric. However, in practice some covariates may be asymmetric or discrete.

For parametric models, methods based on crossvalidation or on measures such as AIC (Akaike, 1974) have been the main focus of attention in model identification and variable selection (Miller, 2002). For linear regression models, the leave-one-out crossvalidation method (Stone, 1974) is inconsistent and tends to select unnecessarily many variables. However, Shao (1993) proved that, if $m/n \to 1$ and $n - m \to \infty$, then leave-m-out crossvalidation is consistent. On the other hand, under nonparametric settings, leave-one-out crossvalidation is consistent; see for example Tjøstheim & Auestad (1994) and Cheng & Tong (1992). An interesting discussion about the different performances of crossvalidation in linear regression models and nonparametric models can also be found in Gao & Tong (2004).

Semiparametric models are different again. We shall show that leave-one-out crossvalidation again fails in selecting the variables of a single-index model. However, leave-m-out crossvalidation is consistent for single-index models provided that $m/n \to c \in [2/3, 1)$, different from the requirements on m in linear regression models. Thus, no more than 1/3 of the samples should be used for model estimation, and this is usually not enough to estimate the model well, resulting in inferior efficiency in variable selection. Furthermore, the computation of leave-m-out crossvalidation is very time-consuming. To overcome these disadvantages, we shall propose the separated crossvalidation method.

2. Optimal model and parameter estimation

We use notation similar to that of Shao (1993). Let $\mathcal{S}$ denote all nonempty subsets of $\{1, \ldots, p\}$. For any $\alpha \in \mathcal{S}$, let d_α be the cardinality of α, and let θ_α and X_α be two $d_\alpha \times 1$ column vectors containing the components of θ and X respectively indexed by the integers in α. Let θ denote the vector which minimises $E\{Y - E(Y|X_\alpha^\top \theta)\}^2$. The corresponding single-index model,

$$Y = g_\alpha(X_\alpha^\top \theta) + \varepsilon_\alpha, \quad \varepsilon_\alpha = Y - E(Y|X_\alpha^\top \theta) = Y - g_\alpha(X_\alpha^\top \theta), \tag{2}$$

is denoted by $\mathcal{M}_\alpha$. If we know whether or not each component of the true θ^0 is zero, then models $\mathcal{M}_\alpha$ can be classified into two categories. In one category, at least one covariate with a nonzero coefficient in (1) is missing in X_α. In the other category, X_α contains all covariates with nonzero coefficients. The optimal model, denoted by $\mathcal{M}_{\alpha_0}$, is defined as the one in the second category with the smallest number d_0 of covariates.

Suppose $\{(X_i, Y_i), i = 1, \ldots, n\}$ is a random sample from model (1). Consider model $\mathcal{M}_\alpha$ with $\alpha \supset \alpha_0$. To guarantee the consistency of estimation, we assume throughout the paper that $X_\alpha^\top \theta$ has a density function for all θ in a small neighbourhood of θ_α^0, a column vector containing the components of θ^0 indexed by the integers in α; see Horowitz & Härdle (1996) for more discussion. The popular method of Härdle et al. (1993) estimates the model as follows. Suppose $A \subseteq R^p$ is a compact convex set such that the density function of $X^\top \theta$ is uniformly bounded away from zero on $\{\theta^\top x : x \in A\}$ for any θ near θ^0. For any given $b > 0$ and $h > 0$, let $A^{bh} = \{x \in R^p : \|x - x_0\| \leqslant bh$ for some $x_0 \in A\}$. Introducing A and A^{bh} is for technical purposes; see Härdle et al. (1993) for more details. Let $g_\alpha(u|\theta) = E(Y|X_\alpha^\top \theta = u^\top \theta)$. Its leave-one-out estimator is given by

$$\hat{g}_\alpha^{\setminus i}(u|\theta) = \frac{\sum_{j \neq i} K_h(X_{j,\alpha}^\top \theta - u^\top \theta) Y_j}{\sum_{j \neq i} K_h(X_{j,\alpha}^\top \theta - u^\top \theta)},$$

where h is a bandwidth, K is a univariate kernel function with support $[-b, b]$ and $K_h(.) = h^{-1}K(./h)$. The index parameter in model $\mathcal{M}_\alpha$ is estimated by minimizing

$$\mathrm{HCV}_\alpha(\theta, h) := \sum_i{}' \{Y_i - \hat{g}_\alpha^{\backslash i}(X_{i,\alpha}|\theta)\}^2, \qquad (3)$$

with respect to θ and $h > 0$ subject to $||\theta|| = 1$, where $\sum_i'$ denotes summation over indices i such that $X_i \in A$. We assume that all $X_i \in A^{bh}$; otherwise one can always completely ignore those data outside A^{bh}. For ease of exposition, suppose that $X_i \in A$ if $1 \leqslant i \leqslant n'$ and $X_i \notin A$ if $i > n'$, which implies that $n - n' = O(nh)$. This estimator has very good asymptotic properties. It needs no under-smoothing for the estimator of θ to achieve the root-n consistency. However, it is not easy to solve the above minimization problem, especially when d_α is large.

Based on local linear approximation (Ruppert & Wand, 1994), Xia et al. (2002) estimated θ_α^0 by

$$\hat{\theta} = \arg \min_{\theta : ||\theta||=1} \sum_{j=1}^n \sum_{i=1}^n (Y_i - a_j - d_j \theta^\top X_{ij,\alpha})^2 w_{ij},$$

where $X_{ij,\alpha} = X_{i,\alpha} - X_{j,\alpha}$ and w_{ij} is a weight depending on the distance between $X_{i,\alpha}$ and $X_{j,\alpha}$. The corresponding algorithm takes the following form. With an initial value θ, calculate

$$\begin{pmatrix} a_j^\theta \\ d_j^\theta h \end{pmatrix} = \left\{ \sum_{i=1}^n K_h(X_{ij,\alpha}^\top \theta) \begin{pmatrix} 1 \\ X_{ij,\alpha}^\top \theta/h \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij,\alpha}^\top \theta/h \end{pmatrix}^\top \right\}^{-1} \sum_{i=1}^n K_h(X_{ij,\alpha}^\top \theta) \begin{pmatrix} 1 \\ X_{ij,\alpha}^\top \theta/h \end{pmatrix} Y_i, \quad (4)$$

and then calculate

$$\theta = \left\{ \sum_{i,j} K_h(X_{ij,\alpha}^\top \theta)(d_j^\theta)^2 X_{ij,\alpha} X_{ij,\alpha}^\top \right\}^{-1} \sum_{i,j} K_h(X_{ij,\alpha}^\top \theta) d_j^\theta X_{ij,\alpha}(y_i - a_j^\theta), \quad \theta =: \mathrm{sign}(\theta_1) \frac{\theta}{||\theta||}. \quad (5)$$

Repeat (4) and (5) until the iteration process converges, to what we call the minimum average variance estimate.

Xia & Tong (2006) proved that the estimator produced by the algorithm can achieve root-n consistency and has the same asymptotic distribution as the estimator of Härdle et al. (1993).

3. CROSSVALIDATION VARIABLE SELECTION

In the crossvalidation method, the data are split into two sets: the training set s^c and the test set s. The training set is used to estimate all candidate models, and the model that best predicts the test set is the preferred model. Note that, in (2), $\theta = \theta_\alpha^0$ for any $\alpha \supset \alpha_0$. For such α and any $s \subset \{1, \ldots, n'\}$ with cardinality m, we first estimate θ_α^0 by $\hat{\theta}_\alpha^{\backslash s}$, the minimum average variance estimate of θ in model (2) using $\{(X_j, Y_j) : 1 \leqslant j \leqslant n, \ j \notin s\}$. We then estimate the link function $g_\alpha(u|\hat{\theta}_\alpha^{\backslash s})$ by the local linear smoother

$$\hat{g}_\alpha^{\backslash s}(u|\hat{\theta}_\alpha^{\backslash s}) = \sum_{j \notin s} M_{\alpha,h}\{(X_{j,\alpha} - u)^\top \hat{\theta}_\alpha^{\backslash s}\}Y_j \Big/ \sum_{j \notin s} M_{\alpha,h}\{(X_{j,\alpha} - u)^\top \hat{\theta}_\alpha^{\backslash s}\}, \qquad (6)$$

where

$$M_{\alpha,h}\{(X_{j,\alpha} - u)^\top \hat{\theta}_\alpha^{\backslash s}\} = S_{\alpha,2}^{\backslash s}(u|\hat{\theta}_\alpha^{\backslash s}) K_h\{(X_{j,\alpha} - u)^\top \hat{\theta}_\alpha^{\backslash s}\}$$

$$- S_{\alpha,1}^{\backslash s}(u|\hat{\theta}_\alpha^{\backslash s})\{(X_{j,\alpha} - u)^\top \hat{\theta}_\alpha^{\backslash s}/h\} K_h\{(X_{j,\alpha} - u)^\top \hat{\theta}_\alpha^{\backslash s}\}$$

with $S_{\alpha,k}^{\backslash s}(u|\theta) = \sum_{j\notin s} K_h\{(X_{j,\alpha} - u)^\top \theta\}[\{(X_{j,\alpha} - u)^\top \theta\}/h]^k$, $k = 0, 1, 2$. We define the leave-m-out crossvalidation function as

$$CV_\alpha(m) = m^{-1}\binom{n'}{m}^{-1}\sum_{s}{}'\sum_{i\in s}\{Y_i - \hat{g}_\alpha^{\backslash s}(X_{i,\alpha}|\hat{\theta}_\alpha^{\backslash s})\}^2, \tag{7}$$

where $\sum_s'$ indicates summation over all possible subsets $s \subset \{1, \dots, n'\}$ with cardinality m. Later, we will use $\sum_{i,s}'$ to denote $\sum_s'\sum_{i\in s}$. The model $\mathcal{M}_\alpha$ with the smallest value of $CV_\alpha(m)$ is the selected model.

THEOREM 1. *Suppose Assumptions* A1–A7 *in the Appendix hold. If* $m \to \infty$ *with* $m/n \to c \in [0, 1)$ *and* $h \propto n^{-1/5}$, *then, for any* $\alpha \supset \alpha_0$ *and* $\delta_d := d_\alpha - d_0$, *we have*

$$\lim_{n\to\infty} \mathrm{pr}\left\{CV_\alpha(m) > CV_{\alpha_0}(m)\right\} = \mathrm{pr}\{\chi^2(\delta_d) > \frac{(2 - 3c)\delta_d}{1 - c}\}.$$

By Theorem 1, for leave-m-out crossvalidation to be consistent, i.e. $\lim_{n\to\infty} \mathrm{pr}\{CV_\alpha(m) > CV_{\alpha_0}(m)\} = 1$, it is required that $2 - 3c \leqslant 0$ and $c < 1$, or $1 > c \geqslant 2/3$. Although we have no conclusion in the case $c = 1$, our conjecture is that consistency does not hold, since $\hat{\theta}_\alpha^{\backslash s}$ is no longer root-n consistent as $n_c := n - m = o(n)$, i.e. the size of the training set is much smaller than n.

The way leave-m-out crossvalidation splits the data is acceptable for a linear regression model since its parameter can be estimated quite well with a small sample. However, the size of the training set used by leave-m-out crossvalidation is usually too small for the nonparametric smoothing methods. Another disadvantage of leave-m-out crossvalidation is its heavy computational burden since there are $\binom{n'}{m}$ possible splitting combinations. To reduce the burden, Monte Carlo leave-m-out crossvalidation randomly draws, with or without replacement, a collection $\mathcal{R}$ of subsets of $\{1, \dots, n'\}$, of size m, and selects a model that minimizes

$$CV_\alpha^{\mathrm{mc}}(m) := \sum_{s\in\mathcal{R}}\sum_{i\in s}\{Y_i - \hat{g}_\alpha^{\backslash s}(X_{i,\alpha}|\hat{\theta}_\alpha^{\backslash s})\}^2. \tag{8}$$

In linear regression models, the performance of this method has been proved to be similar to that of leave-m-out crossvalidation; see Zhang (1993) and Shao (1993). Monte Carlo leave-m-out crossvalidation is thus used in our simulation study instead of leave-m-out crossvalidation.

Although Theorem 1 is proved for the minimum average variance estimator, other model estimation methods can also be used providing that the estimator has a similar stochastic expansion to that in (A1). Examples are the estimator by Härdle et al. (1993), albeit computationally intensive, and the average derivatives estimator investigated by Härdle & Stoker (1989). The method of Hristache et al. (2001) might also work because Xia & Tong (2006) proved that an alternative version has a similar expansion.

4. VARIABLE SELECTION BY SEPARATION

Starting with the full covariate set $\{x_1, \dots, x_p\}$, we need to check whether a certain covariate, x_d say, contributes to the response variable Y. For this purpose, we introduce the model

$$Y = g(X_\alpha^\top \theta, x_d) + e, \quad \alpha \cup \{d\} = \{1, \dots, p\}. \tag{9}$$

Compared with model (1), in which the contribution of x_d is mixed up with that of the other covariates through a linear combination, the contribution of x_d in model (9) is 'separated' and can be assessed more accurately. Another reason for us to introduce model (9) is the different behaviours of crossvalidation for parametric models and nonparametric models. Note that the relationship between Y and x_d is 'nonparametric' in (9). Therefore, leave-one-out crossvalidation can tell whether or not x_d contributes significantly to the response variable Y as proved in Cheng & Tong (1992) and Yao & Tong (1994).

The parameter θ in model (9) can be estimated by the first d_α entries of the minimum average variance estimate of θ in $Y = g(X_{\alpha \cup d}^\top \theta) + e$. For any fixed θ, define $g_{\alpha,d}(u, v | \theta) = E(Y | X_\alpha^\top \theta = u^\top \theta, x_d = v)$. Its leave-one-out estimator $\hat{g}_{\alpha,d}^{\backslash i}(u, v | \theta)$ is the first entry of

$$\left\{ \sum_{j \neq i} K_{h_1,j}^{\alpha,\theta}(u, v) \begin{pmatrix} 1 \\ \theta^\top(X_{j,\alpha} - u) \\ X_{j,d} - v \end{pmatrix} \begin{pmatrix} 1 \\ \theta^\top(X_{j,\alpha} - u) \\ X_{j,d} - v \end{pmatrix}^\top \right\}^{-1} \sum_{j \neq i} K_{h_1,j}^{\alpha,\theta}(u, v) \begin{pmatrix} 1 \\ \theta^\top(X_{j,\alpha} - u) \\ X_{j,d} - v \end{pmatrix}^\top Y_j, \quad (10)$$

where $K_{h_1,j}^{\alpha,\theta}(u, v) = K_{h_1}(X_{j,\alpha}^\top \theta - u) H_{h_1}(x_{j,d} - v)$ is a two-dimensional product kernel, h_1 is a bandwidth, $H = K$ if x_d is continuous and $H_h(v) = I(v = 0)$ if x_d is discrete.

For ease of exposition, we use $\hat{g}_{\alpha_1,d}^{\backslash i}(X_i | \hat{\theta}_{\alpha_1}^{\backslash i})$ and $\hat{g}_{\alpha_1}^{\backslash i}(X_i | \hat{\theta}_{\alpha_1}^{\backslash i})$ to denote $\hat{g}_{\alpha_1,d}^{\backslash i}(X_{i,\alpha_1}, x_{i,d} | \hat{\theta}_{\alpha_1}^{\backslash i})$ and $\hat{g}_{\alpha_1}^{\backslash i}(X_{i,\alpha_1} | \hat{\theta}_{\alpha_1}^{\backslash i})$ respectively. We propose the following algorithm for the variable selection. Start with an initial covariate set α satisfying $\alpha_0 \subset \alpha$.

Step 1. Calculate $\hat{\theta}_\alpha$, the minimum average variance estimate of θ in model $Y = g(X_\alpha^\top \theta) + \varepsilon$ from all data points. Find the entry of $\hat{\theta}_\alpha$ with the smallest absolute value and its corresponding index in α, d say. Set $\alpha_1 = \alpha \setminus \{d\}$.

Step 2. Denote by $\hat{\theta}_\alpha^{\backslash i}$ the minimum average variance estimate of θ in $Y = g(X_\alpha^\top \theta) + \varepsilon$ based on $\{(X_j, Y_j)\}_{j \neq i}$. Eliminate the last entry and denote the rest by $\hat{\theta}_{\alpha_1}^{\backslash i}$.

Step 3. Calculate $\hat{g}_{\alpha_1,d}^{\backslash i}(X_i | \hat{\theta}_{\alpha_1}^{\backslash i})$ as defined in (10) and $\hat{g}_{\alpha_1}^{\backslash i}(X_i | \hat{\theta}_{\alpha_1}^{\backslash i})$ as defined in (6), with α and θ replaced by α_1 and $\hat{\theta}_{\alpha_1}^{\backslash i}$ respectively. Let

$$\mathrm{CV}_{\alpha_1,d} = \frac{1}{n'} \sum_i{}' \{Y_i - \hat{g}_{\alpha_1,d}^{\backslash i}(X_i | \hat{\theta}_{\alpha_1}^{\backslash i})\}^2, \quad \mathrm{CV}_{\alpha_1} = \frac{1}{n'} \sum_i{}' \{Y_i - \hat{g}_{\alpha_1}^{\backslash i}(X_i | \hat{\theta}_{\alpha_1}^{\backslash i})\}^2,$$

where $\sum_i{}'$ is defined in (3). If $\mathrm{CV}_{\alpha_1,d} < \mathrm{CV}_{\alpha_1}$, stop and select α. Otherwise, go to Step 1 with α replaced by α_1.

Repeat the above procedure until no more variables can be eliminated. We call this procedure the separated crossvalidation method.

Step 1 is employed to simplify the calculations. As θ^0 can be estimated with root-n consistency in single-index models, if $\alpha \supset \alpha_0$, i.e. x_d is redundant, then $\hat{\theta}_d = O_p(n^{-1/2})$. If x_d is necessary, then $\hat{\theta}_d = \theta_d^0 + O_p(n^{-1/2})$, which is bounded away from zero in probability. Therefore, if the initial covariate set α contains redundant variables, then with probability tending to 1 only the redundant variables will be considered for elimination from Step 1. It can be used further to simplify the calculation in Steps 2 and 3 by replaying $\hat{\theta}_\alpha^{\backslash i}$ and $\hat{\theta}_{\alpha_1}^{\backslash i}$ with $\hat{\theta}_\alpha$ and $\hat{\theta}_{\alpha_1}$ respectively. Step 2 is employed to estimate the parameters in model (9) assuming that x_d can be removed. Step 3 calculates and compares the leave-one-out crossvalidation values for models (2) and (9) in order to check the importance of x_d; see also Cheng & Tong (1992).

As shown in Härdle et al. (1993) and Xia & Tong (2006), the commonly used bandwidth selection methods for nonparametric regression can be used to estimate the link function

as well as the index parameter. As for the calculation of (10), theoretical justification requires different bandwidths for the estimation of model (9) depending on the type of x_d: $h_1 \propto n^{-1/6}$ if x_d is continuous and $h_1 = h \propto n^{-1/5}$ if x_d is discrete, where h is the bandwidth used in the calculation of CV_{α_1}. Many available bandwidth selection methods, such as crossvalidation or generalized crossvalidation bandwidth selection methods and the rule-of-thumb can be used to choose the bandwidths; see Silverman (1986) and Fan & Gijbels (1996) for more details. More discussion can be found in §5 below. We have the following consistency property for the variable selection procedure.

THEOREM 2. *Suppose Assumptions* A1–A7 *in the Appendix hold and that the bandwidth satisfies the requirements mentioned above.*
  (i) *If* $\alpha \cup d = \alpha_0$, *then* $\lim_{n \to \infty} \mathrm{pr}(CV_{\alpha,d} > CV_\alpha) \to 0$.
  (ii) *If* $\alpha_0 \subseteq \alpha$ *and* $d \notin \alpha_0$, *then* $\lim_{n \to \infty} \mathrm{pr}(CV_{\alpha,d} < CV_\alpha) \to 0$.

5. SIMULATION STUDY

We compare the leave-one-out, leave-m-out and separated crossvalidation by simulations. Since the asymptotic distribution of $\hat\theta$ can be used for variable selection, we also include it in the comparison study. The distributional result is that

$$n^{1/2}(\hat\theta - \theta^0) \to N(0, W_0^+ W_1 W_0^+)$$

in distribution as $n \to \infty$, where $W_0 = E[\{X - E(X|X^\top\theta^0)\}\{X - E(X|X^\top\theta^0)\}^\top g'(X^\top\theta^0)^2]$, $W_1 = E[\{X - E(X|X^\top\theta^0)\}\{X - E(X|X^\top\theta^0)\}^\top g'(X^\top\theta^0)^2\varepsilon^2]$ and W_0^+ denotes the Moore-Penrose inverse. The matrices in the asymptotic distribution can be estimated using kernel smoothing by $\hat W_0 = n^{-1}\sum_{i=1}^n (X_i - \hat\mu_i)(X_i - \hat\mu_i)^\top \hat d_i^2$ and $\hat W_1 = n^{-1}\sum_{i=1}^n (X_i - \hat\mu_i)(X_i - \hat\mu_i)^\top \hat d_i^2(Y_i - \hat a_i)^2$, where $\hat\mu_i = \sum_{j=1}^n K_h(X_{ij}^\top\hat\theta)X_j / \sum_{j=1}^n K_h(X_{ij}^\top\hat\theta)$ with $\hat a_i$ and $\hat d_i$ given by

$$\begin{pmatrix} \hat a_i \\ \hat d_i h \end{pmatrix} = \left\{ \sum_{j=1}^n K_h(X_{ij}^\top\hat\theta) \begin{pmatrix} 1 \\ X_{ij}^\top\hat\theta/h \end{pmatrix} \begin{pmatrix} 1 \\ X_{ij}^\top\hat\theta/h \end{pmatrix}^\top \right\}^{-1} \sum_{j=1}^n K_h(X_{ij}^\top\hat\theta) \begin{pmatrix} 1 \\ X_{ij}^\top\hat\theta/h \end{pmatrix} Y_j.$$

Based on the asymptotic distribution, a variable x_k is selected if $|\hat\theta_k| > 1{\cdot}96(c_{kk}/n)^{1/2}$, where c_{kk} is the (k, k)th entry of $\hat W_0^+ \hat W_1 \hat W_0^+$.

In the calculations below, we use a Gaussian kernel, since we find heuristically that it performs better in estimating the index parameter; see also Seifert & Gasser (1996). After (X_i, y_i) are standardized, the bandwidths are calculated by the rule-of-thumb of Silverman (1986, pp. 45–7) as follows. In (4), $h = 1{\cdot}06 s_{\theta^\top X_\alpha} n^{-1/5}$, where $s_{\theta^\top X_\alpha}$ is the sample standard deviation of $\theta^\top X_{i,\alpha}$. In (10), $h_1 = 1{\cdot}06 s_{\theta^\top X_\alpha} n^{-1/6}$ if x_d is continuous, and $h_1 = h$ if x_d is discrete. The computer code in Matlab for separated crossvalidation is available at http://www.stat.nus.edu.sg/~staxyc.

Example 1. We draw random samples with size $n = 50, 100$ and 200 respectively from a logistic regression model,

$$Y \sim \mathrm{Ber}\{l(X^\top\beta)\}, \qquad l(\mu) = \exp(\mu)/\{1 + \exp(\mu)\},$$

where $\beta = (3, 1{\cdot}5, 0, 0, 2, 0, 0, 0)^\top$. In the model, two designs were used for $X = (x_1, \ldots, x_8)^\top$. In design A, $(x_1, \ldots, x_6)^\top \sim N(0, \Sigma_6)$, where $\Sigma_p = (0{\cdot}5^{|i-j|})_{1 \leqslant i \leqslant j \leqslant p}$, and x_7, x_8, are independently Ber($0{\cdot}5$), independent of $(x_1, \ldots, x_6)^\top$. In design B, $x_{(2k)} = 2I(z_{(2k)} > 0) - 1$ and $x_{(2k-1)} = z_{(2k-1)}$ for $k = 1, 2, 3, 4$, where $Z = (z_1, \ldots, z_8)^\top \sim$

Table 1. *Example* 1. *Relative frequencies of correct selection*

Design	n	CV(1)	CV($0.25n$)	CV($0.5n$)	CV($0.75n$)	SCV	ASD
	50	0·00	0·00	0·18	0·38	0·41	0·27
A	100	0·29	0·46	0·58	0·63	0·66	0·44
	200	0·23	0·47	0·85	0·68	0·90	0·72
	50	0·30	0·37	0·46	0·46	0·51	0·32
B	100	0·37	0·43	0·69	0·77	0·81	0·65
	200	0·67	0·71	0·80	0·87	0·91	0·75

CV(m), leave-m-out crossvalidation; SCV, separated crossvalidation; ASD, asymptotic distribution method.

$N(0, \Sigma_8)$. Design A was investigated by Fan & Li (2001). A single-index model was fitted to the data and the variable selection methods were applied. The relative frequencies of correct selection of the variables among 100 replications are listed in Table 1.

Table 1 shows that the separated crossvalidation method outperforms all the other methods. Its efficiency is even comparable with the results of Fan & Li (2001), where the model is known up to unknown parameters. Also, the table shows that leave-m-out crossvalidation performs better if the data are split in the way according to Theorem 1.

Example 2. The Tobit model is an econometric model in which the dependent variable is censored. In the original model of Tobin (1958), for example, the response is expenditures on consumer durables, and the censoring occurs as negative values are unobservable, i.e.

$$Y = (\beta^\top X + 0.5\varepsilon)I(\beta^\top X + 0.5\varepsilon > 0),$$

where $I(.)$ is the indicator function. See also Nishiyama & Robinson (2005). We consider two designs. In design A, $X = (x_1, \ldots, x_{20})^\top \sim N(0, I_{20})$, and, in design B, $x_{(2k)} = 2I(z_{(2k)} > 0) - 1$ and $x_{(2k-1)} = z_{(2k-1)}$, for $k = 1, \ldots, 10$, where $Z \sim N(0, \Sigma_{20})$. The error term $\varepsilon \sim N(0, 1)$ is independent of X and $\beta = (1, 1, \ldots, 1, 0, \ldots, 0)^\top$, with its first l elements 1 and the others 0.

Table 2 shows the relative frequencies of selecting the variables correctly, based on 100 simulations. The number of covariates in this example is larger than in Example 1. As we mentioned at the beginning of § 4, having a large number of covariates will compromise the efficiency of leave-m-out crossvalidation, and this is clearly reflected in Table 2. In most of Table 2, CV($0.5n$) outperforms CV($0.75n$), suggesting that, for small to medium sample size, the way of splitting the data suggested by Theorem 1 is not applicable, because of the nature of nonparametric smoothing. In contrast, the separated crossvalidation is rather robust and performs better.

We also found from simulations not reported here that the choice of bandwidth is not so sensitive in variable selection as in nonparametric function estimation. This insensitivity was also observed in Cheng & Tong (1992). As mentioned in §3 and §4, other ways of estimating the single-index models can also be used in the procedure of variable selection and performs similarly, but some can be very time-consuming.

Table 2. *Example* 2. *Relative frequencies of correct selection*

Design	l	n	CV(1)	CV(0·5n)	CV(0·75n)	SCV	ASD
	5	50	0·08	0·36	0·02	0·84	0·08
A	10	50	0·17	0·49	0·14	0·60	0·14
	5	100	0·32	0·82	0·78	0·99	0·26
	10	100	0·56	0·90	0·93	1·00	0·33
	5	50	0·12	0·38	0·00	0·85	0·03
B	10	50	0·14	0·32	0·00	0·59	0·17
	5	100	0·42	0·92	0·93	0·97	0·10
	10	100	0·55	0·92	0·90	0·99	0·37

CV(m), leave-m-out crossvalidation; SCV, separated crossvalidation; ASD, asymptotic distribution method.

6. Applications to two real datasets

Example 3. *The Swiss banknotes data.* The data contain 6 explanatory variables which are certain measurements of Swiss banknotes, called Length, Left, Right, Bottom, Top and Diagonal, and denoted by $x_1, \ldots, x_6$ respectively. The response variable Y is coded as 0 or 1, indicating whether a banknote is genuine or not. There are 200 banknotes. The first 100 banknotes are genuine, and the others are counterfeit.

The separated crossvalidation selects x_4, x_5 and x_6 for a single-index model. The fitted values from single-index models based on all variables and on the selected variables are plotted in Fig. 1. The single-index parameters are estimated respectively as $\theta_{\mathrm{ALL}} = (-0.1597, 0.4638, -0.1549, 0.5699, 0.2922, -0.5703)^\top$ when all the variables are used and $\theta_{\mathrm{S}} = (0.8006, 0.3011, -0.5181)^\top$ when the selected variables are used. Both models fit the data very well. To compare their prediction capabilities, we split the data randomly into a training set comprising 50 counterfeit banknotes and 50 genuine banknotes, and a test set containing the rest. We estimate the model with the training set, apply the estimated model to the test set and calculate the number of misspecifications. With different covariate sets, the average numbers of misspecifications based on 10 000 replications of this random splitting are given in Table 3. A single-index model with variables selected by the principle component analysis is also compared; see Härdle & Simar (2003). Apparently, separated crossvalidation gives the best results.

Example 4. *Ozone concentration data.* In this example, we study the relationship between ozone concentration, Y, and radiation level, R, temperature, T, and wind speed, W. From May to September 1973, 111 observations were taken daily in New York. We include the direct interaction between any two covariates in the model as covariates. As a consequence, we have 9 covariates $X = (x_1, \ldots, x_9)^\top = (R, T, W, R^2, R*T, R*W, T^2, T*W, W^2)^\top$. After standardising Y and $x_k, k = 1, \ldots, 9$, we apply separated crossvalidation to

Table 3. *Swiss banknotes data. Average number of misspecifications*

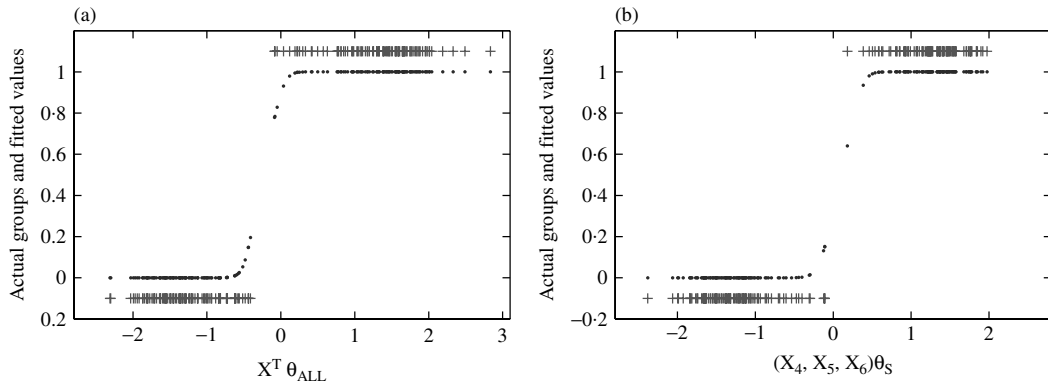| Method	Selected variables	Ave. no. of misspecifications
All variables	$x_1, x_2, x_3, x_4, x_5, x_6$	0·5787
Crossvalidation	x_1, x_4, x_5, x_6	0·6223
Seperated crossvalidation	x_4, x_5, x_6	0·5100
Principle component anal.	x_5, x_6	0·5411

Fig. 1. Estimation of single-index models for the banknotes data (a) based on all covariates, (b) based on only the selected variables. In both panels, '+' denotes the observations and '.' the fitted values. For easy visualization, we re-scaled the values of the observed Y.
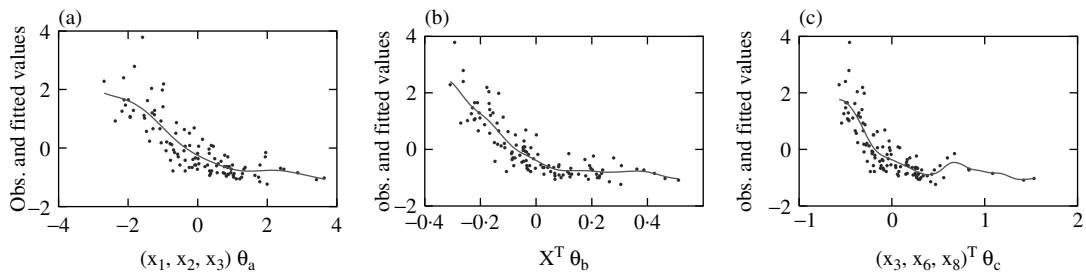

Fig. 2. Estimation of single-index models for the ozone concentration data (a) based on the original covariates (R, T, W), (b) based on the extended variables, (c) based on the selected variables. In all panels, '.' denotes the observations and '−' the fitted values.

Table 4. *Ozone concentration data. Average prediction errors*

Method	Selected variables	Ave. prediction error
All original variables	x_1, x_2, x_3	0·3643
All extended variables	$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9$	0·3621
Separated crossvalidation	x_3, x_6, x_8	0·3403

the data, thereby selecting variables x_3, x_6 and x_8 with estimated index parameter $\theta_c = (0{\cdot}8486, -0{\cdot}0992, -0{\cdot}5196)^\top$. Single-index models having X and the original three variables as predictors are also investigated and the estimated parameters are $\theta_b = (0{\cdot}2147, 0{\cdot}1544, -0{\cdot}7541, -0{\cdot}1245, -0{\cdot}0029, -0{\cdot}0607, -0{\cdot}2292, 0{\cdot}5183, 0{\cdot}1448)^\top$ and $\theta_a = (0{\cdot}3443, 0{\cdot}7051, -0{\cdot}6199)^\top$ respectively. The fitted values are plotted in Fig. 2. To compare the prediction capabilities of single-index models with different covariates, we again split the data randomly into two sets, this time with the training set comprising 56 observations and the test set containing the remaining 55 observations. The prediction errors are defined as the mean residual sum of squares. The results in Table 4 are based on 10 000 replications of this random splitting. Again, Table 4 indicates that separated crossvalidation does best.

APPENDIX

Assumptions and Proofs

First we introduce some notation. Let $\gamma_\alpha(.|\theta)$ and $\gamma_0(.)$ be the density functions of $X_\alpha^\top \theta$ and $X^\top \theta^0$ respectively. Let $\mathcal{U} = \{X^\top \theta^0 : X \in A\}$, $\mathcal{U}_\alpha = \{X_\alpha^\top \theta_\alpha^0 : X \in A\}$, $\mathcal{D}_\alpha = \{x_\alpha : x \in A\}$, where A is defined in §2, $K_1 = \int t^2 K(t) dt$, $K_2 = \int K^2(t) dt$. For any $\alpha \supset \alpha_0$, let $\mu_\alpha(x|\theta) = E(X_\alpha | X_\alpha^\top \theta_\alpha = x_\alpha^\top \theta_\alpha)$, $v_\alpha(x|\theta) = \mu_\alpha(x|\theta) - x_\alpha$, $W_\alpha = \int_A v_\alpha(x|\theta) v_\alpha(x|\theta)^\top g'(x^\top \theta^0)^2 f(x) dx$, and $U_{\alpha,j} = W_\alpha^{\frac{1}{2}+} g'(X_j^\top \theta^0) v_\alpha(X_j|\theta^0)$. Let $\Theta_{n,\alpha} = \{\theta_\alpha : ||\theta|| = 1, ||\theta - \theta^0|| \leqslant rn^{-1/2}\}$, $\mathcal{H}_n = \{h : r_1 n^{-1/5} \leqslant h \leqslant r_2 n^{-1/5}\}$ for some $r > 0$ and $0 < r_1 < r_2 < \infty$. Denote $\Theta_{n,\{1,\ldots,p\}}$ by Θ_n.

We need the following regularity conditions to prove the theorems.

Assumption A1. X has a compact support in $\mathcal{R}^p$, and for any $\alpha \supseteq \alpha_0$, $\inf_{x \in A, \theta \in \Theta_{n,\alpha}} \gamma_\alpha(x^\top \theta | \theta) > 0$.

Assumption A2. Link function $g(.)$ has bounded third-order derivatives on $\mathcal{U}$.

Assumption A3. K is a symmetric density function with a compact support. Without loss of generality, we assume that $K_1 = 1$ and the Fourier transform of $K(t)$ is absolutely integrable.

Assumption A4. $E(\varepsilon_i | X_i) = 0$ and $E(\varepsilon_i^2 | X_i) = \sigma^2$.

Assumption A5. For any α, $\sup_{m \to \infty} \sup_s \left\| m^{-1} \sum_{j \in s} U_{\alpha,j} U_{\alpha,j}^\top - I_{d_\alpha} + \theta_\alpha^0 \theta_\alpha^{0\top} \right\| = o_p(1)$, and

$$\hat{\theta}_\alpha^{\backslash s} - \theta_\alpha^0 = n_c^{-1} W_\alpha^+ \sum_{j \notin s} g'(X_{j,\alpha}^\top \theta_\alpha^0) v_\alpha^\top(X_j | \theta^0) \varepsilon_j + \delta_n^{\backslash s}, \tag{A1}$$

where I_{d_α} is the identity matrix and $\delta_n^{\backslash s} = o_p(n^{-1/2})$ uniformly for all s.

Assumption A6. For any $\alpha \subset \alpha_0$, $g_\alpha(v|\theta) = E(Y|X_\alpha^\top \theta = v^\top \theta)$ has bounded first-order derivative with respect to $\theta \in \Theta_{n,\alpha}$; $\sigma_\alpha^2(\theta) := E\{g_\alpha(X_\alpha|\theta) - Y\}^2$ with $\inf_{\theta \in \Theta_{n,\alpha}} \sigma_\alpha^2(\theta) > \sigma^2$.

Assumption A7. For any $\alpha \cup d \supseteq \alpha_0$, if x_d is continuous, the joint density function of $(X_\alpha^\top \theta, x_d)$, $f_{X_\alpha^\top \theta, x_d}(u^\top \theta, v)$, is uniformly bounded away from zero for $\theta \in \Theta_{n,\alpha}$, $u \in \mathcal{D}_\alpha$ and $v \in A_d := \{x_d : (x_1, \ldots, x_d, \ldots, x_p) \in A\}$; if x_d is discrete, the conditional density function of $X_\alpha^\top \theta$ given $x_d = v$, $f_{X_\alpha^\top \theta | x_d = v}(.)$, satisfies $\inf_{u \in \mathcal{D}_\alpha, \theta \in \Theta_{n,\alpha}} f_{X_\alpha^\top \theta_\alpha^0 | x_d = v}(u^\top \theta) > 0$.

Assumptions A1–A4 are required for the consistency of estimations; see Härdle et al. (1993) and Xia & Tong (2006). For Assumption A5, while Xia & Tong (2006) has proved (A1) with $\delta_n^{\backslash s} = o_p(n^{-1/2})$ for any given s, the uniform convergence rate here is necessary to guarantee the validity of leave-m-out crossvalidation and is parallel to the balanced block design assumption in linear regression; see Zhang (1993). The requirement on the Fourier transform of $K(t)$ in Assumption A3 is to ensure the difference between the minimum average variance estimate $\hat{\theta}$ and θ^0 admits the form in (A1). Many kernel functions meet this demand, such as the triweight kernel.

Gaussian kernel is also permissible at the expense of a longer proof. Assumption A6 is a common assumption if the optimal model exists and is unique; see Yao & Tong (1994). Assumption A7 is used to ensure the denominators of kernel smoothers is bounded away from zero.

We outline the proofs here. Detailed derivation is available upon request.

Proof of Theorem 1. Write $Y_i - \hat{g}_\alpha^{\backslash s}(X_i|\hat{\theta}_\alpha^{\backslash s}) = Y_i - \hat{g}_\alpha^{\backslash s}(X_i|\theta_\alpha^0) + \hat{g}_\alpha^{\backslash s}(X_i|\theta_\alpha^0) - \hat{g}_\alpha^{\backslash s}(X_i|\hat{\theta}_\alpha^{\backslash s})$. Then, by (A1), Lemma 7 and Lemma 9 in Xia & Tong (2006), we can prove that

$$\mathrm{CV}_\alpha(m) = m^{-1}\binom{n'}{m}^{-1}\sum_{i,s}'\{Y_i - \hat{g}_{\alpha_0}^{\backslash s}(X_i|\theta_\alpha^0)\}^2 + m^{-1}\binom{n'}{m}^{-1}\left\{\frac{2}{n_c}\sum_{i,s}'\varepsilon_i U_{\alpha,i}^\top \sum_{j\notin s}U_{\alpha,j}\varepsilon_j\right.$$

$$\left.+\frac{1}{n_c^2}\sum_{i,s}'U_{\alpha,i}^\top\left(\sum_{j\notin s}U_{\alpha,j}\varepsilon_j\right)\left(\sum_{j\notin s}U_{\alpha,j}^\top\varepsilon_j\right)U_{\alpha,i}\right\} + o_p\left(\frac{1}{n}\right)$$

$$:= \mathrm{RSS}(m) + m^{-1}\binom{n'}{m}^{-1}(T_1 + T_2) + o_p\left(\frac{1}{n}\right),$$

where the term $o_p(n^{-1})$ is quantified by computing corresponding second moments using Assumptions A1–A4 and the facts that $\{X_i, Y_i\}$ are independent observations. Let $\mathrm{e}_{s^c} = (\varepsilon_j)_{j\notin s}$ and $U_{\alpha,s^c} = (U_{\alpha,j_1}, \ldots, U_{\alpha,j_{n_c}})^\top$, where $j_i \notin s$. By Assumptions A2 and A5, we have

$$T_2 = \frac{m}{n_c^2}\sum_s'\left(\mathrm{e}_{s^c}^\top U_{\alpha,s^c}\right)\left(I_{d_\alpha} - \theta_\alpha^0\theta_\alpha^{0\top}\right)\left(\mathrm{e}_{s^c}^\top U_{\alpha,s^c}\right)^\top\{1 + o_p(1)\}$$

$$= \frac{m}{n_c^2}\sum_s'\left(\mathrm{e}_{s^c}^\top U_{\alpha,s^c}\right)\left(\mathrm{e}_{s^c}^\top U_{\alpha,s^c}\right)^\top\{1 + o_p(1)\}.$$

The last equality holds since $U_{\alpha,j}^\top\theta_\alpha^0 = 0$ for all j. Combinatoric calculation leads to

$$T_1 + T_2 = \binom{n'-2}{m-1}\frac{1}{n_c^2}\left\{(2n + n' - 3m - 1)\left(\sum_i'U_{\alpha,i}\varepsilon_i\right)^\top\left(\sum_i'U_{\alpha,i}\varepsilon_i\right)\right.$$

$$\left.+(3m - 2n)\sum_i'U_{\alpha,i}^\top U_{\alpha,i}\varepsilon_i^2\right\}\{1 + o_p(1)\}.$$

Note that $n_c = n - m$, and both m/n' and m/n tend to c. By the law of large numbers and Assumption A5, $n'^{-1}\sum_i\varepsilon_i^2 U_{\alpha,i}^\top U_{\alpha,i} \to \sigma^2 E\{tr(I_{d_\alpha} - \theta_\alpha^0\theta_\alpha^{0\top})\} = \sigma^2(d_\alpha - 1)$ in probability and $n'^{-1}(\sum_i'U_{\alpha,i}\varepsilon_i)^\top(\sum_i'U_{\alpha,i}\varepsilon_i) \to \sigma^2\chi^2(d_\alpha - 1)$ in distribution. Therefore,

$$n\{\mathrm{CV}_\alpha(m) - \mathrm{RSS}(m)\} \to \sigma^2\left\{3\chi^2(d_\alpha - 1) + \frac{(3c - 2)(d_\alpha - 1)}{(1 - c)}\right\} \tag{A2}$$

in distribution. As $\mathrm{RSS}(m)$ is independent of α and $U_{\alpha_0,i}$ is a subvector of $U_{\alpha,i}$ if $\alpha \supset \alpha_0$, thus the proof is completed. □

Proof of Theorem 2. First, we quantify CV_α. If $\alpha \supset \alpha_0$, let $m = 1$ in (A2) and we have $\mathrm{CV}_\alpha = n'^{-1}\sum_i'\{Y_i - \hat{g}_\alpha^{\backslash i}(X_i|\theta_\alpha^0)\}^2 + O_p(n'^{-1})$. Mimicking the steps leading to Lemma 1 in Yao & Tong (1994), we have

$$\mathrm{CV}_\alpha = \frac{1}{n'}\sum_i'\varepsilon_i^2 + \frac{c_1}{n'h} + c_2h^4 + o_p\left(\frac{1}{n'h}\right), \tag{A3}$$

where $c_1 = \sigma^2 K_2 E\{\gamma_0^{-1}(X^\top\theta^0)\} = \sigma^2 K_2 L(\mathcal{U})$, $c_2 = Eg''^2(X^\top\theta^0)/4$ and $L(\mathcal{U})$ is the Lebesgue measure of $\mathcal{U}$.

If $\alpha \cup d = \alpha_0$, Step 2 in § 4 with (A1) indicates that $\hat{\theta}_\alpha^{\backslash i} - \theta_\alpha^0 = O_p(n^{-1/2})$ uniformly in i. Then, by Theorem 6 in Masry (1996) and similar arguments leading to (A3), we have

$$\mathrm{CV}_\alpha = n^{'-1} \sum_i \{Y_i - \hat{g}_\alpha^{\backslash i}(X_i|\theta_\alpha^0)\}^2 + o_p(1) = \sigma^2(\theta_\alpha^0) + o_p(1). \tag{A4}$$

Next, we consider $\mathrm{CV}_{\alpha,d}$ with $\alpha \cup d \supseteq \alpha_0$. If x_d is continuous, then, similarly to (A3), $\mathrm{CV}_{\alpha,d}$ admits the following expansion

$$\mathrm{CV}_{\alpha,d} = \frac{1}{n'} {\sum_i}' \{Y_i - \hat{g}_{\alpha,d}^{\backslash i}(X_i|\theta_\alpha^0)\}^2 + o_p\left(\frac{1}{n'h_1^2}\right) = \frac{1}{n'} {\sum_i}' \varepsilon_i^2 + \frac{c_3}{n'h_1^2} + 4c_2 h_1^4 + o_p\left(\frac{1}{n'h_1^2}\right), \tag{A5}$$

where $c_3 = \sigma^2 K_2^2 L(\mathcal{U}_\alpha) L(A_d)$ with A_d defined in Assumption A7.

For discrete x_d with M values, $v_1, \ldots, v_M$, we classify $\{(X_i, Y_i)\}_{i=1}^{n'}$ into M groups based on the value of $x_d : i \in G_k \Leftrightarrow x_{id} = v_k$. Let n_k be the number of elements in G_k and $n_k = O(n')$, $k = 1, \ldots, M$. If $i \in G_k$, by (10), $\hat{g}_{\alpha,d}^{\backslash i}(X_i|\hat{\theta}_\alpha^{\backslash i})$ equals $\hat{g}_\alpha^{\backslash i}(X_i|\hat{\theta}_\alpha^{\backslash i})$, which is defined in (6) with θ replaced by $\hat{\theta}_\alpha^{\backslash i}$ and subindex $\{j \notin s\}$ by $\{j \in G_k, j \ne i\}$. Thus $\mathrm{CV}_{\alpha,d} = n^{'-1} \sum_{k=1}^M n_k \mathrm{CV}_\alpha^k$, where $\mathrm{CV}_\alpha^k := n_k^{-1} \sum_{i \in G_k} \{Y_i - \hat{g}_\alpha^{\backslash i}(X_i|\hat{\theta}_\alpha^{\backslash i})\}^2$ is the $\mathrm{CV}_\alpha(1)$ in (7) using data $\{(X_{i\alpha}, Y_i) : i \in G_k\}$. Since $\alpha \cup d \supseteq \alpha_0$, $E(Y|X)$ only depends on X_α within each G_k. Therefore, similarly to (A3), by Assumption A7 we have

$$\mathrm{CV}_\alpha^k = \frac{1}{n_k} \sum_{i \in G_k} \varepsilon_i^2 + c_4 h_1^4 + \frac{\sigma^2 K_2}{n_k h_1} L(\mathcal{U}_\alpha^k) + o_p\left(\frac{1}{n_k h_1}\right), \; k = 1, \ldots, M,$$

where $c_4 = E\{g^{''2}(X_\alpha^\top \theta_\alpha^0)|x_d = v_k\}/4$, and $\mathcal{U}_\alpha^k$ is the support of $X_\alpha^\top \theta_\alpha^0$ given that $x_d = v_k$. Therefore,

$$\mathrm{CV}_{\alpha,d} = \frac{1}{n'} {\sum_i}' \varepsilon_i^2 + \frac{\sigma^2 K_2}{n'h_1} \sum_{k=1}^M L(\mathcal{U}_\alpha^k) + c_2 h_1^4 + o_p\left(\frac{1}{n'h_1}\right). \tag{A6}$$

Note that if x_d is redundant, i.e. $\beta_d^0 = 0$, then $\mathcal{U}_\alpha^k$ is also the support of $X^\top \theta^0$ given that $x_d = v_k$. By the discussion about the identification of single-index models with discrete covariates (Ichimura, 1993), we have $\sum_{k=1}^M L(\mathcal{U}_\alpha^k) > L(\mathcal{U})$.

By the conditions on h and h_1, Theorem 2 follows from (A3), (A4), (A5) and (A6).

$\square$

REFERENCES

AKAIKE, H. (1974). A new look at statistical model identification. *IEEE Trans. Auto. Cont.* **19**, 716–23.
ALTHAM, P. M. E. (1984). Improving the precision of estimation by fitting a generalized linear model and quasi-likelihood. *J. R. Statist. Soc.* B **46**, 118–9.
CHENG, B. & TONG, H. (1992). On consistent nonparametric order determination and chaos (with Discussion). *J. R. Statist. Soc.* B **54**, 427–49.
FAN, J. & GIJBELS, I. (1996). *Local Polynomial Modeling and its Applications*. London: Chapman and Hall.
FAN, J. & LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.
GAO, J. & TONG, H. (2004). Semiparametric nonlinear time series model selection. *J. R. Statist. Soc.* B **66**, 321–36.
HÄRDLE, W., HALL, P. & ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157–78.
HÄRDLE, W. & SIMAR, L. (2003). *Applied Multivariate Statistical Analysis*. Berlin: Springer.
HÄRDLE, W. & STOKER, T. M. (1989). Investigating smooth multiple regression by method of average derivatives. *J. Am. Statist. Assoc.* **84**, 986–95.
HOROWITZ, J. L. & HÄRDLE, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *J. Am. Statist. Assoc.* **91**, 1632–40.
HRISTACHE, M., JUDITSKI, A. & SPOKOINY, V. (2001). Direct estimation of the index coefficients in a single-index model. *Ann. Statist.* **29**, 595–623.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Economet.* **58**, 71–120.

Klein, R. W. & Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61**, 387–421.

Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal.* **17**, 571–99.

Miller, A. J. (2002). *Subset Selection in Regression*. London: Chapman and Hall.

Naik, P. A. & Tsai, C.-L. (2001). Single-index model selections. *Biometrika* **88**, 821–32.

Nishiyama, Y. & Robinson, P. M. (2005). The bootstrap and the Edgeworth correction for semiparametric averaged derivatives. *Econometrica* **73**, 903–48.

Powell, J. L., Stock, J. H. & Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57**, 1403–30.

Ruppert, D. & Wand, M. P. (1994). Multivariate locally weighted least-squared regression. *Ann. Statist.* **22**, 1346–70.

Seifert, B. & Gasser, T (1996). Finite sample variance of local polynomials: analysis and solutions. *J. Am. Statist. Assoc.* **91**, 267–75.

Shao, J. (1993). Linear model selection by cross-validation. *J. Am. Statist. Assoc.* **88**, 486–94.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Stone, M. (1974). Cross-Validation choice and assessment of statistical prediction (with Discussion). *J. R. Statist. Soc.* B **36**, 111–47.

Tjøstheim, D. & Auestad, B. H. (1994). Nonparametric identification of nonlinear time series: selecting significant lags. *J. Am. Statist. Assoc.* **89**, 1410–9.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica* **26**, 24–36.

Xia, Y. & Tong, H. (2006). On efficiencies of estimations for single index models. In *Frontiers in Statistics*, Ed. J. Fan & K. Hourl, pp. 63–86, New York: London: World Scientific/Imperial College Press.

Xia, Y., Tong, H., Li, W. K. & Zhu, L. (2002). An adaptive estimation of dimension reduction space (with Discussion). *J. R. Statist. Soc.* B **64**, 363–410.

Yao, Q. & Tong, H. (1994). On subset selection in nonparametric stochastic regression. *Statist. Sinica.* **4**, 51–70.

Zhang, P. (1993). Model selection via multifold cross validation. *Ann. Statist.* **21**, 299–313.