Science in China Series A: Mathematics
Sep., 2007, Vol. 50, No. 9, 1289–1302
www.scichina.com
www.springerlink.com

© 2007 SCIENCE IN CHINA PRESS
Springer

On spline approximation of sliced inverse regression

Li-ping ZHU[†] & Zhou YU

Department of Statistics, East China Normal University, Shanghai 200062, China
(email: lpzhu@stat.ecnu.edu.cn)

Abstract The dimension reduction is helpful and often necessary in exploring the nonparametric regression structure. In this area, Sliced inverse regression (SIR) is a promising tool to estimate the central dimension reduction (CDR) space. To estimate the kernel matrix of the SIR, we herein suggest the spline approximation using the least squares regression. The heteroscedasticity can be incorporated well by introducing an appropriate weight function. The root-n asymptotic normality can be achieved for a wide range choice of knots. This is essentially analogous to the kernel estimation. Moreover, we also propose a modified Bayes information criterion (BIC) based on the eigenvalues of the SIR matrix. This modified BIC can be applied to any form of the SIR and other related methods. The methodology and some of the practical issues are illustrated through the horse mussel data. Empirical studies evidence the performance of our proposed spline approximation by comparison of the existing estimators.

Keywords: asymptotic normality, spline, Bayes information criterion, dimension reduction, sliced inverse regression, structural dimensionality

MSC(2000): 62H12, 62J02

1 Introduction

Consider a regression problem with a response Y and a p-dimensional predictor vector $X = (X_1, \dots, X_p)^{\mathrm{T}}$. A high dimensional predictor vector often makes the statistical analysis difficult. To tackle this problem, we consider a sub-dimensional model proposed in [1, 2] such that

$$Y \perp\!\!\!\perp X | B^{\mathrm{T}} X, \tag{1.1}$$

where $\perp\!\!\!\perp$ stands for independence. This means that $B^{\mathrm{T}}X$ is a sufficient statistic for the regression of Y on X. Note that the model specification (1.1) does not uniquely determine the $p \times K$ matrix B. To address this, Cook[2] defined the central dimension reduction subspace (CDR), indicated with $S_{Y|\boldsymbol{x}}$, which is the intersection of all dimension reduction subspaces satisfying (1.1). The existence of the CDR space has been investigated in [3]. In this article, we assume the existence of the CDR space so K is the smallest of all possible integer for (1.1) to hold. Sliced inverse regression (SIR, see [1]) is a promising tool for identifying and estimating CDR

Received November 14, 2005; accepted January 8, 2007; published online June 26, 2007
DOI:10.1007/s11425-007-0085-5

[†] Corresponding author

This work was supported by the special fund (2006) for selecting and training young teachers of universities in Shanghai (Grant No. 79001320) and an FRG grant (FRG/06-07/I-06) from Hong Kong Baptist University, China, and a grant (HKU 7058/05P) from the Research Grants Council of Hong Kong, China

subspace by using the conditional mean of X given Y. Let $\Lambda = \mathrm{Cov}(E(X|Y))$. Under the linearity condition, that is,

$$E(X|P_{S_{Y|\boldsymbol{x}}}X) = P_{S_{Y|\boldsymbol{x}}}X, \tag{1.2}$$

where $P_{(\cdot)}$ stands for the projection operator in the standard inner product (see [2]). It can be shown that $\mathrm{Span}(\Lambda) \subseteq S_{Y|\boldsymbol{x}}$ in $[1, 2]$ where $\mathrm{Span}(\Lambda)$ denotes the space spanned by the column vectors of Λ. Extension of SIR and a review of other inference techniques are given in [2].

Several methods have been proposed in the literature to estimate Λ based on the local averages or the local covariances computed from the points with neighboring Y. The so-called "slicing" estimator proposed in [1] is a very simple and useful estimation scheme which has become one of the standard estimations in this area. For the consistency of the slicing estimation, Hsing and Carroll[4] and Zhu and Ng[5] proved the asymptotic normality of the SIR matrix estimator. Note that the local smoothing methods are also applicable. Therefore, Zhu and Fang[6] established the asymptotical normality for the kernel estimation of the SIR. When the bandwidth is confined in a range from a rate $n^{-1/2}$ to $n^{-1/(2m)}$, the asymptotic normality holds true, where $m \geqslant 2$ presents the degree of smoothness of the inverse regression functions. Fung, He and Liu, et al.[7] considered using the canonical variables from the design space whose correlations with a spline basis in the response space are significant. This could be viewed as a variant of SIR. The idea of using the splines to estimate the SIR was mentioned briefly in the discussion of Li[1] by Kent[8]. The relationship between the SIR and the canonical correlation was also explored by Chen and Li[9].

However, the aforementioned estimation procedures might lose efficiency in the heteroscedastic models. To circumvent this issue, we suggest the least square spline approximation with the implement of the weight function. Much research work has been devoted to choosing an appropriate weight function in literature. Among these, see, Cook and Weisberg[10]. The root-n asymptotic normality of the spline approximation of the SIR matrix can be achieved for a wide range choice of knots in this article. This phenomenon is essentially the same as the kernel estimation in [6].

The second part of the paper is to consider a criterion to determine the dimensionality K that is asymptotically valid for rather general predictors. In practice, the estimation of the matrix B is independent of the structural dimension K. However, the determination of an available dimension is also crucial to achieve the goal of reducing the dimensionality. For SIR, Li[1] suggested a sequential chi-squared test procedure to determine the dimension, that is, evaluating K by successively testing the nullity of the $p - k$ smallest eigenvalues, starting by $k = 0$. He proposed a chi-square test under the normal distribution assumption. Schott[11] considered using both the first and second moments of the conditional distribution of X given Y, and developed a chi-square test which is valid for any elliptically symmetric predictor distribution. Velilla[12] and Fung, He and Liu, et al.[7] also considered a sequential test. Bura and Cook[13] suggested a general weighted chi-squared sequential test that does not require the normality of covariates. Ferré[14] proposed a new approach for the SIR and the pHd to determine the dimension of B. He suggested that the determination of the structural dimension is measured by the squared trace correlation between the subspaces of the CDR space and their estimates. However, the sequential test might be inefficient because the significant levels at each step do not determine the significant level of the entire procedure, which remains unknown. Furthermore, the retained dimension depends on the choice of the significant level.

An alternative approach to the problem of choosing K is by studying the quality of estimation through a convenient discrepancy measure. This is the case in model or variable selection in linear regression, where numerous criteria have been proposed, including Mallows's C_p (see [15]), the Akaike information criterion (AIC, see [16]), the Bayes information criterion (BIC, see [17]). In this dimension reduction sense where the link function may be nonlinear and unknown, Zhu, Miao and Peng[18] suggested a new procedure of the BIC type for the determination of dimensions. This is a general method which can be applied to many dimension reduction tools. However, in their criterion, the choice of the penalty function C_n is difficult to choose in practice. Therefore, Zhu and Zhu[19] proposed to choose $C_n = \ln(n)$. It works efficiently in most cases. This methodology has some merits: only convergence of the estimator of the relevant matrix is needed and the estimator of the dimension is consistent. Borrowing the idea of the BIC, we propose another modified BIC based on the eigenvalues of the SIR matrix whose performance will be illustrated by simulations.

The rest of this paper is organized as follows. In the next section, the asymptotical normality for SIR matrix in two cases, the homoscedastic model and the heteroscedastic model, are achieved at the root-n rate. So are the non-zeros eigenvalues and their corresponding eigenvectors. We introduce our motivation of proposing the BIC method in sec. 3. The consistency of the determination of the structural dimension is also discussed. Some illustrative examples by simulation and real data application are reported in sec. 4 to show the performance of the spline estimation and to compare with the existing methods. The tedious proofs are delayed in the Appendix.

2 Spline approximation

2.1 Asymptotic properties of SIR matrix

In this section, we will establish the asymptotical normality for the spline approximation of the SIR matrix. Firstly, we introduce some notations. Denote by $Q(y)$, the distribution function of Y. Its corresponding density function of Y is $q(y)$. Let X and its independent copies x_j be

$$X = (X_1, \dots, X_p)^{\mathrm{T}}, \quad x_j = (x_{1j}, \dots, x_{pj})^{\mathrm{T}}, \quad j = 1, \dots, n.$$

Without loss of generality, we assume that $Q(y)$ has a support on [0,1] and X is standardized, that is, $E(X) = 0$ and $\mathrm{Cov}(X) = I_p$. This is due to the fact that the CDR subspace based on the standardized variable can be easily back transformed to that based on the original predictor vector X (see [2]). Then the SIR matrix is defined as follows:

$$\Lambda = \mathrm{Cov}(E(X|Y)) = E(E(X|Y)E(X^{\mathrm{T}}|Y)).$$

Our objective is then to estimate, based on (x_j, y_j)'s, the SIR matrix Λ, its eigenvalues and the corresponding eigenvectors.

2.1.1 Homoscedastic model case

For simplicity, we first consider the homoscedastic model. To estimate the SIR matrix, we herein suggest the spline approximation. There are a large amount of papers on regression splines. Among these, Agarwal and Studden[20] and Huang and Studden[21] considered the rates of convergence and the connection between splines and kernels in the univariate case. More recently, Zhou, Shen and Wolfe[22] studied the asymptotic distribution of the regression spline.

Specifically, a spline is defined as a piecewise polynomial that is smoothly connected at its knots. More specifically, for any fixed integer $m > 1$, denote $S(m, \underline{t})$ to be the set of spline functions with knots $\underline{t} = \{0 = t_0 < t_1 < \cdots < t_{k_0+1} = 1\}$. Then for $m \geqslant 2$, $S(m, \underline{t}) = \{s \in C^{m-2}[0,1] : s(y)$ is a polynomial degree $(m-1)$ on each subinterval $[t_i, t_{i+1}]\}$. Here k_0 is referred to as the number of internal knots. The common choices of m are 2 for linear splines, 3 for quadratic splines and 4 for cubic splines. In this paper, we use t_i as uniform partitions of $[0, 1]$ or as the (i/k_0)-th quantile of the observed y values so they are uniform in percentile ranks. The former is used in all of our empirical investigations reported in sec. 4. The minimum size of partition k_0 should be chosen such that $k_0 + m \geqslant K$, where K is the number of effective dimensions being sought. Here we do not need the exact value K, but a reasonable upper bound will also be helpful. Consider the spline estimation of the condition expectation, $f(y) =: E(X|Y = y)$ based on the sample. The j-th element $E(X_j|Y = y)$ is denoted by $f_j(y)$ throughout the paper. To estimate $f_j(y)$, we use the least squares criterion. For each $1 \leqslant j \leqslant p$, the estimator of order m for $f_j(y)$ is defined to be the least squares minimizer $\hat{f}_j(y) \in S(m, \underline{t})$ corresponding to

$$\sum_{i=1}^{n}(x_{ji} - \hat{f}_j(y_i))^2 = \min_{s_j(y)\in S(m,\underline{t})} \sum_{i=1}^{n}(x_{ji} - s_j(y_i))^2.$$

It is convenient to express the elements in $S(m, \underline{t})$ in terms of B-splines. For any fixed m and $\underline{t}$, let

$$N_{i,m}(y) = (t_i - t_{i-m})[t_{i-m}, \ldots, t_i](t - y)_+^{m-1}, \quad i = 1, \ldots, J = k_0 + m,$$

where $[t_{i-m}, \ldots, t_i]g$ denotes the mth-order divided difference of the function g and $t_i = t_{\min(\max(i,0),k_0+1)}$ for any $i = 1 - m, \ldots, J$. Then $\{N_{i,m}(\cdot)\}_{i=1}^J$ forms a basis for $S(m, \underline{t})$ (see [23, p. 124]); that is, for any $s(y) \in S(m, \underline{t})$, there exists an $\underline{\alpha}$ such that $s(y) = \underline{\alpha}' N_m(y)$, where $N_m(y) = (N_{1,m}(y), \ldots, N_{J,m}(y))^{\mathrm{T}}$. For notational convenience, in the sequel, $N_m(\cdot)$ will be abbreviated as $N(\cdot)$. Write $G_{J,n} = \frac{1}{n}\sum_{j=1}^n N(y_j)N^{\mathrm{T}}(y_j)$ and its expectation $G(q) = E(N(Y)N^{\mathrm{T}}(Y))$.

The kl-th element of Λ can be written as $\lambda_{kl} = E(E(X_k|Y)E(X_l|Y)), 1 \leqslant k, l \leqslant p$. Its corresponding element $\lambda_{n,kl}$ is defined as, by replacing the unknowns by their estimators,

$$\begin{aligned}\lambda_{n,kl} &= \frac{1}{n}\sum_{j=1}^{n} \hat{f}_k(y_j)\hat{f}_l(y_j) = \frac{1}{n}\sum_{j=1}^{n} \hat{E}(X_l|y_j)\hat{E}(X_k|y_j)\\ &= \frac{1}{n}\sum_{j=1}^{n}\left(\sum_{i=1}^{n} N^{\mathrm{T}}(y_j)G_{J,n}^{-1}N(y_i)x_{li}\right)\left(\sum_{i_1=1}^{n} N^{\mathrm{T}}(y_j)G_{J,n}^{-1}N(y_{i_1})x_{ki_1}\right).\end{aligned}$$

To present our main results, we adopt the vectorization of a matrix. For a symmetric $(p \times p)$ matrix $C = (c_{kl})_{p\times p}$, let $\mathrm{Vech}(C) = (c_{11}, \ldots, c_{p1}, c_{22}, \ldots, c_{p2}, \ldots, c_{pp})$ be a $p(p+1)/2$ dimensional vector.

We are now in the position to introduce the theoretical results. Define the kl-th element of matrix $H(X, Y)$ as

$$H_{kl}(X, Y) = f_l(Y)X_k + f_k(Y)X_l + f_k(Y)f_l(Y) - 3E(f_k(Y)f_l(Y))$$

and for any $\lambda \in \mathbb{R}^{p(p+1)/2}$, $\sigma^2_\lambda(\Lambda) = \lambda^{\mathrm{T}}\mathrm{Cov}(\mathrm{Vech}(H(X,Y)))\lambda$. The asymptotic normality is stated in the following theorem.

Theorem 1. *In addition to* (1.2), *assume that the conditions* (i)–(v) *in the following Subsection* 5.1 *hold. Then as* $n \to \infty$, *we have*

$$\sqrt{n}(\Lambda_n - \Lambda) \to H, \quad \textit{in distribution}, \tag{2.1}$$

where $\lambda^{\mathrm{T}}\mathrm{Vech}(H)$ *is distributed as* $N(0, \sigma^2_\lambda)$ *for any* $\lambda \neq 0$.

2.1.2 Heteroscedastic model case

In many situations, the inverse error $e = X - E(X|Y)$ is not homoscedastic, that is, $e_i = X_i - E(X|Y = y_i)$ are uncorrelated with mean 0 and $\mathrm{Var}(e_i) = w(y_i)\sigma^2$, where $w(.)$ is a positive continuous weight function on [0,1]. In such settings, it is more appropriate to consider a weighted sum of squares criterion, such as

$$\sum_{i=1}^{n} w_j^{-1}(y_i)(x_{ji} - s_j(y_i))^2.$$

Recall the definition of H in Theorem 1. Similarly, we can obtain the following theorem.

Theorem 2. *In addition to* (1.2), *assume that the conditions* (i)–(v) *in the following Subsection* 5.1 *hold. Then as* $n \to \infty$, *we have*

$$\sqrt{n}(\Lambda_{wn} - \Lambda) \to H, \quad \textit{in distribution}, \tag{2.2}$$

where $\lambda^{\mathrm{T}}\mathrm{Vech}(H)$ *is distributed as* $N(0, \sigma^2_\lambda)$ *for any* $\lambda \neq 0$.

Compared with the result in Theorem 1, an interesting finding is that both converge to the same form of the random matrix H.

2.2 Asymptotic properties of eigenvalues and of eigenvectors

From Theorem 1 (Theorem 2), we can derive the asymptotic normality of the eigenvalues and of the corresponding eigenvectors by using the standard perturbation theory. The following result is parallel to that of the SIR presented by Zhu and Fang[6]. We omit the detail of the proof in this article.

Let $\lambda_1(A) \geqslant \lambda_2(A) \geqslant \cdots \geqslant \lambda_p(A) \geqslant 0$ and $b_i(A) = (b_{1i}(A), \ldots, b_{pi}(A))^{\mathrm{T}}, i = 1, \ldots, p$, denote, respectively, the eigenvalues and their corresponding eigenvectors of a $p \times p$ matrix A.

Theorem 3. *In addition to the conditions of Theorem* 1 (*Theorem* 2), *assume that the nonzero* $\lambda_l(\Lambda)$*'s are distinct. Then for each nonzero eigenvalue* $\lambda_i(\Lambda)$ *and the corresponding eigenvector* $b_i(\Lambda)$, *we have*

$$\sqrt{n}(\lambda_i(\Lambda_n) - \lambda_i(\Lambda)) = \sqrt{n} b_i(\Lambda)^{\mathrm{T}}(\Lambda_n - \Lambda)b_i(\Lambda) + o_p(\sqrt{n}\|\Lambda_n - \Lambda\|) = b_i(\Lambda)^{\mathrm{T}} H b_i(\Lambda) + O_p(1)$$

and

$$\begin{aligned}\sqrt{n}(b_i(\Lambda_n) - b_i(\Lambda)) &= \sqrt{n}\sum_{l=1, l\neq i}^{p} \frac{b_i(\Lambda)b_i(\Lambda)^{\mathrm{T}}(\Lambda_n - \Lambda)b_i(\Lambda)}{\lambda_j(\Lambda) - \lambda_l(\Lambda)} + o_p(\sqrt{n}\|\Lambda_n - \Lambda\|) \\ &= \sum_{l=1, l\neq i}^{p} \frac{b_i(\Lambda)b_i(\Lambda)^{\mathrm{T}} H b_i(\Lambda)}{\lambda_j(\Lambda) - \lambda_l(\Lambda)} + O_p(1),\end{aligned}$$

where $\|\Lambda_n - \Lambda\| = \sum_{1\leqslant i,j\leqslant p} |a_{ij}|$, *and* H *is given in Theorem* 1 (*or* H_w *is given in Theorem* 2).

It is important to note that the asymptotic normality holds when $\lambda_i(\Lambda) > 0$. Otherwise, $\lambda_i(\Lambda)$ converges to 0 faster than the root-n rate by the direct application of Theorem 3.1 in [24, p. 264]). The same phenomenon holds for Li[1], Zhu and Ng[5] and Fung, He and Liu, et al.[7]

3 Estimating the structural dimensionality

The determination of the dimension of $S_{y|\boldsymbol{x}}$ is another important issue in this area. In this section, we aim at determining the structural dimensionality. To overcome the shortcomings of the sequential tests, Zhu, Miao and Peng[18] and Zhu and Zhu[19] suggested a new procedure of BIC (see [17]) type for determining the dimension. In this paper, we suggest another modified version of BIC.

Note that the determination of the dimension of $S_{Y|\boldsymbol{x}}$ is equivalent to the estimation of K, the number of the eigenvalues of Λ being greater than 0. Recall the definition of $\lambda_i(A)$. Define

$$G(k) = n\frac{\sum_{i=1}^{k}\lambda_i(\Lambda_n)^2}{\sum_{i=1}^{p}\lambda_i(\Lambda_n)^2} - 2\ln(n)\frac{k}{p}. \tag{3.1}$$

The second term of $G(k)$ is a penalty function and k equals the number of nonzeros eigenvalues $\lambda_i(\Lambda)$ needed to estimate. Similar to Schwarz[17], we choose $\ln(n)$ in the penalty. Then the estimator of K is defined as the maximizer $\hat{K}$ of $G(k)$ over $k \in \{1,\ldots,p\}$, that is,

$$G(\hat{K}) = \max_{1\leqslant k\leqslant p} G(k). \tag{3.2}$$

Theorem 4. *Under the conditions of Theorem* 1, $\hat{K}$ *converges to* K *in probability.*

4 Illustrative examples

4.1 Simulations

In this subsection, we conduct a small simulation study to evidence the finite-sample performance of our proposed BIC and to compare it with the existing methods. Three models are selected here with $p = 5$ and $n = 100$ or 500. These models were also considered in [1] to show that the efficiency of the slicing estimator of SIR is insensitive to the number of slices. To study its performance in the highly skewed or heavy tailed case, Fung, He and Liu et al.[7] also varied the predictor distributions. We will adopt the models used in their paper.

The one-dimension model considered here is

$$y = x_1 + x_2 + e, \tag{4.1}$$

and the two-dimensional models take one of the forms

$$y = x_1(1 + x_1 + x_2) + e, \tag{4.2}$$

$$y = x_1/(1 + x_1 + x_2) + 0.5e, \tag{4.3}$$

where $x_i, i = 1,\ldots,5$ is distributed as F_i but e comes from some distribution G. A total of seven cases are reported in Table 1.

Columns 2–6 of Table 1 specify the distribution of F_i and G, where Z stands for the standard normal, B for Bernoulli, C for $\chi_1^2 - 1$, L for lognormal, and t_v for the student's distribution

Table 1. Specification of Cases 1–7

Case	F_1	F_2	F_3	F_4	F_5	G	Model	K	(1.2)
Case 1	Z	Z	Z	Z	Z	t_5	(4.1)	1	Yes
Case 2	Z	B	B	Z	L	$0.05Z$	(4.2)	2	Yes
Case 3	C	C	C	C	C	Z	(4.1)	1	No
Case 4	C	C	C	C	C	C	(4.2)	2	Yes
Case 5	t_3	t_3	t_3	t_3	t_3	t_3	(4.1)	1	Yes
Case 6	Z	Z	Z	Z	Z	Z	(4.3)	2	Yes
Case 7	Z	Z	Z	Z	$*$	Z	(4.1)	1	Yes

with v-degrees of freedom. All the x_i and e_i are independent of one another with the exception $*$ in the table for F_5 of Case 7. In this case, x_5 is taken from the distribution $N(x_1+x_2, 10^{-6})$. This indicates that a strong collinearity among predictors exists. Column 8 specifies the form of the model used in the corresponding case. Column 9 gives the structural dimensionality of each case, and Column 10 indicates whether the linear condition (1.2) holds or not. The central space exists in all the above cases, although the linearity condition fails in Case 3, and the linear dependence among predictors is present in Case 7. Fung, He and Liu et al.[7] made this simulation by drawing 300 samples of size $n = 100$ and 300 samples of size $n = 500$. Their simulation results are cited here to compare BIC with three sequential tests called CHSQ, ASNM and RANK. One can refer to Fung, He and Liu et al.[7] for details about CHSQ, ASNM and RANK tests. The results are reported in Table 2–Table 8 for each case.

Table 2. Frequencies of selected model dimensions with Case 1

Test	n	K=0	K=1	K=2	K=3	K=4
BIC	100	0	265	32	3	0
	500	0	300	0	0	0
CHSQ	100	0	292	8	0	0
	500	0	284	15	0	0
ASNM	100	0	249	50	1	0
	500	0	232	64	4	0
RANK	100	21	279	0	0	0
	500	0	300	0	0	0

Table 3. Frequencies of selected model dimensions with Case 2

Test	n	K=0	K=1	K=2	K=3	K=4
BIC	100	0	16	261	23	0
	500	0	1	295	4	0
CHSQ	100	0	47	246	7	0
	500	0	0	289	12	0
ASNM	100	78	156	54	10	2
	500	0	55	212	32	1
RANK	100	104	80	93	22	1
	500	0	28	254	18	0

Table 4. Frequencies of selected model dimensions with Case 3

Test	n	K=0	K=1	K=2	K=3	K=4
BIC	100	0	0	291	9	0
	500	0	0	298	2	0
CHSQ	100	0	1	280	18	1
	500	0	0	279	21	0
ASNM	100	36	31	197	35	1
	500	0	0	248	52	0
RANK	100	0	77	187	36	0
	500	0	0	300	0	0

Table 5. Frequencies of selected model dimensions with Case 4

Test	n	K=0	K=1	K=2	K=3	K=4
BIC	100	0	36	222	42	0
	500	0	46	251	3	0
CHSQ	100	0	2	270	27	1
	500	0	0	280	20	0
ASNM	100	53	36	175	36	0
	500	0	0	246	52	2
RANK	100	101	150	49	0	0
	500	0	3	297	10	0

Fung, He and Liu et al.[7] claimed that CHSQ test proposed by Li[1] is a simple and reliable choice except when the variables are highly skewed or heavy tailed, the RANK test is more robust but tends to be conservative for small to modest sample sizes, and the ASNM test is less predictable. Our proposed BIC behaves similar as the CHSQ and RANK tests.

In particular, Table 2 with all normal predictors implies that all tests perform well while Table 3 with all chi-square predictors indicates that all tests tend to point to 2 dimensions although the first SIR direction is close to $(1, 1, 0, 0, 0)$. Table 4 and Table 7 report that BIC work is much better than other criteria in the small sample size. For the large sample size, BIC also works as RANK. Table 5 shows that BIC performs well in the small sample size.

Table 6. Frequencies of selected model dimensions with Case 5

Test	n	K=0	K=1	K=2	K=3	K=4
BIC	100	0	98	200	2	0
	500	0	108	190	2	0
CHSQ	100	0	181	112	7	0
	500	0	106	192	2	0
ASNM	100	58	190	49	3	0
	500	22	254	21	3	0
RANK	100	66	233	1	0	0
	500	9	290	1	0	0

Table 7. Frequencies of selected model dimensions with Case 6

Test	n	K=0	K=1	K=2	K=3	K=4
BIC	100	0	7	230	62	1
	500	0	0	299	1	0
CHSQ	100	0	101	191	8	0
	500	0	0	291	8	1
ASNM	100	1	58	187	54	0
	500	0	0	231	69	0
RANK	100	86	188	26	0	0
	500	0	0	300	0	0

Table 8. Frequencies of selected model dimensions with Case 7

Test	n	K=0	K=1	K=2	K=3	K=4
BIC	100	0	254	46	0	0
	500	0	298	2	0	0
CHSQ	100	0	289	11	0	0
	500	0	286	12	2	0
ASNM	100	0	287	13	0	0
	500	0	282	18	0	0
RANK	100	0	300	0	0	0
	500	0	300	0	0	0

However, in Case 5, CHSQ and BIC often pick some extra dimensions even when the right direction has already been well estimated. There is severe collinearity in Case 7. All tests are hardly affected by collinearity, even though the estimated direction is unable to choose between $(1,1,0,0,0)$ and $(0,0,0,0,1)$.

Conclusively, as a non-sequential test with an easy computation, our BIC is a competitive approach to determining the dimensionality.

4.2 Real-data application: horse mussel data

Theoretically, the spline approximation can handle the heteroscedastic model. However, its efficiency might depend on the choice of the weight function. Usually we need to estimate the variance function in practice. In this subsection, we will illustrate this issue by the horse mussel data.

The sample of 201 horse mussels was collected at 5 sites in the Malborough Sounds at the Northeast of New Zealand's South Island (see [25]). The response variable is muscle mass Y, the edible portion of the mussel, in grams. The quantitative predictors are all the related characteristics of mussel shells: shell length L, shell width W, shell height H, each in mm, and shell mass S in grams. Indicator predictors for site may be relevant when the variation in muscle mass from site to site is of interest. See also the data description in [2].

For simplicity, we only consider the regression problem with the response Y and predictors (L, W, S). Cook [2] suggested that the predictors be transformed to $X^{\mathrm{T}} = (L, W^{0.36}, S^{0.11}) =:$

(X_1, X_2, X_3) to comply with the linearity requirement. They claimed that the data can be possibly heteroscedastic. This can also be viewed via the scatter plots given in Figure 1. It indicates that the variance of X_1 fluctuates a little bit whereas the variances of X_2 and of X_3 tend to be smaller with the increase of Y.

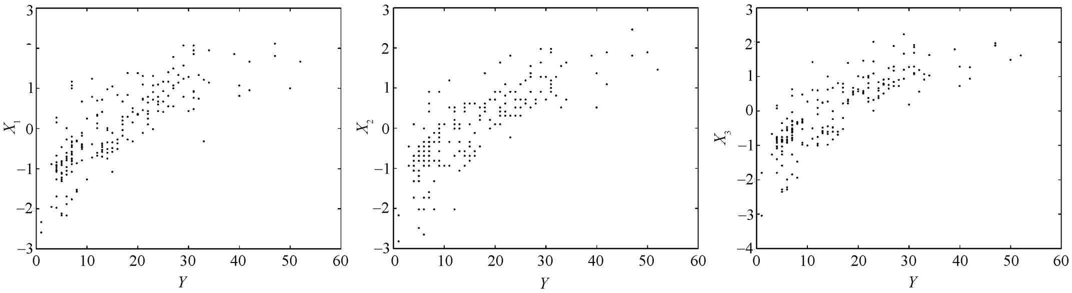

Figure 1. The scatter plots of X versus Y for the horse mussel data. The left plot Y against X_1; the middle plot Y against X_2; the right plot scatters Y against X_3

Specifically, we use the spline estimation to estimate the variance function. Write $E(\cdot|Y = y) = E(\cdot|y)$, then we have $\widehat{\mathrm{Var}}(X|y) = \widehat{E}(X^2|y) - \widehat{E}^2(X|y)$. In our study, the cubic spline function is used. The data-driven method, the well-known generalized cross validation (GCV), is used here to select the optimal bandwidth. This estimation procedure should be consistent. See, for instance, [26] and [27] and the references therein. The spline estimation of the variance function reported in Figure 2 verifies the heteroscedasticity. Hence, the weight functions can be chosen as $w_i(y) = \mathrm{Var}(X_i|y), i = 1, 2, 3$. Figure 3 shows that scatter plots of y_j versus the adjusted predictors $x_{ij}/\sqrt{\widehat{\mathrm{Var}}(x_i|y_j)}$. It should be viewed as homoscedasticity now. The data analysis can proceed based on the adjusted data.

5 Appendix

5.1 Some Assumptions

(i) $f(y) = E(X|Y = y) \in C^m[0, 1]$;

(ii) $E\|X_l\|^4 < \infty$ for all $l = 1, \dots, p$;

(iii) the B-Spline function $N(\cdot)$ satisfies: $\max_{1\leqslant i\leqslant k_0} |h_{i+1} - h_i| = o(k_0^{-1})$ and $h/\min_{1\leqslant i\leqslant k_0} h_i \leqslant M$ where $h_i = t_i - t_{i-1}$, $h = \max_{1\leqslant i\leqslant k_0} h_i$ and $M > 0$ is a predetermined constant;

(iv) as $n \to \infty$, $h \sim n^{-c_1}$ with positive numbers c_1 satisfying $\frac{1}{2m} < c_1 < \frac{1}{2}$, and the notation " $\sim$ " means that the two quantities have the same convergence order;

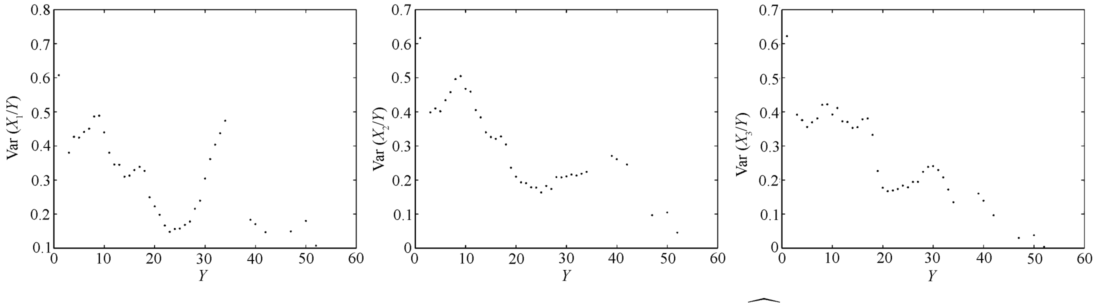

Figure 2. The estimated variance functions. The left plot Y against $\widehat{\mathrm{Var}}(X_1|Y)$; the middle plot Y against $\widehat{\mathrm{Var}}(X_2|Y)$; the right plot Y against $\widehat{\mathrm{Var}}(X_3|Y)$

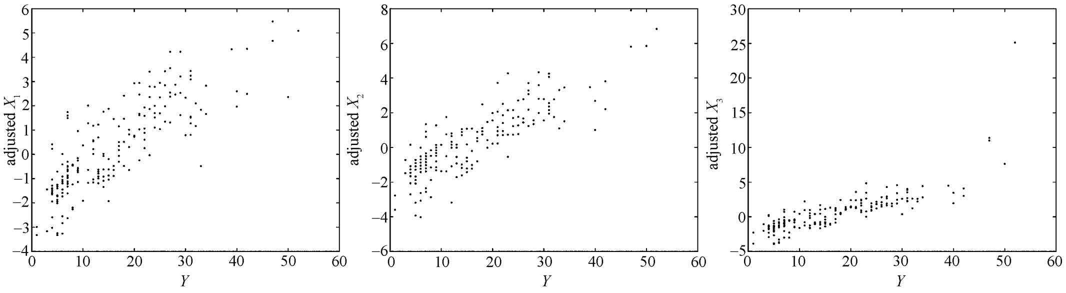

Figure 3. The scatter plots of the adjusted predictors X versus Y. The left plot Y against X_1; the middle plot Y against X_2; the right plot Y against X_3

(v) the marginal density of Y is bounded away from 0 and infinity on [0,1].

Remark 5.1. Condition (i) is concerned with the smoothness of the inverse regression curve $E(X|Y=y)$. Condition (ii) is necessary for the asymptotic normality of Λ_n. Condition (iii) is usually used in the spline approximation. Such an assumption assures that $M^{-1} < k_0 h < M$, which is necessary for numerical computations. These conditions are also commonly used. The number of knots plays a similar role as the number of slices for SIR or the bandwidth of its kernel estimation in [6]. Condition (iv) shows the range of knots for asymptotic normality. Clearly, it is fairly wide, but an undersmoothing is needed because the optimal number of knots $O(n^{\frac{1}{2m+1}})$ is not in this range. This phenomenon is essentially the same as the kernel estimation in [6]. Condition (v) may appear to be stringent requirement on the distribution of Y, but note that in sense of (1.1) we will have an invariant property with a monotone transformation on the Y response so Condition (v) always holds if an appropriate transformation is used.

5.2 Proofs of theorems

Proof of Theorem 1. The final proof of Theorem 1 can be easily derived from Lemma 5.1 and the standard U-statistic theory. Without loss of generality, we assume that $E(X)=0$. By invoking Lemma 5.1, this theorem will hold if we write $\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{f}_k(y_i)f_l(y_i)$ into

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(E(N^{\mathrm{T}}(Y)f_l(Y))G^{-1}(q)N(y_j)x_{kj} + N^{\mathrm{T}}(y_j)G^{-1}(q)E(N(Y)X_k)f_l(y_j)) \\ - E(f_k(Y)f_l(Y)) + O_p(\sqrt{n}h^m). \tag{5.1}$$

Firstly, we write the term $\frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{f}_k(y_i)f_l(y_i)$ as a U-statistic U_n.

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{f}_k(y_i)f_l(y_i) &= \frac{1}{\sqrt{n}n}\sum_{i=1}^{n}\sum_{j=1}^{n}N^{\mathrm{T}}(y_i)G_{J,n}^{-1}N(y_j)x_{kj}f_l(y_i) \\ &= \frac{\sqrt{n}}{C_n^2}\sum_{i<j}\frac{N^{\mathrm{T}}(y_i)G_{J,n}^{-1}N(y_j)x_{kj}f_l(y_i) + N^{\mathrm{T}}(y_j)G_{J,n}^{-1}N(y_i)x_{ki}f_l(y_j)}{2} + o_p(1) \\ &= \frac{\sqrt{n}}{C_n^2}\sum_{i<j}\frac{N^{\mathrm{T}}(y_i)G^{-1}(q)N(y_j)x_{kj}f_l(y_i) + N^{\mathrm{T}}(y_j)G^{-1}(q)N(y_i)x_{ki}f_l(y_j)}{2} + o_p(1) \\ &=: \frac{\sqrt{n}}{C_n^2}\sum_{i<j}u(x_i,y_i,x_j,y_j) + o_p(1) =: \sqrt{n}U_n + o_p(1). \end{aligned} \tag{5.2}$$

The first equality holds because the sum of all terms with $i=j$ is $o_p(1)$. Write $\Delta = G_{J,n} - G(q)$.

Lemma 6.4 in [22] proved that $\|\Delta\| = o(h)$. Hence, the second equality holds by invoking the relationship $G_{J,n}^{-1} = G^{-1}(q) - G^{-1}(q)\Delta(I + G^{-1}(q)\Delta)^{-1}G^{-1}(q)$.

To prove this theorem, we need to show that U_n can be approximated by its projection, $\hat{U}_n = \sum_{j=1}^{n} E(U_n|x_{kj}, y_j) - (n-1)Eu(X_{k1}, Y_1, X_{k2}, Y_2)$, where $u(\cdot)$ is the kernel of the U-statistic U_n. In what follows we will verify that U_n can be approximated by its projection $\hat{U}_n$ at a rate $\frac{1}{\sqrt{n}}h$, that is

$$\sqrt{n}(\hat{U}_n - U_n) = O_p\left(\frac{1}{\sqrt{n}}h\right). \tag{5.3}$$

Similar to the proof of Lemma 5.3 of Zhu and Zhu [19], we only need to show $E(u(X_{k1}, Y_1, X_{k2}, Y_2))^2 = O(1/h^2)$ where $u(\cdot)$ is defined in (5.2). Clearly,

$$\begin{aligned}&E(u(X_{k1}, Y_1, X_{k2}, Y_2))^2\\ &\quad\leqslant 2E((N^{\mathrm{T}}(Y_1)G^{-1}(q)N(Y_2)X_{k2}f(Y_1))^2 + (N^{\mathrm{T}}(Y_2)G^{-1}(q)N(Y_1)X_{k1}f(Y_2))^2).\end{aligned}$$

These two terms are symmetric, so we only need to deal with the first term. Noting that $N^{\mathrm{T}}(Y)G^{-1}(q)N(Y) = O_P(J)$, we have

$$\begin{aligned}&E(N^{\mathrm{T}}(Y_1)G^{-1}(q)N(Y_2)X_{k2}f_l(Y_1))^2\\ &\quad= E((f_l(Y_1)N^{\mathrm{T}}(Y_1)G^{-1}(q)N(Y_2))^2E(X_{k2}^2|Y_2))\\ &\quad= E(N^{\mathrm{T}}(Y_1)f_l(Y_1)G^{-1}(q)N(Y_2)E(X_{k2}^2|Y_2)N^{\mathrm{T}}(Y_2)G^{-1}(q)f_l(Y_1)N(Y_1))\\ &\quad= \mathrm{trace}(E(E(X_{k2}^2|Y_2)N(Y_2)N^{\mathrm{T}}(Y_2)G^{-1}(q))E(f_l^2(Y_1)N(Y_1)N^{\mathrm{T}}(Y_1)G^{-1}(q)))\\ &\quad\leqslant \mathrm{trace}(E(E(X_{k2}^2|Y_2)N(Y_2)N^{\mathrm{T}}(Y_2)G^{-1}(q))E(E(X_{l1}^2|Y_1)N(Y_1)N^{\mathrm{T}}(Y_1)G^{-1}(q)))\\ &\quad\leqslant \mathrm{trace}(E\|X_2^4\|E(N(Y_2)N^{\mathrm{T}}(Y_2)G^{-1}(q))^2) = O(J^2) = O(1/h^2),\end{aligned}$$

where the operator trace(A) is the sum of the diagonal elements of matrix A. Therefore, (5.3) holds. It means that $\sqrt{n}U_n$ and $\sqrt{n}\hat{U}_n$ are asymptotically equivalent. We are now in the position to express the projection $\hat{U}_n$ into sum of i.i.d random variables. The procedure is the same as in the proof of Lemma 5.3 in [19]. We have

$$E(N^{\mathrm{T}}(y_i)G^{-1}(q)N(y_j)x_{kj}f_l(y_i)|x_{kj}, y_j) = E(N^{\mathrm{T}}(Y)f_l(Y))G^{-1}(q)N(y_j)x_{kj} = f_l(y_j)x_{kj},$$

and $E(N^{\mathrm{T}}(y_i)G^{-1}(q)N(y_j)x_{kj}f_l(y_i)|x_{ki}, y_i) = f_l(y_i)x_{ki}$. Moreover, by invoking (2.7) of Barrow and Smith[28], we can easily obtain the expectation of the above two equations. Both equal $E(f_k(Y)f_l(Y)) + O(h^m)$.

After some basic calculations, the centered projection $\hat{U}_n - E(\hat{U}_n)$ of U_n can be written into a sum of i.i.d random variables, (5.1) holds.

Proof of Theorem 2. The proof is almost identical to that of Theorem 1, the details are omitted.

Proof of Theorem 3. This result is parallel to that of the SIR presented by Zhu and Fang[6], hence we skip the details.

Proof of Theorem 4. Let K be the true value of the dimension of B. As stated in Theorem 3, we have $\lambda_k(\Lambda_n) - \lambda_k(\Lambda) = O_p(1/\sqrt{n})$. Therefore, if $K > k$,

$$G(K) - G(k) = \left(n\frac{\sum_{i=1}^{K}\lambda_i{}^2(\Lambda_n)}{\sum_{i=1}^{p}\lambda_i{}^2(\Lambda_n)} - 2\ln(n)\frac{K}{p}\right) - \left(n\frac{\sum_{i=1}^{k}\lambda_i{}^2(\Lambda_n)}{\sum_{i=1}^{p}\lambda_i{}^2(\Lambda_n)} - 2\ln(n)\frac{k}{p}\right)$$

$$
\begin{aligned}
&= n\frac{\sum_{i=k+1}^{K}\lambda_i{}^2(\Lambda_n)}{\sum_{i=1}^{p}\lambda_i^2(\Lambda_n)} + 2\ln(n)\frac{k-K}{p} \\
&\to n\frac{\sum_{i=k+1}^{K}\lambda_i{}^2(\Lambda)}{\sum_{i=1}^{K}\lambda_i^2(\Lambda)} + 2\ln(n)\frac{k-K}{p} > 0. \qquad (5.4)
\end{aligned}
$$

If $K < k$,

$$
\begin{aligned}
G(K) - G(k) &= \left(n\frac{\sum_{i=1}^{K}\lambda_i{}^2(\Lambda_n)}{\sum_{i=1}^{p}\lambda_i{}^2(\Lambda_n)} - 2\ln(n)\frac{K}{p}\right) - \left(n\frac{\sum_{i=1}^{k}\lambda_i{}^2(\Lambda_n)}{\sum_{i=1}^{p}\lambda_i{}^2(\Lambda_n)} - 2\ln(n)\frac{k}{p}\right) \\
&= -n\frac{\sum_{i=K+1}^{k}\lambda_i{}^2(\Lambda_n)}{\sum_{i=1}^{p}\lambda_i{}^2(\Lambda_n)} + 2\ln(n)\frac{k-K}{p} \\
&\to \ln(n)\frac{k-K}{p} > 0. \qquad (5.5)
\end{aligned}
$$

It follows from (5.4) and (5.5) that $\hat{K} \to K$ in probability.

5.3 A useful lemma

Lemma 5.1. *Under Assumptions* (i), (ii), (iv) *and* (v) *illustrated in Subsection* 5.1*,*

$$
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{f}_l(y_j) - f_l(y_j))^2 = o_p(1). \qquad (5.6)
$$

Proof of Lemma 5.1. To prove this lemma, we only need to show the expectation is $o(1)$. Express the LHS into three terms:

$$
\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(\hat{f}_l(y_j) - f_l(y_j))^2 = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\hat{f}_l^2(y_j) - 2\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\hat{f}_l(y_j)f_l(y_j) + \frac{1}{\sqrt{n}}\sum_{j=1}^{n}f_l^2(y_j). \qquad (5.7)
$$

First, we write the first term in a U-statistic first.

$$
\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\hat{f}_l^2(y_j) &= \frac{1}{\sqrt{n}n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\sum_{k=1}^{n}N(x_j)G_{J,n}^{-1}N(x_{li})y_iN(x_{lj})G_{J,n}^{-1}N(x_{lk})y_k \\
&= \frac{1}{\sqrt{n}n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\sum_{k=1}^{n}N(x_{lj})G^{-1}(q)N(x_{li})y_iN(x_{lj})G^{-1}(q)N(x_{lk})y_k + o_p(1) \\
&= \frac{\sqrt{n}}{C_n^3}\sum_{i<j<k}u_1(x_{li}, y_i, x_{lj}, y_j, x_{lk}, y_k) + o_p(1) =: \sqrt{n}U_1 + o_p(1).
\end{aligned}
$$

The expectation of this U-statistic can be obtained as follows.

$$
\begin{aligned}
E(U_1) &= Eu_1(X_{l1}, Y_1, X_{l2}, Y_2, X_{l3}, Y_3) \\
&= E(N^{\mathrm{T}}(X_{l2})G^{-1}(q)N(X_{l1})Y_1N^{\mathrm{T}}(X_{l2})G^{-1}(q)N(X_{l3})Y_3) \\
&= E(Y_1N^{\mathrm{T}}(X_{l1})G^{-1}(q)N(X_{l2})N^{\mathrm{T}}(X_{l2})G^{-1}(q)N(X_{l3})Y_3) \\
&= E(Y_1N^{\mathrm{T}}(X_1))E(G^{-1}(q)N(X_{l2})N^{\mathrm{T}}(X_{l2}))E(G^{-1}(q)N(X_{l3})Y_3) \\
&= E(f_l^2(X)) + O(h^m).
\end{aligned}
$$

The computation of the expectation of the second term in (5.7) is essentially the same. That is, $E(\hat{f}_l(y_j)f_l(y_j)) = E(f_l^2(Y)) + O(h^m)$. This completes the proof of Lemma 5.1.

Acknowledgements The authors cordially thank two anonymous referees for their valuable comments which lead to the improvement of this paper.

References

[1] Li K. Sliced inverse regression for dimension reduction (with discussion). *J Amer Statist Assoc,* **86**: 316–342 (1991)

[2] Cook R. Regression graphics: Ideas for studying regressions through graphics. New York: Wiley & Sons, 1998

[3] Zhu L, Zhu L. Model checking for the conditional independence in dimension reduction models. Working Paper, 2005b

[4] Hsing T, Carroll R. An asymptotic theory for sliced inverse regression. *Ann Statist,* **20**: 1040–1061 (1992)

[5] Zhu L, Ng K. Asymptotics of sliced inverse regression. *Statist Sinica,* **5**: 727–736 (1995)

[6] Zhu L, Fang K. Asymptotics for kernel estimate of sliced inverse regression. *Ann Statist,* **24**: 1053–1068 (1996)

[7] Fung W, He X, Liu L, Shi P. Dimension reduction based on canonical correlation, *Statist Sinica,* **12**: 1093–1113 (2002)

[8] Kent J. Sliced inverse regression for dimension reduction: comment. *J Amer Statist Assoc,* **86**: 336–337 (1991)

[9] Chen C, Li K. Can SIR be as popular as multiple linear regression? *Statist Sinica,* **8**: 289–316 (1998)

[10] Cook R, Weisberg S. Residuals and Influence in Regression. New York: Chapman and Hall, 1982

[11] Schott J. Determining the dimensionality in sliced inverse regression. *J Amer Statist Assoc*, **89**: 141–148 (1994)

[12] Velilla S. Assessing the number of linear components in a general regression problem. *J Amer Statist Assoc*, **93**: 1088–1098 (1998)

[13] Bura E, Cook R. Estimating the structural dimension of regressions via parametric inverse regression. *J Roy Stat Soc B,* **63**: 393–410 (2001)

[14] Ferré L. Determination of the dimension in SIR and related methods. *J Amer Statist Assoc,* **93**: 132–140 (1998)

[15] Mallows C. Bounds on distribution functions in terms of expectations of order-statistics. *Ann Probability,* **1**: 297–303 (1973)

[16] Akaike H. Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika,* **60**: 255–265 (1973)

[17] Schwarz G. Estimating the dimension of a model. *Ann Math Statist*, **30**: 461–464 (1978)

[18] Zhu L, Miao B, Peng H. Sliced Inverse Regression with Large Dimensional Covariates. *J Amer Statist Assoc,* **101**: 630–643 (2006)

[19] Zhu L, Zhu L. On kernel method for sliced average variance estimation. *J Multi Ana,* **98**: 970–991 (2007)

[20] Agarwal G, Studden W. Asymptotic integrated mean squares error using least squares and bias minimizing splines. *Ann Statist,* **8**: 1307–1325 (1980)

[21] Huang S, Studden W. An equivalent kernel method for least squares spline regression. *Statist Decisions,* **3**: 179–201 (Supp) (1993)

[22] Zhou X, Shen X, Wolfe D. Local asymptotics for regression splines and confidence regions. *Ann Statist,* **26**: 1760–1782 (1998)

[23] Schumaker L. Spline Functions. New York: John Wiley, 1981

[24] Eaton M, Tyler D. On Wielandt's inequality and its applications to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *Ann Statist,* **19**: 260–271 (1991)

[25] Camden M. The Data Bundle. Wellington: New Zealand Statistical Association, 1989

[26] Yao Q, Tong H. Quantifying the influence of initial values on nonlinear prediction. *J Roy Statist Soc Ser B,* **56**: 701–726 (1994)

[27] Härdle W, Tsybakov A. Local polynomial estimators of the volatility function in nonparametric autoregression. *J Economet,* **81**: 233–242 (1997)

[28] Barrow D, Smith P. Asymptotic properties of best $L_2[0,1]$ approximation by spline with variable knots. *Q Appl Math,* **36**: 293–304 (1978)