J. R. Statist. Soc. B (2003)
65, *Part* 1, *pp.* 57–80

Adaptive varying-coefficient linear models

Jianqing Fan,

University of North Carolina, Chapel Hill, USA

Qiwei Yao

London School of Economics and Political Science, UK

and Zongwu Cai

University of North Carolina, Charlotte, USA

[Received June 2000. Final revision July 2002]

Summary. Varying-coefficient linear models arise from multivariate nonparametric regression, non-linear time series modelling and forecasting, functional data analysis, longitudinal data analysis and others. It has been a common practice to assume that the varying coefficients are functions of a given variable, which is often called an *index*. To enlarge the modelling capacity substantially, this paper explores a class of varying-coefficient linear models in which the index is unknown and is estimated as a linear combination of regressors and/or other variables. We search for the index such that the derived varying-coefficient model provides the least squares approximation to the underlying unknown multidimensional regression function. The search is implemented through a newly proposed hybrid backfitting algorithm. The core of the algorithm is the alternating iteration between estimating the index through a one-step scheme and estimating coefficient functions through one-dimensional local linear smoothing. The locally significant variables are selected in terms of a combined use of the t-statistic and the Akaike information criterion. We further extend the algorithm for models with two indices. Simulation shows that the methodology proposed has appreciable flexibility to model complex multivariate non-linear structure and is practically feasible with average modern computers. The methods are further illustrated through the Canadian mink–muskrat data in 1925–1994 and the pound–dollar exchange rates in 1974–1983.

Keywords: Akaike information criterion; Backfitting algorithm; Generalized cross-validation; Local linear regression; Local significant variable selection; One-step estimation; Smoothing index

1. Introduction

Suppose that we are interested in estimating the multivariate regression function $G(\mathbf{x}) \equiv E(Y|\mathbf{X}=\mathbf{x})$, where Y is a random variable and $\mathbf{X}$ is a $d \times 1$ random vector. In this paper, we propose to *approximate* the regression function $G(\mathbf{x})$ by a varying-coefficient model

$$g(\mathbf{x}) = \sum_{j=0}^{d} g_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})x_j, \tag{1.1}$$

where $\boldsymbol{\beta} \in \Re^d$ is an unknown direction, $\mathbf{x} = (x_1, \ldots, x_d)^{\mathrm{T}}$, $x_0 = 1$ and coefficients $g_0(\cdot), \ldots, g_d(\cdot)$

Address for correspondence: Qiwei Yao, Department of Statistics, London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, UK.
E-mail: q.yao@lse.ac.uk

1369–7412/03/65057

are unknown functions. We choose the direction $\boldsymbol{\beta}$ and coefficient functions $g_j(\cdot)$ such that $E\{G(\mathbf{X}) - g(\mathbf{X})\}^2$ is minimized. The appeal of this model is that, once $\boldsymbol{\beta}$ has been given, we can directly estimate $g_j(\cdot)$ by the standard one-dimensional kernel regression localized around $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$. Furthermore, the coefficient functions $g_j(\cdot)$ can be easily displayed graphically, which may be particularly helpful to visualize how the surface $g(\cdot)$ changes. Model (1.1) appears linear in each co-ordinate of $\mathbf{x}$ when the index $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ is fixed. It may include quadratic and cross-product terms of x_j (or more generally any given functions of x_j) as 'new' components of $\mathbf{x}$. Hence it has considerable flexibility to cater for complex multivariate non-linear structure.

We develop an efficient backfitting algorithm to estimate $g(\cdot)$. The virtue of the algorithm is the alternating iteration between estimating $\boldsymbol{\beta}$ through a one-step estimation scheme (Bickel, 1975) and estimating functions $g_j(\cdot)$ through one-dimensional local linear smoothing. Since we apply smoothing on a scalar $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ only, the method suffers little from the so-called 'curse of dimensionality' which is the innate difficulty associated with multivariate nonparametric fitting. The generalized cross-validation (GCV) method for bandwidth selection is incorporated in the algorithm in an efficient manner. To avoid overfitting, we delete local insignificant variables in terms of a combined use of the t-statistic and the Akaike information criterion (AIC), which is adopted for its computational efficiency. The deletion of insignificant variables is particularly important when we include, for example, quadratic functions of x_j as new components in the model, which could lead to overparameterization. The method proposed has been further extended to estimate varying-coefficient models with two indices, one of which is known.

Varying-coefficient models arise from various statistical contexts in slightly different forms. The vast amount of literature includes, among others, Cleveland *et al.* (1992), Hastie and Tibshirani (1993), Carroll *et al.* (1998), Kauermann and Tutz (1999), Xia and Li (1999a), Zhang and Lee (1999, 2000) and Fan and W. Zhang (1999, 2000) on local multidimensional regression, Ramsay and Silverman (1997) on functional data analysis, Hoover *et al.* (1998), Wu *et al.* (1998) and Fan and J. Zhang (2000) on longitudinal data analysis, Nicholls and Quinn (1982), Chen and Tsay (1993) and Cai, Fan and Yao (2000) on non-linear time series and Cai, Fan and Li (2000) on generalized linear models with varying coefficients. The form of model (1.1) is not new. It was proposed in Ichimura (1993). Recently, Xia and Li (1999b) extended the idea and the results of Härdle *et al.* (1993) from the single-index model to the adaptive varying-coefficient model (1.1). They proposed to estimate the coefficient functions with a given bandwidth and a direction $\boldsymbol{\beta}$, and then to choose the bandwidth and the direction by cross-validation. Some theoretical results were derived under the assumption that the bandwidth was of the order $O(n^{-1/5})$ and the direction $\boldsymbol{\beta}$ was within an $O_p(n^{-1/2})$ consistent neighbourhood of the true value. However, the approach suffers from heavy computational expense. This to some extent explains why most previous work assumed a known direction $\boldsymbol{\beta}$. The new approach in this paper differs from those in three key aspects:

(a) only a one-dimensional smoother is used in estimation,
(b) the index coefficient $\boldsymbol{\beta}$ is estimated by data and
(c) within a local region around $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ we select significant variables x_j's to avoid overfitting.

Aspect (b) is different from Härdle *et al.* (1993) and Xia and Li (1999b) since we estimate the coefficient functions and the direction simultaneously; no cross-validation is needed. This idea is similar *in spirit* to that of Carroll *et al.* (1997) who showed that a semiparametric efficient estimator of the direction $\boldsymbol{\beta}$ can be obtained. Further we provide a theorem (i.e. theorem 1, part (b), in Section 2) on the model identification problem of the form (1.1) which has not been addressed before.

The rest of the paper is organized as follows. Section 2 deals with the adaptive varying-coefficient model (1.1). The extension to the case with two indices is outlined in Section 3. The numerical results of two simulated examples are reported in Section 4.1, which demonstrate that the methodology proposed can capture complex non-linear structure with moderate sample sizes, and further the required computation typically takes less than a minute on a Pentium II 350 MHz personal computer. The methodology is further illustrated in Section 4.2 through the Canadian mink–muskrat data in 1925–1944 and the pound–dollar exchange rates in 1974–1983. The technical proofs are relegated to Appendix A.

2. Adaptive varying-coefficient linear models

2.1. Approximation and identifiability

Since $G(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ is a conditional expectation, it holds that

$$E\{Y - g(\mathbf{X})\}^2 = E\{Y - G(\mathbf{X})\}^2 + E\{G(\mathbf{X}) - g(\mathbf{X})\}^2$$

for any $g(\cdot)$. Therefore, the search for the least squares approximation $g(\cdot)$ of $G(\cdot)$, as defined in model (1.1), is equivalent to the search for such a $g(\cdot)$ that $E\{Y - g(\mathbf{X})\}^2$ obtains the minimum. Theorem 1, part (a), below indicates that there is always such a $g(\cdot)$ under mild conditions. Obviously, if $G(\mathbf{x})$ is in the form of the right-hand side of model (1.1), $g(\mathbf{x}) \equiv G(\mathbf{x})$. The second part of the theorem points out that the coefficient vector $\boldsymbol{\beta}$ is unique up to a constant unless $g(\cdot)$ is in a class of special quadratic functions (see equation (2.2) below). In fact, model (1.1) is an overparameterized form in the sense that one of the functions $g_j(\cdot)$ can be represented in terms of the others. Theorem 1, part (b), confirms that, once the direction $\boldsymbol{\beta}$ has been specified, the function $g(\cdot)$ has a representation with at most d (instead of $d+1$) $g_j(\cdot)$. Furthermore, those $g_j(\cdot)$ functions are identifiable.

Theorem 1.

(a) Assume that the distribution function of $(\mathbf{X}, Y)$ is continuous, and $E\{Y^2 + ||\mathbf{X}||^2\} < \infty$. Then, there is a $g(\cdot)$ defined by model (1.1) for which

$$E\{Y - g(\mathbf{X})\}^2 = \inf_{\boldsymbol{\alpha}} \inf_{f_0, \ldots, f_d} \left[E\left\{ Y - \sum_{j=0}^{d} f_j(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}) X_j \right\}^2 \right], \tag{2.1}$$

where the first infimum is taken over all unit vectors in $\Re^d$ and the second over all measurable functions $f_0(\cdot), \ldots, f_d(\cdot)$.

(b) For any given twice-differentiable $g(\cdot)$ of the form (1.1), if we choose $||\boldsymbol{\beta}|| = 1$, and the first non-zero component of $\boldsymbol{\beta}$ positive, such a $\boldsymbol{\beta}$ is unique unless $g(\cdot)$ is of the form

$$g(\mathbf{x}) = \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x} + c, \tag{2.2}$$

where $\boldsymbol{\alpha}, \boldsymbol{\gamma} \in \Re^d$, $c \in \Re$ are constants and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are not parallel to each other. Furthermore, once $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^{\mathrm{T}}$ has been given and $\beta_d \neq 0$, we may let $g_d(\cdot) \equiv 0$. Consequently, all the other $g_j(\cdot)$ are uniquely determined.

Remark 1. If the conditional expectation $G(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ cannot be expressed in the form of the right-hand side of model (1.1), there may be more than one $g(\mathbf{x})$ of the form (1.1), for which equation (2.1) holds. For example, let $Y = X_1^2 + X_2^2$, where both X_1 and X_2 are independent random variables uniformly distributed on [0, 1]. Then $G(x_1, x_2) = x_1^2 + x_2^2$, which is not in the form (1.1). However, equation (2.1) holds for both $g(x_1, x_2) = 1.25x_1^2$ and $g(x_1, x_2) = 1.25x_2^2$.

Without loss of generality, we always assume from now on that, in model (1.1), $||\boldsymbol{\beta}|| = 1$ and the first non-zero component of $\boldsymbol{\beta}$ is positive. To avoid the complication caused by the lack of uniqueness of the index direction $\boldsymbol{\beta}$, we always assume that $G(\cdot)$ admits a unique least squares approximation of $g(\cdot)$ which cannot be expressed in the form (2.2).

2.2. Estimation

Let $\{(\mathbf{X}_t, Y_t);\ 1 \leqslant t \leqslant n\}$ be observations from a strictly stationary process with the same marginal distribution as $(\mathbf{X}, Y)$. To estimate the surface $g(\cdot)$ defined by equations (1.1) and (2.1), we need to search for the minimizers of $\{f_j(\cdot)\}$ for any given direction $\boldsymbol{\alpha}$ and then to find the direction at which the mean-squared error is minimized. An exhaustive search is intractable. We develop a backfitting algorithm for this optimization problem.

Let $\beta_d \neq 0$. It follows from theorem 1, part (b), that we only search for an approximation in the form

$$g(\mathbf{x}) = \sum_{j=0}^{d-1} g_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})x_j. \tag{2.3}$$

Our task can be formally split into two parts—estimation of functions $g_j(\cdot)$ with a given $\boldsymbol{\beta}$ and estimation of the index coefficient $\boldsymbol{\beta}$ with given $g_j(\cdot)$. We also discuss how to choose the smoothing parameter h in terms of GCV (Wahba, 1977), and how to apply backward deletion to choose locally significant variables in terms of a combined use of the t-statistic and AIC. The algorithm for practical implementation will be summarized at the end of this section.

The computer-intensive nature of the problem prevents us from exploring more sophisticated methods which may lead to an improvement in performance at the cost of computing time. For example, various plug-in methods (chapter 4 of Fan and Gijbels (1996)), thorough GCV (see step 3 in Section 2.3 below) or the corrected AIC (Hurvich *et al.*, 1998; Simonoff and Tsai, 1999) would lead to better bandwidth selectors. The local variable selection could also be based solely on the AIC or the corrected AIC.

2.2.1. *Local linear estimators for the $g_j(\cdot)$ with given $\boldsymbol{\beta}$*

For given $\boldsymbol{\beta}$ with $\beta_d \neq 0$, we need to estimate

$$g(\mathbf{X}) = \arg \min_{f \in \mathcal{F}(\boldsymbol{\beta})} (E[\{Y - f(\mathbf{X})\}^2 \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}]), \tag{2.4}$$

where

$$\mathcal{F}(\boldsymbol{\beta}) = \left\{ f(\mathbf{x}) = \sum_{j=0}^{d-1} f_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})x_j \,\middle|\, f_0(\cdot), \ldots, f_{d-1}(\cdot) \text{ measurable, and } E\{f(\mathbf{X})\}^2 < \infty \right\}.$$

The least squares property of equation (2.4) leads to the estimators $\hat{g}_j(z) = \hat{b}_j$, $j = 0, \ldots, d-1$, where $(\hat{b}_0, \ldots, \hat{b}_{d-1})$ is the minimizer of the sum of weighted squares

$$\sum_{t=1}^{n} \left(Y_t - \sum_{j=0}^{d-1} b_j X_{tj} \right)^2 K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t - z)\, w(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t),$$

where $w(\cdot)$ is a bounded weight function with a bounded support, which is introduced to control the boundary effect. Note that only one-dimensional kernel smoothing is used here.

The above estimation procedure is based on the local constant approximation $g_j(y) \approx g_j(z)$ for y in a neighbourhood of z. Since local constant regression has several drawbacks compared with local linear regression (Fan and Gijbels, 1996), we consider the local linear estimators for the functions $g_0(\cdot), \ldots, g_{d-1}(\cdot)$. This leads to minimizing the sum

$$\sum_{t=1}^{n} \left[Y_t - \sum_{j=0}^{d-1} \{b_j + c_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t - z)\} X_{tj} \right]^2 K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t - z)\, w(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t) \tag{2.5}$$

with respect to $\{b_j\}$ and $\{c_j\}$. Define $\hat{g}_j(z) = \hat{b}_j$ and $\hat{\dot{g}}_j(z) = \hat{c}_j$ for $j = 0, \ldots, d-1$ and set

$$\hat{\boldsymbol{\theta}} \equiv (\hat{b}_0, \ldots, \hat{b}_{d-1},\ \hat{c}_0, \ldots, \hat{c}_{d-1})^{\mathrm{T}}.$$

It follows from least squares theory that

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \boldsymbol{\Sigma}(z)\, \mathcal{X}^{\mathrm{T}}(z)\, \mathcal{W}(z) \mathcal{Y}, \\ \boldsymbol{\Sigma}(z) &= \{\mathcal{X}^{\mathrm{T}}(z)\, \mathcal{W}(z)\, \mathcal{X}(z)\}^{-1}, \end{aligned} \tag{2.6}$$

where $\mathcal{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\mathcal{W}(z)$ is an $n \times n$ diagonal matrix with $K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i - z)\, w(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)$ as its ith diagonal element, $\mathcal{X}(z)$ is an $n \times 2d$ matrix with $(\mathbf{U}_i^{\mathrm{T}}, (\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i - z)\mathbf{U}_i^{\mathrm{T}})$ as its ith row and $\mathbf{U}_t = (1, X_{t1}, \ldots, X_{t,d-1})^{\mathrm{T}}$.

2.2.2. *Search for $\boldsymbol{\beta}$-direction with the $g_j(\cdot)$ fixed*

The minimization property of equation (2.1) suggests that we should search for $\boldsymbol{\beta}$ to minimize

$$R(\boldsymbol{\beta}) = \frac{1}{n} \sum_{t=1}^{n} \left\{ Y_t - \sum_{j=0}^{d-1} g_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t) X_{tj} \right\}^2 w(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t). \tag{2.7}$$

We employ a one-step estimation scheme (see, for example, Bickel (1975)) to estimate $\boldsymbol{\beta}$, which is in the spirit of one-step Newton–Raphson estimation. We expect that the estimator derived will be good if the initial value is reasonably good (see Fan and Chen (1999)).

Suppose that $\hat{\boldsymbol{\beta}}$ is the minimizer of equation (2.7). Then $\dot{R}(\hat{\boldsymbol{\beta}}) = 0$, where $\dot{R}(\cdot)$ denotes the derivative of $R(\cdot)$. For any $\boldsymbol{\beta}^{(0)}$ close to $\hat{\boldsymbol{\beta}}$, we have the approximation

$$0 = \dot{R}(\hat{\boldsymbol{\beta}}) \approx \dot{R}(\boldsymbol{\beta}^{(0)}) + \ddot{R}(\boldsymbol{\beta}^{(0)})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)}),$$

where $\ddot{R}(\cdot)$ is the Hessian matrix of $R(\cdot)$. This leads to the one-step iterative estimator

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} - \ddot{R}(\boldsymbol{\beta}^{(0)})^{-1}\, \dot{R}(\boldsymbol{\beta}^{(0)}), \tag{2.8}$$

where $\boldsymbol{\beta}^{(0)}$ is the initial value. We rescale $\boldsymbol{\beta}^{(1)}$ such that it has unit norm with first non-vanishing element positive. It is easy to see from equation (2.7) that

$$\begin{aligned} \dot{R}(\boldsymbol{\beta}) &= -\frac{2}{n} \sum_{t=1}^{n} \left\{ Y_t - \sum_{j=0}^{d-1} g_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t) X_{tj} \right\} \left\{ \sum_{j=0}^{d-1} \dot{g}_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t) X_{tj} \right\} \mathbf{X}_t\, w(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t), \\ \ddot{R}(\boldsymbol{\beta}) &= \frac{2}{n} \sum_{t=1}^{n} \left\{ \sum_{j=0}^{d-1} \dot{g}_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t) X_{tj} \right\}^2 \mathbf{X}_t \mathbf{X}_t^{\mathrm{T}}\, w(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t) \\ &\quad - \frac{2}{n} \sum_{t=1}^{n} \left\{ Y_t - \sum_{j=0}^{d-1} g_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t) X_{tj} \right\} \left\{ \sum_{j=0}^{d-1} \ddot{g}_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t) X_{tj} \right\} \mathbf{X}_t \mathbf{X}_t^{\mathrm{T}}\, w(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t). \end{aligned} \tag{2.9}$$

In this derivation, the derivative of the weight function $w(\cdot)$ is assumed to be 0 for simplicity. In practice, we usually let $w(\cdot)$ be an indicator function.

In case the matrix $\ddot{R}(\cdot)$ is singular or nearly so, we adopt a ridge regression (Seifert and Gasser, 1996) approach as follows: use estimator (2.8) with $\ddot{R}$ replaced by $\ddot{R}_r$ which is defined by the right-hand side of equation (2.9) with $\mathbf{X}_t\mathbf{X}_t^{\mathrm{T}}$ replaced by $\mathbf{X}_t\mathbf{X}_t^{\mathrm{T}} + q_n\mathbf{I}_d$ for some positive ridge parameter q_n.

Now we briefly state two alternative methods for estimating $\boldsymbol{\beta}$, although they may not be as efficient as the above method. The first is based on a random search method, which is more direct and tractable when d is small. The basic idea is to keep drawing $\boldsymbol{\beta}$ randomly from the d-dimensional unit sphere and then to compute $R(\boldsymbol{\beta})$. Stop the algorithm if the minimum fails to decrease significantly in, say, every 100 new draws. The second approach is to adapt the average derivative method of Newey and Stoker (1993) and Samarov (1993). Under model (1.1), the direction $\boldsymbol{\beta}$ is parallel to the expected difference between the gradient vector of the regression surface and $(g_1(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}), \ldots, g_{d-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}), 0)^{\mathrm{T}}$ and hence can be estimated by the average derivative method via iteration.

2.2.3. Bandwidth selection

We apply the GCV method, proposed by Wahba (1977) and Craven and Wahba (1979), to choose the bandwidth h in the estimation of $\{g_j(\cdot)\}$. The criterion can be described as follows. For given $\boldsymbol{\beta}$, let

$$\hat{Y}_t = \sum_{j=0}^{d-1} \hat{g}_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t) X_{tj}.$$

It is easy to see that all those predicted values are linear combinations of $\mathcal{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$ with coefficients depending on $\{\mathbf{X}_t\}$ only. Namely, we may write

$$(\hat{Y}_1, \ldots, \hat{Y}_n)^{\mathrm{T}} = \mathbf{H}(h)\mathcal{Y},$$

where $\mathbf{H}(h)$ is the $n \times n$ hat matrix, independent of $\mathcal{Y}$. The GCV method selects h minimizing

$$\mathrm{GCV}(h) \equiv \frac{1}{n[1 - n^{-1}\,\mathrm{tr}\{\mathbf{H}(h)\}]^2} \sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2 \, w(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t),$$

which in fact is an estimate of the weighted mean integrated square errors. Under some regularity conditions, it holds that

$$\mathrm{GCV}(h) = a_0 + a_1 h^4 + \frac{a_2}{nh} + o_p(h^4 + n^{-1}h^{-1}).$$

Thus, up to first-order asymptotics, the optimal bandwidth is $h_{\mathrm{opt}} = (a_2/4na_1)^{1/5}$. The coefficients a_0, a_1 and a_2 will be estimated from $\mathrm{GCV}(h_k)$ via least squares regression. This bandwidth selection rule, inspired by the empirical bias method of Ruppert (1997), will be applied outside the loops between estimating β and $\{g_j(\cdot)\}$. See Section 2.2.5.

To calculate $\mathrm{tr}\{\mathbf{H}(h)\}$, we note that, for $1 \leqslant i \leqslant n$,

$$\hat{Y}_i = \frac{1}{n} \sum_{t=1}^{n} Y_t K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \, w(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t)(\mathbf{U}_t^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})\boldsymbol{\Sigma}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\begin{pmatrix} \mathbf{U}_t \\ \mathbf{U}_t\boldsymbol{\beta}^{\mathrm{T}}(\mathbf{X}_t - \mathbf{X}_i)/h \end{pmatrix},$$

where $\mathbf{0}$ denotes the $d \times 1$ vector with all components 0 and $\boldsymbol{\Sigma}(\cdot)$ is defined as in expression (2.6). The coefficient of Y_i on the right-hand side of the above expression is

$$\gamma_i \equiv \frac{1}{n} K_h(0) \, w(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)(\mathbf{U}_i^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})\boldsymbol{\Sigma}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\begin{pmatrix} \mathbf{U}_i \\ \mathbf{0} \end{pmatrix}.$$

Now, we have that $\mathrm{tr}\{\mathbf{H}(h)\} = \Sigma_{i=1}^{n}\gamma_i$.

2.2.4. Choosing locally significant variables

As we discussed before, model (2.3) can be overparameterized. Thus, it is necessary to select significant variables for each given z after an initial fitting. In our implementation, we use a backward stepwise deletion technique which relies on a modified AIC and t-statistics. More precisely, we delete the least significant variable in a given model according to t-values, which yields a new and reduced model. We select the best model according to the AIC.

We start with the full model

$$g(\mathbf{x}) = \sum_{j=0}^{d-1} g_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})x_j. \tag{2.10}$$

For fixed $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} = z$, model (2.10) could be viewed as a (local) linear regression model. The least squares estimator $\hat{\boldsymbol{\theta}} \equiv \hat{\boldsymbol{\theta}}(z)$ given in expression (2.6) entails

$$\mathrm{RSS}_d(z) = \sum_{t=1}^{n} \left[Y_t - \sum_{j=0}^{d-1} \{\hat{g}_j(z) + \hat{\dot{g}}_j(z)(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t - z)\}X_{tj} \right]^2 K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t - z)\, w(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t).$$

The 'degrees of freedom' of $\mathrm{RSS}_d(z)$ are $m(d, z) = n_z - p(d, z)$ where $n_z = \mathrm{tr}\{\mathcal{W}(z)\}$ may be viewed as the number of observations used in the local estimation and

$$p(d, z) = \mathrm{tr}\{\boldsymbol{\Sigma}(z)\, \mathcal{X}^{\mathrm{T}}(z)\, \mathcal{W}^2(z)\, \mathcal{X}(z)\}$$

as the number of local parameters. Now we define the AIC for this model as

$$\mathrm{AIC}_d(z) = \log\{\mathrm{RSS}_d(z)/m(d, z)\} + 2\, p(d, z)/n_z.$$

To delete the least significant variable among $x_0, x_1, \ldots, x_{d-1}$, we search for x_k such that both $g_k(z)$ and $\dot{g}_k(z)$ are close to 0. The t-statistics for those two variables in the (local) linear regression are

$$t_k(z) = \frac{\hat{g}_k(z)}{\surd\{c_k(z)\,\mathrm{RSS}(z)/m(d, z)\}},$$

$$t_{d+k} = \frac{\hat{\dot{g}}_k(z)}{\surd\{c_{d+k}(z)\,\mathrm{RSS}(z)/m(d, z)\}}$$

respectively, where $c_k(z)$ is the $(k+1, k+1)$th element of matrix $\boldsymbol{\Sigma}(z)\, \mathcal{X}^{\mathrm{T}}(z)\, \mathcal{W}^2(z)\, \mathcal{X}(z)\, \boldsymbol{\Sigma}(z)$. Discarding a common factor, we define

$$T_k^2(z) = \hat{g}_k(z)^2/c_k(z) + \hat{\dot{g}}_k(z)^2/c_{d+k}(z).$$

Letting j be the minimizer of $T_k^2(z)$ over $0 \leqslant k < d$, we delete x_j from the full model (2.10). This leads to a model with $d - 1$ 'linear terms'. Repeating the above process, we may define $\mathrm{AIC}_l(z)$ for all $1 \leqslant l \leqslant d$. If

$$\mathrm{AIC}_k(z) = \min_{1 \leqslant l \leqslant d} \{\mathrm{AIC}_l(z)\},$$

the model selected should have $k - 1$ linear terms x_j.

2.3. Implementation

We now outline the algorithm.

Step 1: standardize the data $\{\mathbf{X}_t\}$ such that they have sample mean 0 and sample covariance matrix $\mathbf{I}_d$. Specify an initial value of $\boldsymbol{\beta}$.

Step 2: for each prescribed bandwidth value h_k, $k = 1, \ldots, q$, repeat (a) and (b) below until two successive values of $R(\boldsymbol{\beta})$ defined in equation (2.7) differ insignificantly.

(a) For a given $\boldsymbol{\beta}$, estimate $g_j(\cdot)$ by expression (2.6).
(b) For given $g_j(\cdot)$, search $\boldsymbol{\beta}$ using the algorithm described in Section 2.2.2.

Step 3: for $k = 1, \ldots, q$, calculate $\mathrm{GCV}(h_k)$ with $\boldsymbol{\beta}$ equal to its estimated value, where $\mathrm{GCV}(\cdot)$ is defined in Section 2.2.3. Let $\hat{a}_1$ and $\hat{a}_2$ be the minimizer of $\Sigma_{k=1}^q \{\mathrm{GCV}(h_k) - a_0 - a_1 h_k^4 - a_2/nh_k\}^2$. Define $\hat{h} = (\hat{a}_2/4n\hat{a}_1)^{1/5}$ if $\hat{a}_1$ and $\hat{a}_2$ are positive, and $\hat{h} = \arg\min_{h_k}\{\mathrm{GCV}(h_k)\}$ otherwise.
Step 4: for $h = \hat{h}$ selected in step 3, repeat (a) and (b) in step 2 until two successive values of $R(\boldsymbol{\beta})$ differ insignificantly.
Step 5: for $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ selected in step 4, apply the stepwise deletion of Section 2.2.4 to select significant variables X'_{tj}s for each fixed z.

Some additional remarks are now in order.

Remark 2.

(a) The standardization in step 1 also ensures that the sample mean of $\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t\}$ is 0 and the sample variance is 1 for any unit vector $\boldsymbol{\beta}$. This effectively rewrites model (2.3) as

$$\sum_{j=0}^{d} g_j\{\boldsymbol{\beta}^{\mathrm{T}}\hat{\Sigma}^{-1/2}(\mathbf{x} - \hat{\mu})\}x_j,$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are the sample mean and sample variance respectively. In the numerical examples in Section 4, we report $\hat{\Sigma}^{-1/2}\hat{\boldsymbol{\beta}}/||\hat{\Sigma}^{-1/2}\hat{\boldsymbol{\beta}}||$ as the estimated value of $\boldsymbol{\beta}$ defined in model (2.3).
(b) We may let $w(z) = I(|z| \leqslant 2 + \delta)$ for some small $\delta \geqslant 0$. To speed up the computation further, we estimate the functions $g_j(\cdot)$ in step 3 on 101 regular grids in the interval $[-1.5, 1.5]$ first, and then estimate the values of the functions on this interval by linear interpolation. Finally, in step 4, we estimate the $g_j(\cdot)$ on the interval $[-2, 2]$.
(c) With the Epanechnikov kernel, we let $q = 15$ and $h_k = 0.2 \times 1.2^{k-1}$ in step 3. The specified values for h practically cover the range of 0.2–2.57 times the standard deviation of the data. If we use the Gaussian kernel, we may select the range of the bandwidth between 0.1 and 1.5 times the standard deviation.
(d) To stabilize the search for $\boldsymbol{\beta}$ further in step 2(b), we replace an estimate of $g_j(\cdot)$ on a grid point by a weighted average on its five nearest neighbours with weights $\{1/2, 1/6, 1/6, 1/12, 1/12\}$. The edge points are adjusted accordingly.
(e) In searching for $\boldsymbol{\beta}$ in terms of the one-step iterative algorithm, we estimate the derivatives of $g_j(\cdot)$ based on their adjusted estimates on the grid points as follows:

$$\hat{\dot{g}}_j(z) = \{\hat{g}_j(z_1) - \hat{g}_j(z_2)\}/(z_1 - z_2), \qquad j = 0, \ldots, d,$$
$$\hat{\ddot{g}}_j(z) = \{\hat{g}_j(z_1) - 2\,\hat{g}_j(z_2) + \hat{g}_j(z_3)\}/(z_1 - z_2)^2, \qquad j = 0, \ldots, d,$$

where $z_1 > z_2 > z_3$ are three nearest neighbours of z among the 101 regular grid points (see (b) above). Equation (2.8) should be iterated a few times instead of just once.
(f) Although the algorithm proposed works well with the examples reported in Section 4, its convergence requires further research. In practice, we may detect whether an estimated $\boldsymbol{\beta}$ is likely to be the global minimum by using multiple initial values.

3. Varying-coefficient linear models with two indices

In this section, we consider varying-coefficient models with two indices but one of them known. We assume knowledge of one index to keep computations practically feasible.

Let Y and V be two random variables and $\mathbf{X}$ be a $d \times 1$ random vector. We use V to denote the known index, which could be a (known) linear combination of $\mathbf{X}$. The goal is to approximate the conditional expectation $G(\mathbf{x}, v) = E(Y|\mathbf{X} = \mathbf{x}, V = v)$, in the mean-square sense (see equation (2.1)), by a function of the form

$$g(\mathbf{x}, v) = \sum_{j=0}^{d-1} g_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, v)x_j, \tag{3.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^{\mathrm{T}}$ is a $d \times 1$ unknown unit vector. Similar to theorem 1, part (b), it may be proved that, under some mild conditions on $g(\mathbf{x}, v)$, the expression on the right-hand side of equation (3.1) is unique if the first non-zero β_k is positive and $\beta_d \neq 0$. Let $\{(\mathbf{X}_t, V_t, Y_t); 1 \leqslant t \leqslant n\}$ be observations from a strictly stationary process; $(\mathbf{X}_t, V_t, Y_t)$ has the same distribution as $(\mathbf{X}, V, Y)$.

The estimation for $g(\mathbf{x}, v)$ can be carried out in a similar manner to that for the one-index case (see Section 2.3). We outline the algorithm below briefly.

Step 1: standardize the data $\{\mathbf{X}_t\}$ such that it has sample mean 0 and sample covariance matrix $\mathbf{I}_d$. Standardize the data $\{V_t\}$ such that V_t has sample mean 0 and sample variance 1. Specify an initial value of $\boldsymbol{\beta}$.
Step 2: for each prescribed bandwidth value h_k, $k = 1, \ldots, q$, repeat (a) and (b) below until two successive values of $R(\boldsymbol{\beta})$ defined in equation (3.2) differ insignificantly.

(a) For a given $\boldsymbol{\beta}$, estimate $g_j(\cdot, \cdot)$ in terms of local linear regression.
(b) For given $g_j(\cdot, \cdot)$, search for $\boldsymbol{\beta}$ by using a one-step iteration algorithm.

Step 3: for $k = 1, \ldots, q$, calculate $\mathrm{GCV}(h_k)$ with $\boldsymbol{\beta}$ equal to its estimated value, where $\mathrm{GCV}(\cdot)$ is defined as in Section 2.2.3. Let $\hat{a}_1$ and $\hat{a}_2$ be the minimizer of $\Sigma_{k=1}^{q}\{\mathrm{GCV}(h_k) - a_0 - a_1 h_k^4 - a_2/nh_k^2\}^2$. Define $\hat{h} \equiv (\hat{a}_2/2n\hat{a}_1)^{1/6}$.
Step 4: for $h = \hat{h}$ selected in step 3, repeat (a) and (b) in step 2 until two successive values of $R(\boldsymbol{\beta})$ differ insignificantly.
Step 5: for $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ from step 4, select local significant variables for each fixed (z, v).

Remark 3.

(a) In step 2(a) above, local linear regression estimation leads to the problem of minimizing the sum

$$\sum_{t=1}^{n}\left[Y_t - \sum_{j=0}^{d-1}\{a_j + b_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t - z) + c_j(V_t - v)\}X_{tj}\right]^2 K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t - z, V_t - v)\, w(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t, V_t),$$

where $K_h(z, v) = h^{-2}\, K(z/h, v/h)$, $K(\cdot, \cdot)$ is a kernel function on $\Re^2$ and $w(\cdot, \cdot)$ is a bounded weight function with a bounded support in $\Re^2$. We use a common bandwidth h for simplicity. The estimators derived are $\hat{g}_j(z, v) = \hat{a}_j$, $\hat{\dot{g}}_{j,z}(z, v) = \hat{b}_j$ and $\hat{\dot{g}}_{j,v}(z, v) = \hat{c}_j$ for $j = 0, \ldots, d-1$, where $\dot{g}_{j,z}(z, v) = \partial g_j(z, v)/\partial z$ and $\dot{g}_{j,v}(z, v) = \partial g_j(z, v)/\partial v$.

(b) In step 2(b), we search for $\boldsymbol{\beta}$ which minimizes the function

$$R(\boldsymbol{\beta}) = \frac{1}{n}\sum_{t=1}^{n}\left\{Y_t - \sum_{j=0}^{d-1} g_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t, V_t)X_{tj}\right\}^2 w(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_t, V_t). \tag{3.2}$$

A one-step iterative algorithm may be constructed for this purpose in a manner similar to that in Section 2.2.2. The required estimates for the second derivatives of $g_j(z, v)$ may be obtained via a partially local quadratic regression.

(c) In step 3, the estimated $g(\mathbf{x}, v)$ is linear in $\{Y_t\}$ (for a given $\boldsymbol{\beta}$). Thus, the GCV method outlined in Section 2.2.3 is still applicable.

(d) Locally around given indices $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ and v, equation (3.1) is approximately a linear model. Thus, the local variable selection technique outlined in Section 2.2.4 is still applicable in step 5.

4. Numerical properties

We use the Epanechnikov kernel in our calculation. The one-step iterative algorithm that was described in Section 2.2.2 is used to estimate the index $\boldsymbol{\beta}$, in which we iterate the ridge version of equation (2.8) 2–4 times. We stop the search in step 2 when either the two successive values of $R(\boldsymbol{\beta})$ differ by less than 0.001 or the number of replications of (a) and (b) in step 2 exceeds 30. Setting initially the ridge parameter $q_n = 0.001n^{-1/2}$, we keep doubling its value until $\ddot{R}_r(\cdot)$ is no longer ill conditioned with respect to the precision of computers.

4.1. Simulation

We demonstrate the finite sample performance of the varying-coefficient model with one index in example 1, and with two indices in example 2. We use the absolute inner product $|\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\beta}}|$ to measure the goodness of the estimated direction $\hat{\boldsymbol{\beta}}$. Their inner product represents the cosine of the angles between the two directions. For example 1, we evaluate the performance of the estimator in terms of the mean absolute deviation error

$$\mathcal{E}_{\mathrm{MAD}} = \frac{1}{101d} \sum_{j=0}^{d-1} \sum_{k=1}^{101} |\hat{g}_j(z_k) - g_j(z_k)|,$$

where $z_k,\ k = 1, \ldots, 101$, are the regular grid points on $[-2, 2]$ after the standardization. For example 2, $\mathcal{E}_{\mathrm{MAD}}$ is calculated on the observed values instead.

4.1.1. Example 1

Consider the regression model

$$Y_t = 3\ \exp(-Z_t^2) + 0.8Z_tX_{t1} + 1.5\ \sin(\pi Z_t)X_{t3} + \varepsilon_t,$$

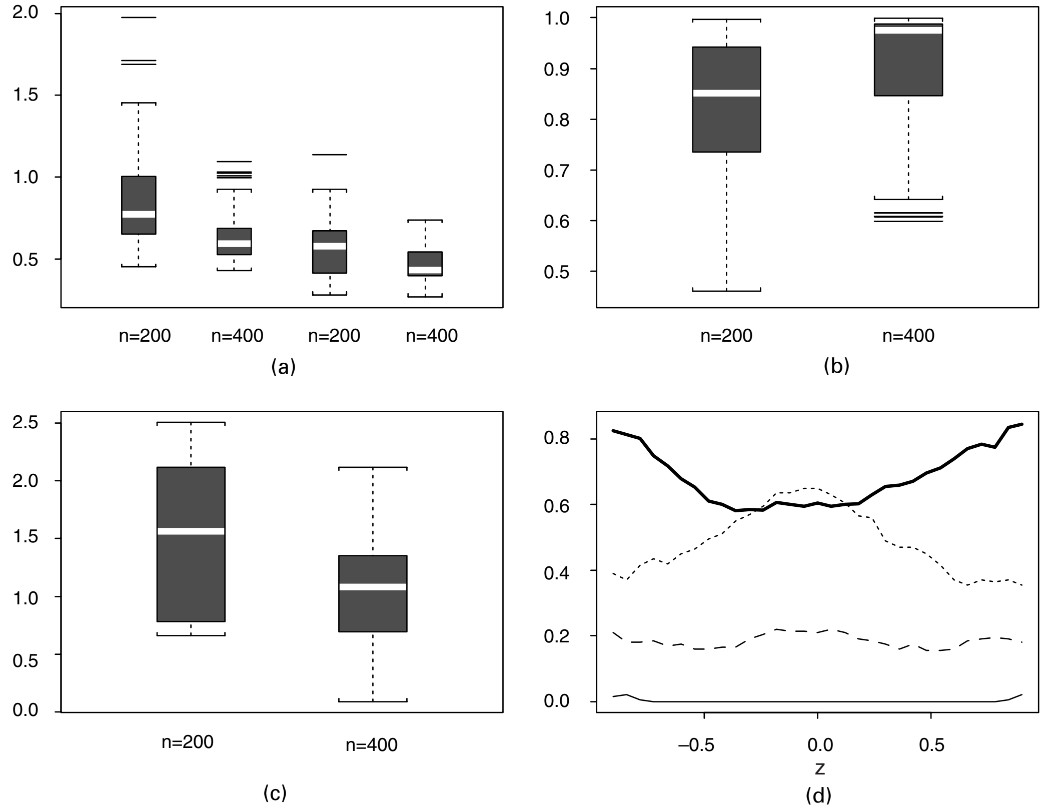
with

$$Z_t = \tfrac{1}{3}(X_{t1} + 2X_{t2} + 2X_{t4}),$$

where $\mathbf{X}_t \equiv (X_{t1}, \ldots, X_{t4})^{\mathrm{T}}$, for $t \geqslant 1$, are independent random vectors uniformly distributed on $[-1, 1]^4$, and the ε_t are independent $N(0, 1)$ random variables. The regression function in this model is of form (2.3) with $d = 4$, $\boldsymbol{\beta} = \frac{1}{3}(1, 2, 0, 2)^{\mathrm{T}}$ and

$$\begin{aligned} g_0(z) &= 3\exp(-z^2), \\ g_1(z) &= 0.8z, \\ g_2(z) &\equiv 0, \\ g_3(z) &= 1.5\ \sin(\pi z). \end{aligned}$$

We conducted two simulations with sample size 200 and 400 respectively, each with 200 replications. The central processor unit time for each replication with sample size 400 is about 18 s on a Pentium II 350 MHz personal computer (Linux). The results are summarized in Fig. 1. Fig. 1(a) displays box plots of the mean absolute deviation errors. We also plot the errors obtained by using the true direction $\boldsymbol{\beta}$. The deficiency due to unknown $\boldsymbol{\beta}$ decreases when the sample size increases. Fig. 1(b) shows that the estimator $\hat{\boldsymbol{\beta}}$ derived from the one-step iterative algorithm is close to the true $\boldsymbol{\beta}$ with high frequency. The average iteration time in the search for $\boldsymbol{\beta}$ is 14.43 s for $n = 400$ and 18.25 s for $n = 200$. Most outliers in Figs 1(a) and 1(b) correspond to the cases where the search for $\boldsymbol{\beta}$ does not converge within 30 iterations. Fig. 1(c) indicates that the bandwidth selector proposed is stable. We also apply the method in Section 2.2.4 to choose the local significant variables at the 31 regular grid points in the range from -1.5 to 1.5 times the standard deviations of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$. The relative frequencies of deletion are depicted in Fig. 1(d). There is overwhelming evidence for including the 'intercept' $g_0(z) = 3\exp(-z^2)$ in the model for all the values of z. In contrast, we tend to delete most often the term X_{t2} which has 'coefficient' $g_2(z) \equiv 0$. There is strong evidence for keeping the term X_{t3} in the model. Note that the term X_{t1} is less significant, as the magnitude of its coefficient $g_1(z) = 0.8z$ is smaller than that of both $g_0(z)$ and $g_3(z)$.



(a)

(b)

(c)

(d)

Fig. 1. Simulation results for example 1: (a) box plots of $\mathcal{E}_{\mathrm{MAD}}$ (the two plots on the left are based on $\hat{\beta}$, and the two plots on the right are based on the true β); (b) box plots of $|\beta^{\mathrm{T}}\hat{\beta}|$; (c) box plots of selected bandwidths; (d) plots of the relative frequencies for deletion of locally insignificant terms at z against z (———, intercept; -------, X_{t1}; ———, X_{t2}; – – –, X_{t3})

Fig. 2 presents three typical examples of the estimated coefficient functions with the sample size $n = 400$. The curves are plotted in the range from -1.5 to 1.5 times the standard deviation of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$. The three examples are selected with the corresponding $\mathcal{E}_{\mathrm{MAD}}$ at the first quartile, the median and the third quartile among the 200 replications. For the example with $\mathcal{E}_{\mathrm{MAD}}$ at the median, we also plot the estimated functions obtained by using the true index $\boldsymbol{\beta}$. For that example, $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{\beta} = 0.946$. The deficiency due to unknown $\boldsymbol{\beta}$ is almost negligible. Note that the biases of the estimators for the coefficient functions $g_0(\cdot)$, $g_1(\cdot)$ and $g_2(\cdot)$ are large near to boundaries. We believe that this is due to the collinearity of functions g and small effective local sample sizes near the tails. Nevertheless, there seems no evidence that the problem has distorted the estimation for the target function $g(\mathbf{x})$.

We also repeated the exercise with $\varepsilon_t \sim N(0, \sigma^2)$ for different values of σ^2. Although the results have the same pattern as before, estimations for both coefficient functions and the direction $\boldsymbol{\beta}$ are more accurate for the models with smaller noise.

4.1.2. Example 2

We consider the regression model

$$Y_t = 3\ \exp(-Z_t^2 + X_{t1}) + (Z_t + X_{t1}^2)X_{t1} - \log(Z_t^2 + X_{t1}^2)X_{t2} + 1.5\ \sin(\pi Z_t + X_{t1})X_{t3} + \varepsilon_t,$$

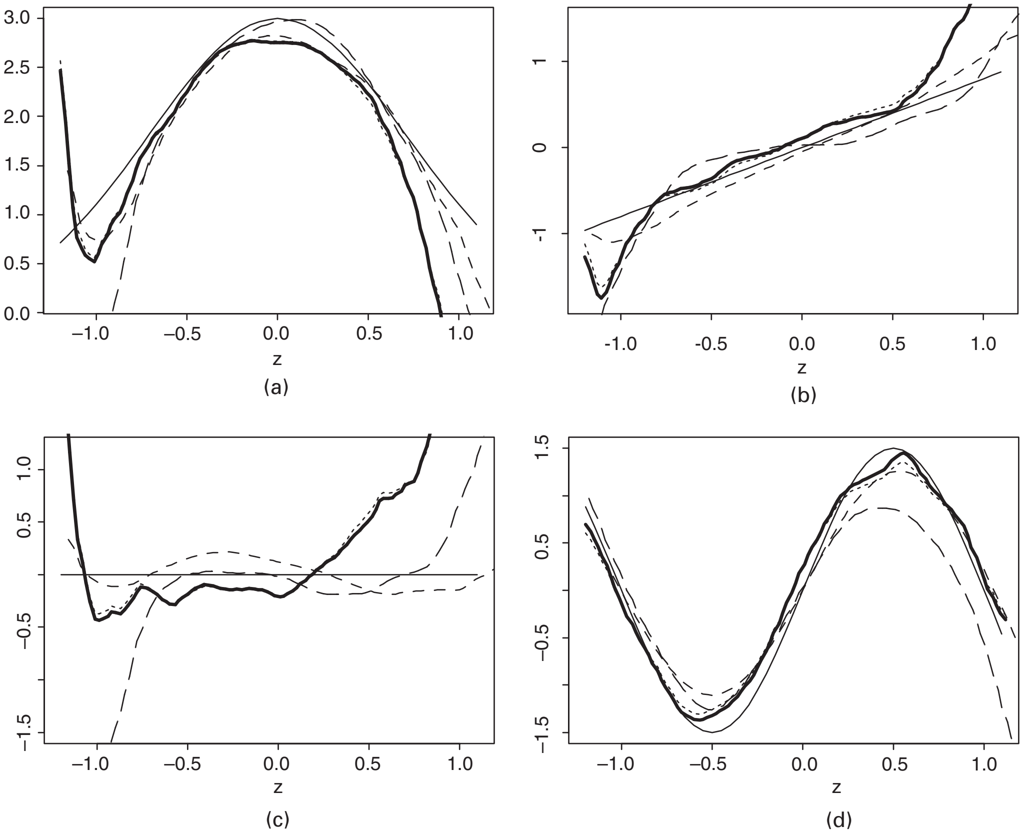


Fig. 2. Simulation results for example 1 ($n = 400$)—estimated coefficient functions with $\mathcal{E}_{\mathrm{MAD}}$ at the first quartile (– – –), third quartile (— — —) and median (———) with the true β (·······), together with the true functions (———): (a) $g_0(z) = 3\ \exp(-z^2)$; (b) $g_1(z) = 0.8z$; (c) $g_2(z) = 0$; (d) $g_3(z) = 1.5\ \sin(\pi z)$

with

$$Z_t = \tfrac{1}{2}(X_{t1} + X_{t2} + X_{t3} + X_{t4}),$$

where $\{X_{t1}, \ldots, X_{t4}\}$ and $\{\varepsilon_t\}$ are the same as in example 1. Obviously, the regression function in this model is of form (3.1) with $d = 4$, $\boldsymbol{\beta} = \tfrac{1}{2}(1, 1, 1, 1)^{\mathrm{T}}$, $V_t = X_{t1}$ and

$$\begin{aligned} g_0(z, v) &= 3\ \exp(-z^2 + v), \\ g_1(z, v) &= z + v^2, \\ g_2(z, v) &= -\log(z^2 + v^2), \\ g_3(z, v) &= 1.5\ \sin(\pi z + v). \end{aligned}$$

We conducted three simulations with sample size 200, 400 and 600 respectively, each with 100 replications. The central processor unit time for each realization, for a Sun Ultra-10 300 MHz workstation, was about 18 s for $n = 200$, 80 s for $n = 400$ and 190 s for $n = 600$. Fig. 3(a) shows that the mean absolute deviation error $\mathcal{E}_{\mathrm{MAD}}$ decreases when n increases. For comparison, we also present $\mathcal{E}_{\mathrm{MAD}}$ based on the true $\boldsymbol{\beta}$. Fig. 3(b) displays the box plots of the absolute inner product $|\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{\beta}}|$, which indicates that the one-step iteration algorithm works reasonably well. The box plots of bandwidths selected by the GCV method are depicted in Fig. 3(c).

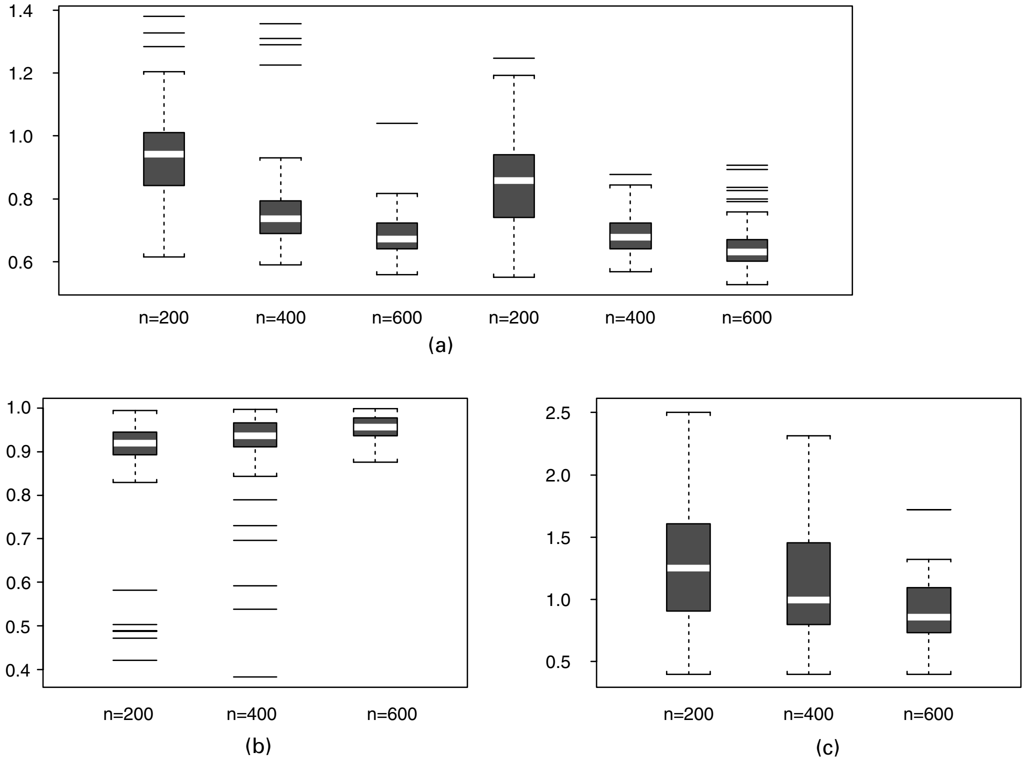


Fig. 3. Simulation results for example 2: box plots of (a) $\mathcal{E}_{\mathrm{MAD}}$, (b) $|\beta^{\mathrm{T}}\hat{\beta}|$ and (c) the bandwidths selected (the three plots on the left of (a) are based on $\hat{\beta}$, and the three on the right are based on the true β)

4.2. Real data examples

4.2.1. Example 3

The annual numbers of muskrats and mink caught over 82 trapping regions have been recently extracted from the records compiled by the Hudson Bay Company on fur sales at auction in 1925–1949. Fig. 4 indicates the 82 posts where furs were collected. Fig. 5 plots the time series for the mink and the muskrat (on the natural logarithmic scale) from eight randomly selected posts. There is clear synchrony between the fluctuations of the two species with a delay of 1 or 2 years, indicating the food chain interaction between prey (i.e. muskrat) and predator (i.e. mink); see Errington (1963). A simple biological model for the food chain interaction proposed by May (1981) and Stenseth *et al.* (1997) is of the form

$$\begin{aligned} X_{t+1} - X_t &= a_0(\theta_t) - a_1(\theta_t) X_t - a_2(\theta_t) Y_t, \\ Y_{t+1} - Y_t &= b_0(\theta_t) - b_1(\theta_t) Y_t + b_2(\theta_t) X_t, \end{aligned} \tag{4.1}$$

where X_t and Y_t denote the population abundances, on a natural logarithmic scale, of prey (muskrat) and predator (mink) respectively at time t, $a_i(\cdot)$ and $b_i(\cdot)$ are non-negative functions and θ_t is an indicator representing the regime effect at time t, which is determined by X_t and/or Y_t. The term 'regime effect' collectively refers to the non-linear effect due to, among other things,

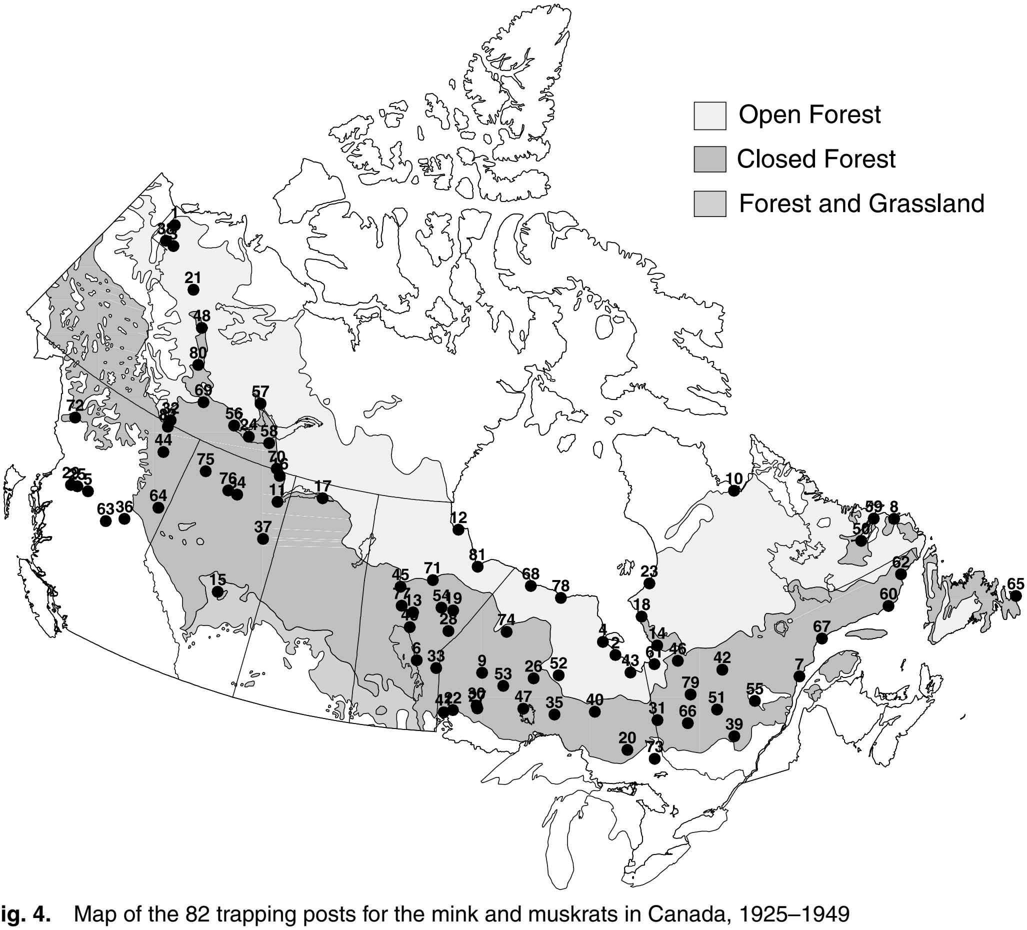


Fig. 4. Map of the 82 trapping posts for the mink and muskrats in Canada, 1925–1949

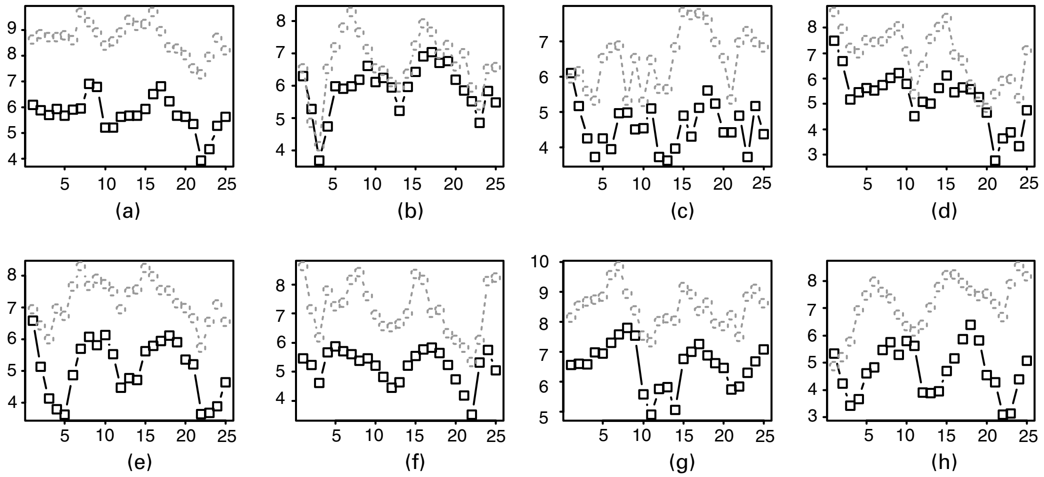


Fig. 5. Time series plots of the mink and muskrat data from eight randomly selected posts (———, mink; - - - - - -, muskrat): (a) post 4; (b) post 7; (c) post 20; (d) post 26; (e) post 43; (f) post 62; (g) post 74; (h) post 79

the different hunting or escaping behaviour and reproduction rates of animals at different stages of population fluctuation (Stenseth *et al.*, 1998). In fact, $a_1(\theta_t)$ and $b_1(\theta_t)$ reflect within-species regulation whereas $a_2(\theta_t)$ and $b_2(\theta_t)$ reflect the food chain interaction between the two species.

Model (4.1), with added random noise, would be in the form of varying-coefficient linear models if we let θ_t be a linear combination X_t and Y_t. However, each mink and muskrat time series has only 25 points, which is too short for fitting such a non-linear model. On the basis of some statistical tests on the common structure for each pair among those 82 posts, Yao *et al.* (2000) suggested a grouping with three clusters: the eastern area consisting of post 10, post 67 and the other six posts on its right in Fig. 4, the western area consisting of the 30 posts on the left in Fig. 4 (i.e. post 17 and those on its left) and the central area consisting of the remaining 43 posts in the middle. Since some data are missing at post 15, we exclude it from our analysis. The sample size for the eastern, central and western areas are therefore 207, 989 and 667 respectively. With the new technique proposed in this paper, we fitted the pooled data for each of the three areas with the model

$$\begin{aligned} X_{t+1} &= f_0(Z_t) + f_1(Z_t)Y_{t-1} + f_2(Z_t)Y_t + f_3(Z_t)X_{t-1} + \varepsilon_{1,t+1}, \\ Y_{t+1} &= g_0(Z_t) + g_1(Z_t)Y_{t-1} + g_2(Z_t)Y_t + g_3(Z_t)X_{t-1} + \varepsilon_{2,t+1}, \end{aligned} \tag{4.2}$$

where $Z_t = \beta_1 Y_{t-1} + \beta_2 Y_t + \beta_3 X_{t-1} + \beta_4 X_t$ with $\boldsymbol{\beta} \equiv (\beta_1, \beta_2, \beta_3, \beta_4)^{\mathrm{T}}$ selected by the data. Comparing with model (4.1), we include further lagged values X_{t-1} and Y_{t-1} in model (4.2). To eliminate the effect of different sampling weights in different regions and for different species, we first standardized mink and muskrat series separately for each post. We apply the local deletion technique presented in Section 2.2.4 to detect *local* redundant variables at 31 regular grid points over the range from -1.5 to 1.5 times the standard deviation of Z_t. As a first attempt, we also selected the *global* model based on the local deletion. A more direct approach would be, for example, based on the generalized likelihood ratio tests of Fan *et al.* (2001). We denote by R_{MSE} the ratio of the MSEs from the fitted model over the sample variance of the variable to be fitted.

First, we use the second of equations (4.2) to model mink population dynamics in the central area. The selected $\boldsymbol{\beta}$ is $(0.424, 0.320, 0.432, 0.733)^{\mathrm{T}}$, the selected bandwidth is 0.415 and $R_{\mathrm{MSE}} = 0.449$. The local variable selection indicates that X_{t-1} is the least significant overall, for it is significant at only seven out of 31 grid points. By leaving it out, we reduce the model to

$$Y_{t+1} = g_0(Z_t) + g_y(Z_t)Y_t + g_x(Z_t)X_t + \varepsilon_{2,t+1}, \tag{4.3}$$

where $Z_t = \beta_1 Y_t + \beta_2 X_t + \beta_3 Y_{t-1}$. Our algorithm selects

$$Z_t = 0.540Y_t - 0.634Y_{t-1} + 0.553X_t, \tag{4.4}$$

which suggests that the non-linearity is dominated by the growth rate of mink (i.e. $Y_t - Y_{t-1}$) and the population of muskrat (i.e. X_t) in the previous year. The estimated coefficient functions are plotted in Fig. 6(a). The coefficient function $g_x(\cdot)$ is positive, which reflects the fact that a large muskrat population will facilitate the growth of the mink population. The coefficient function $g_y(\cdot)$ is also positive, which reflects the natural reproduction process of the mink population. Both $g_y(\cdot)$ and $g_x(\cdot)$ are approximately increasing with respect to the sum of the growth rate of mink and the population of muskrat; see equation (4.4). All the terms in model (4.3) are significant in most places; the number of significant grid points for the intercept, Y_t and X_t are 21, 31 and 26 (out of 31 in total). The selected bandwidth is 0.597 and $R_{\mathrm{MSE}} = 0.461$.

Starting with the first of equations (4.2), the fitted model for the muskrat dynamics in the central area is

$$X_{t+1} = f_0(Z_t) + f_y(Z_t)Y_t + f_x(Z_t)X_t + \varepsilon_{1,t+1} \tag{4.5}$$

with $Z_t = 0.542Y_t + 0.720X_t + 0.435X_{t-1}$, $\hat{h} = 0.498$ and $R_{\mathrm{MSE}} = 0.559$. The estimated coefficient functions are plotted in Fig. 6(b). The coefficient function $f_y(\cdot)$ is always negative, which reflects the fact that mink is the key predator of the muskrat in this core of the boreal forest in Canada. The coefficient $f_x(\cdot)$ is positive, as expected.

We repeated the exercise for pooled data in the western area, which yielded similar results. In fact, model (4.3) appears appropriate for mink dynamics with $Z_t = 0.469Y_t + 0.723X_t + 0.507Y_{t-1}$, $R_{\mathrm{MSE}} = 0.446$, $\hat{h} = 0.415$ and the estimated coefficient functions plotted in Fig. 6(c). Model (4.5) appears appropriate for muskrat dynamics with $Z_t = 0.419Y_t + 0.708X_t + 0.569X_{t-1}$, $\hat{h} = 0.415$, $R_{\mathrm{MSE}} = 0.416$ and the estimated coefficient functions plotted in Fig. 6(d).

Fitting the data in the eastern area leads to drastically different results from the previous results. The fitting for the mink dynamics with model (4.3) gives $Z_t = 0.173Y_t - 0.394X_t + 0.901Y_{t-1}$, $\hat{h} = 0.597$ and $R_{\mathrm{MSE}} = 0.681$. Out of 31 grid points, the intercept, Y_t and X_t are significant at 15, 31 and four points respectively. There is clear autodependence in the mink series $\{Y_t\}$ whereas the muskrat data $\{X_t\}$ carry little information about minks. The estimated coefficients, depicted in Fig. 6(e), reinforce this observation. The fitting of the muskrat dynamics shows again that there seems little interaction between mink and muskrat in this area. For example, the term Y_t in model (4.5) is not significant at all 31 grid points. The estimated coefficient function $f_y(\cdot)$ is plotted as the bold curve in Fig. 6(f), which is always close to 0. We fitted the data with the further simplified model

$$X_{t+1} = f_0(Z_t) + f_x(Z_t)X_t + \varepsilon_{1,t+1},$$

resulting in $Z_t = 0.667X_t - 0.745X_{t-1}$, $\hat{h} = 0.498$ and $R_{\mathrm{MSE}} = 0.584$. The estimated coefficient functions are superimposed on Fig. 6(f). Note that the different ranges of z-values are due to different Z_ts used in the above model and model (4.5).

In summary, we have facilitated the data analysis of the biological food chain interaction model of Stenseth *et al.* (1997) by portraying the non-linearity through varying-coefficient linear forms. The selection of the index in our algorithm is equivalent in this context to the selection of the regime effect indicator, which in itself is of biological interest. The numerical results indicate that there is strong evidence of predator–prey interactions between the minks and the

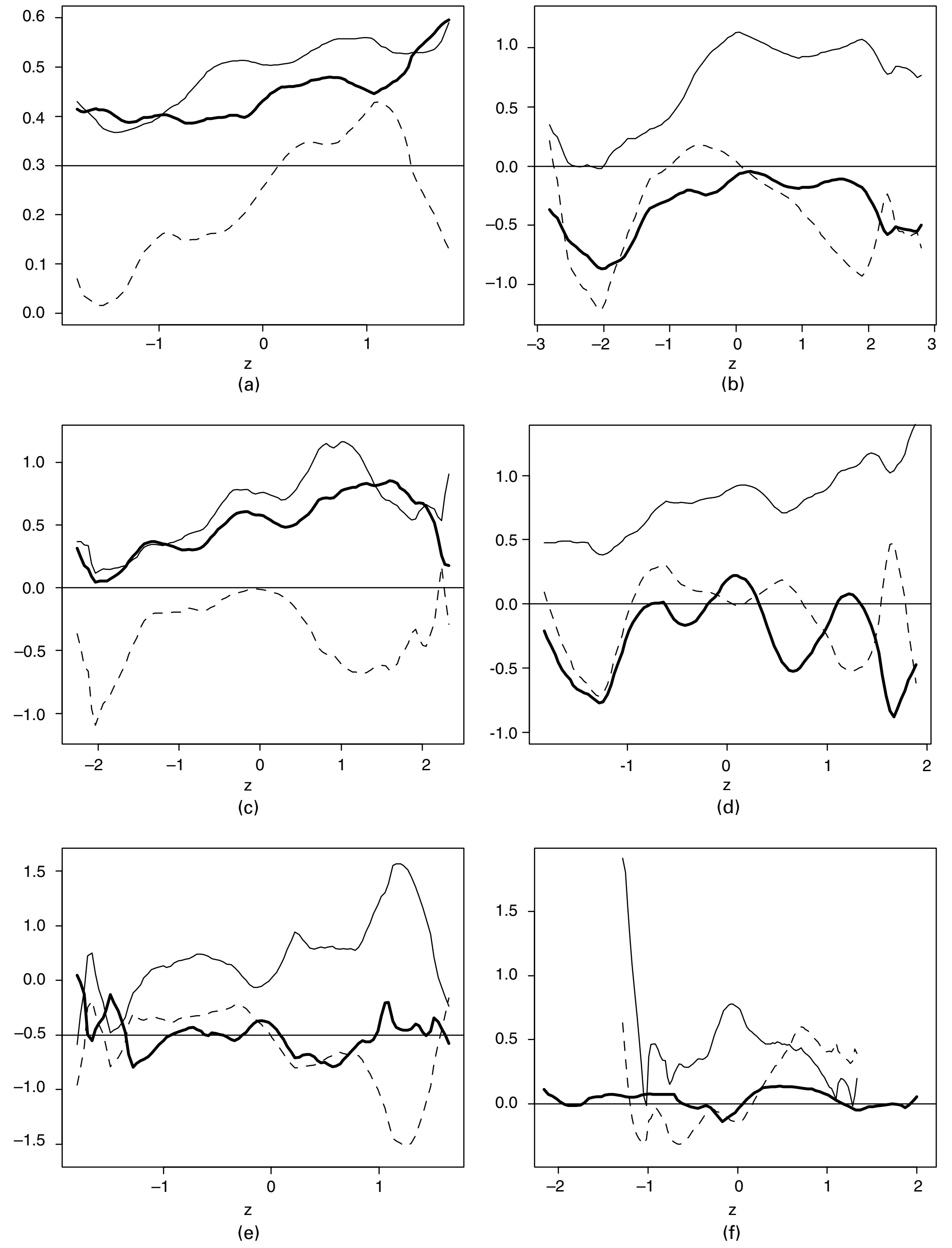
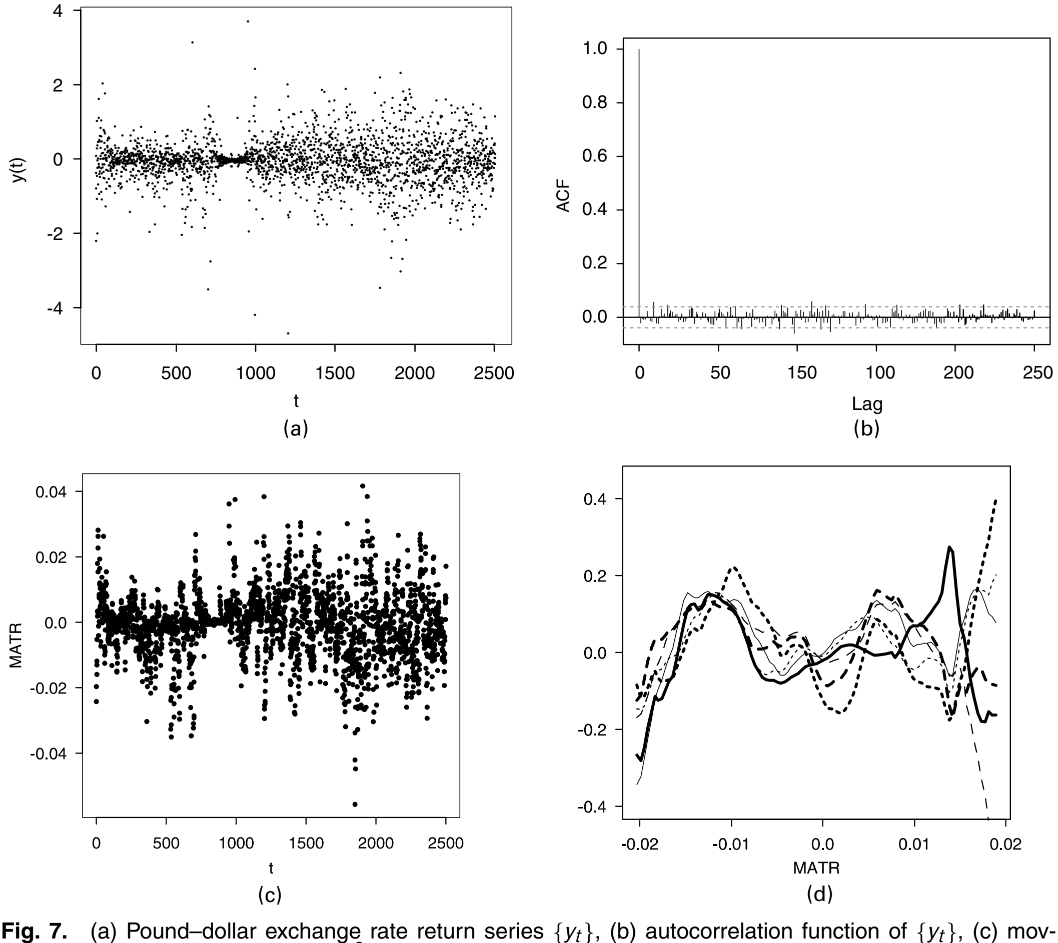


Fig. 6. Estimated coefficient functions for the Canadian mink–muskrat data: (a) mink model for the central area (———, $g_x(\cdot)$; ———, $g_y(\cdot)$; - - - - - - -, $g_0(\cdot)$); (b) muskrat model for the central area (———, $f_y(\cdot)$; ———, $f_x(\cdot)$; - - - - - - -, $f_0(\cdot)$); (c) mink model for the western area (———, $g_x(\cdot)$; ———, $g_y(\cdot)$; - - - - - - -, $g_0(\cdot)$); (d) muskrat model for the western area (———, $f_y(\cdot)$; ———, $f_x(\cdot)$; - - - - - - -, $f_0(\cdot)$); (e) mink model for the eastern area (———, $g_x(\cdot)$; ———, $g_y(\cdot)$; - - - - - - -, $g_0(\cdot)$); (f) muskrat model for the eastern area (———, $f_y(\cdot)$; ———, $f_x(\cdot)$; - - - - - - -, $f_0(\cdot)$)

muskrats in the central and western areas. However, no evidence for such an interaction exists in the eastern area. In the light of what is known about the eastern area, this is not surprising. There is a larger array of prey species for the mink to feed on, making it less dependent on muskrat (see Elton (1942)).

4.2.2. Example 4

Example 4 concerns the daily closing bid prices of the pound sterling in terms of the US dollar from January 2nd, 1974, to December 30th, 1983, which forms a time series of length 2510. These data are available from `ftp://ftp.econ.duke.edu/pub/arg/data`. The previous analysis of this 'particularly difficult' data set can be found in Gallant *et al.* (1991) and the references therein. Let X_t be the exchange rate on the tth day. We model the return series $Y_t = 100 \log(X_t/X_{t-1})$, plotted in Fig. 7(a), by using the techniques developed in this paper. Typically classical financial theory would treat $\{Y_t\}$ as a martingale difference process. Therefore Y_t would be unpredictable. Fig. 7(b) shows that there is almost no significant autocorrelation in $\{Y_t\}$.



(a)

(b)

(c)

(d)

Fig. 7. (a) Pound–dollar exchange rate return series $\{y_t\}$, (b) autocorrelation function of $\{y_t\}$, (c) moving average trading rule $U_t = Y_t/(\Sigma_{i=0}^{9} Y_{t-i}/10)$ and (d) estimated coefficient functions of model (4.6) with $Z_t = U_{t-1}$ and $m = 5$ (——, $g_0(\cdot)$; ········, $g_1(\cdot)$; -------, $g_2(\cdot)$; ——, $g_3(\cdot)$; ·······, $g_4(\cdot)$; -------, $g_5(\cdot)$)

First, we approximate the conditional expectation of Y_t (given its past) by

$$g_0(Z_t) + \sum_{i=1}^{m} g_j(Z_t) Y_{t-i}, \tag{4.6}$$

where $Z_t = \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 X_{t-1} + \beta_4 U_{t-1}$, and

$$U_{t-1} = X_{t-1} \left(L^{-1} \sum_{j=1}^{L} X_{t-j} \right)^{-1} - 1.$$

The variable U_{t-1} defines the *moving average trading rule* (MATR) in finance, and $U_{t-1} + 1$ is the ratio of the exchange rate at time $t - 1$ to the average rate over the past period of length L. The MATR signals 1 (the position to *buy* sterling) when $U_{t-1} > 0$ and -1 (the position to *sell* sterling) when $U_{t-1} < 0$. For a detailed discussion of the MATR, we refer to LeBaron (1997, 1999) and Hong and Lee (1999). We use the first 2410 sample points for estimation and last 100 points for post-sample forecasting. We evaluate the post-sample forecast by the *mean trading return* defined as

$$\text{MTR} = \frac{1}{100} \sum_{t=1}^{100} S_{2410+t-1} Y_{2410+t},$$

where S_t is a signal function taking values -1, 0 and 1. The mean trading return measures the real profits in a financial market, ignoring interest differentials and transaction costs (for simplicity). It is more relevant than the conventional mean-squared predictive errors or average absolute predictive errors for evaluating the performance of forecasting for market movements; see Hong and Lee (1999). Under this criterion, we need to predict the direction of market movement rather than its magnitude. For the MATR, the mean trading return is defined as

$$\text{MTR}_{\text{MA}} = \frac{1}{100} \sum_{t=1}^{100} \{ I(U_{2410+t-1} > 0) - I(U_{2410+t-1} < 0) \} Y_{2410+t}.$$

Let $\hat{Y}_t$ be defined as the estimated function (4.6). The mean trading return for the forecasting based on our varying-coefficient modelling is defined as

$$\text{MTR}_{\text{VC}} = \frac{1}{100} \sum_{t=1}^{100} \{ I(\hat{Y}_{2410+t} > 0) - I(\hat{Y}_{2410+t} < 0) \} Y_{2410+t}.$$

However, ideally we would buy at time $t - 1$ when $Y_t > 0$ and sell when $Y_t < 0$. The mean trading return for this 'ideal' strategy is

$$\text{MTR}_{\text{ideal}} = \frac{1}{100} \sum_{t=1}^{100} |Y_{2410+t}|,$$

which serves as a bench-mark for assessing forecasting procedures. For example, for this particular data set, $\text{MTR}_{\text{MA}}/\text{MTR}_{\text{ideal}} = 12.58\%$ if we let $L = 10$.

Now we are ready to proceed. First, we let $m = 5$ and $L = 10$ in expression (4.6), i.e. we use 1-week data in the past as the 'regressors' in the model and define the MATR by comparing with the average rate in the previous 2 weeks. The selected β is $(0.0068, 0.0077, 0.0198, 0.9998)^{\text{T}}$, which suggests that U_t plays an important role in the underlying non-linear dynamics. The ratio of the MSE of the fitted model to the sample variance of $\{Y_t\}$ is 93.67%, which reflects the presence of high level 'noise' in the financial data. The bandwidth selected is 0.24. The ratio $\text{MTR}_{\text{VC}}/\text{MTR}_{\text{ideal}} = 5.53\%$. The predictability is much lower than that of the MATR. If we include rates in the previous 2 weeks as regressors in the model (i.e. $m = 10$ in expression

(4.6)), the ratio $\mathrm{MTR}_{\mathrm{VC}}/\mathrm{MTR}_{\mathrm{ideal}}$ increases to 7.26% which is still a considerable distance from $\mathrm{MTR}_{\mathrm{MA}}/\mathrm{MTR}_{\mathrm{ideal}}$, whereas the ratio of the MSE of the fitted model to the sample variance of $\{Y_t\}$ is 87.96%. The bandwidth selected is still 0.24, and $\hat{\beta} = (0.0020, 0.0052, 0.0129, 0.9999)^{\mathrm{T}}$.

These calculations (and also others not reported here) suggest that U_t could be the dominant component in the index selected. This leads us to use model (4.6) with fixed $Z_t = U_{t-1}$, which was the approach adopted by Hong and Lee (1999). For $m = 5$, the fit to the data used in the estimation becomes worse; the ratio of the MSE of the fitted model to the sample variance of $\{Y_t\}$ is 97.39%. But it provides better post-sample forecasting; $\mathrm{MTR}_{\mathrm{VC}}/\mathrm{MTR}_{\mathrm{ideal}}$ is 23.76%. The bandwidth selected is 0.24. The plots of estimated coefficient functions indicate possible undersmoothing. By increasing the bandwidth to 0.40, $\mathrm{MTR}_{\mathrm{VC}}/\mathrm{MTR}_{\mathrm{ideal}}$ is 31.35%. The estimated coefficient functions are plotted in Fig. 7(d). The rate of correct predictions for the direction of market movement (i.e. the sign of Y_t) is 50% for the MATR and 53% and 58% for the varying-coefficient model with bandwidths 0.24 and 0.40 respectively.

A word of caution: we should not take for granted the above improvement in forecasting from using U_t as the index. Hong and Lee (1999) conducted empirical studies of this approach with several financial data sets with only partial success. In fact, for this particular data set, model (4.6) with $Z_t = U_t$ and $m = 10$ gives a negative value of $\mathrm{MTR}_{\mathrm{VC}}$. Note that the 'superdominant' position of U_t in the selected smoothing variable $\hat{\beta}^{\mathrm{T}}\mathbf{X}_t$ is partially due to the scaling difference between U_t and (Y_t, X_t); see also Fig. 7(a) and Fig. 7(c). In fact, if we standardize U_t, Y_t and X_t separately beforehand, the resulting $\hat{\beta}$ is $(0.59, -0.52, 0.07, 0.62)^{\mathrm{T}}$ when $m = 5$, which is dominated by U_{t-1} and the contrast between Y_{t-1} and Y_{t-2}. ($\mathrm{MTR}_{\mathrm{VC}}/\mathrm{MTR}_{\mathrm{ideal}} = 1.42\%$. The ratio of the MSE of the fitted model to the sample variance of Y_t is 96.90%.) By doing this, we effectively use a different class of models to approximate the unknown conditional expectation of Y_t; see remark 2, part (a). Finally, we remark that a different modelling approach should be adopted if our primary target is to maximize the mean trading return, which is obviously beyond the scope of this paper.

Acknowledgements

We thank Professor N. C. Stenseth and Professor A. R. Gallant for making available the Canadian mink–muskrat data and pound–dollar exchange data analysed in Section 4.2. We are grateful to the reviewers for very helpful comments.

Jianqing Fan was supported partially by National Science Foundation grant DMS-0196041, Qiwei Yao was supported partially by Engineering and Physical Sciences Research Council grant L16385 and Biological and Biotechnology Sciences Research Council–Engineering and Physical Sciences Research Council grant 96/MMI09785 and Zongwu Cai was supported partially by National Science Foundation grant DMS-0072400 and funds provided by the University of North Carolina at Charlotte.

Appendix A: Proof of theorem 1

We use the same notation as in Section 2.

A.1. Proof of part (a)

It follows from ordinary least squares theory that there is a minimal value of

$$E[\{Y - f(X)\}^2 \mid \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X} = z]$$

over the class of functions of the form $f(\mathbf{x}) = \Sigma_{i=0}^{d} f_i(\boldsymbol{\alpha}^{\tau}\mathbf{x})x_i$ with all f_i measurable. Let $f_0^*(z), \ldots, f_{d-1}^*(z)$

be the minimizer. Then

$$\{f_1^*(z), \ldots, f_d^*(z)\}^{\mathrm{T}} = \{\mathrm{var}(\mathbf{X}|\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X} = z)\}^{-} \,\mathrm{cov}(\mathbf{X}, Y|\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X} = z),$$

$$f_0^*(z) = E(Y|\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X} = z) - \sum_{j=1}^{d} f_j^*(z)\, E(X_j|\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X} = z).$$

In the first expression, A^- denotes a generalized inverse matrix of A for which $AA^-A = A$. It follows immediately from least squares theory that

$$E\left[\left\{Y - f_0^*(z) - \sum_{j=1}^{d} f_j^*(z) X_j\right\}^2 \middle| \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X} = z\right] \leqslant \mathrm{var}(Y|\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X} = z).$$

Consequently,

$$R(\boldsymbol{\alpha}) \equiv E\left\{Y - f_0^*(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}) - \sum_{j=1}^{d} f_j^*(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}) X_j\right\}^2$$

is bounded from above by $\mathrm{var}(Y)$ and is continuous on the compact set $\{\boldsymbol{\alpha} \in R^d| \,||\boldsymbol{\alpha}|| = 1\}$. Hence, there is a $\boldsymbol{\beta}$ in the above set such that $R(\boldsymbol{\alpha})$ obtains its minimum at $\boldsymbol{\alpha} = \boldsymbol{\beta}$. Therefore, $g(\cdot)$ fulfilling equation (2.1) exists.

A.2. Proof of part (b)

Theorem 1, part (b), follows from the following two lemmas immediately.

Lemma 1. Suppose that $F(\cdot) \not\equiv 0$ is a twice-differentiable function defined on R^d, and

$$F(\mathbf{x}) = g_0(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) + \sum_{j=1}^{d} g_j(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) x_j \tag{A.1}$$

$$= f_0(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}) + \sum_{j=1}^{d} f_j(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}) x_j, \tag{A.2}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are non-zero and non-parallel vectors in R^d. Then $F(\mathbf{x}) = c_1\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x} + c_0$, where $\boldsymbol{\gamma} \in R^d$ and $c_0, c_1 \in R$ are constants.

Proof. Without loss of generality we assume that $\boldsymbol{\beta} = (c, 0, \ldots, 0)^{\mathrm{T}}$. Then, it follows from equation (A.1) that $\partial^2 F(\mathbf{x})/\partial x_i^2 = 0$ for $i = 2, \ldots, d$. Write $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x} = z$. Choose $2 \leqslant i \leqslant d$ fixed for which $\alpha_i \neq 0$. Then, from equation (A.2), we have that

$$\frac{\partial^2 F(\mathbf{x})}{\partial x_i^2} = \alpha_i^2 \ddot{f}_0(z) + \alpha_i^2 \sum_{j=1}^{d} \ddot{f}_j(z) x_j + 2\alpha_i \dot{f}_i(z) = 0,$$

namely

$$\alpha_i\{\alpha_i \ddot{f}_0(z) + z\ddot{f}_i(z) + 2\dot{f}_i(z)\} + \alpha_i^2 \sum_{j \neq i} \left\{\ddot{f}_j(z) - \frac{\alpha_j}{\alpha_i}\ddot{f}_i(z)\right\} x_j = 0. \tag{A.3}$$

Letting $x_j = 0$ for $j \neq i$ and $x_i = x/\alpha_i$ in the above equation, we have

$$\alpha_i \ddot{f}_0(x) + x\ddot{f}_i(x) + 2\dot{f}_i(x) = 0. \tag{A.4}$$

Hence, equation (A.3) reduces to

$$\sum_{j \neq i} \left\{\ddot{f}_j(z) - \frac{\alpha_j}{\alpha_i}\ddot{f}_i(z)\right\} x_j = 0,$$

which leads to the equalities below if we let $x_k = x/\alpha_k$ and all other $x_j = 0$ for $k \neq i$ and $\alpha_k \neq 0$, or $x_k \neq 0$, $x_i = x/\alpha_i$ and all other $x_j = 0$ for $k \neq i$ and $\alpha_k = 0$:

$$\ddot{f}_k(x) = \ddot{f}_i(x)\frac{\alpha_k}{\alpha_i}, \qquad 1 \leqslant k \leqslant d.$$

This implies that $f_k(z) = f_i(z)\alpha_k\alpha_i^{-1} + a_k z + b_k$ with $a_i = b_i = 0$. Substituting this into equation (A.2), we have

$$\begin{aligned} F(\mathbf{x}) &= f_0(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}) + \alpha_i^{-1} f_i(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x} + \sum_{j\neq i}(a_j\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x} + b_j)x_j \\ &\equiv f_0^*(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}) + \sum_{j\neq i}(a_j\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x} + b_j)x_j. \end{aligned}$$

Now, an application of argument (A.4) to the last expression above shows that $f_0^*(z) = a_0 z + b_0$. Thus

$$F(\mathbf{x}) = a_0\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x} + b_0 + \sum_{j\neq i}(a_j\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x} + b_j)x_j.$$

Now, $\partial^2 F(\mathbf{x})/\partial x_i\,\partial x_j = a_j\alpha_i$ for any $j \geqslant 2$, which should be 0 according to equation (A.1) since $\boldsymbol{\beta} = (c, 0, \ldots, 0)^{\mathrm{T}}$. Hence, all a_js ($j \geqslant 2$) in the above expression are 0. This implies that

$$F(\mathbf{x}) = \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x} + b_0 + a_1x_1\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x} = \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x} + b_0 + c^{-1}a_1\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x},$$

where $\boldsymbol{\gamma} = a_0\boldsymbol{\alpha} + \mathbf{b}$, and $\mathbf{b} = (b_1, \ldots, b_d)^{\mathrm{T}}$.

Lemma 2. For any

$$F(\mathbf{x}) \equiv F(x_1, \ldots, x_d) = f_0(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}) + \sum_{j=1}^{d} f_j(\alpha^{\mathrm{T}}\mathbf{x})x_j \not\equiv 0,$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^{\mathrm{T}} \in R^d$ and $\alpha_d \neq 0$, $F(\cdot)$ can be expressed as

$$F(\mathbf{x}) = g_0(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}) + \sum_{j=1}^{d-1} g_j(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})x_j, \tag{A.5}$$

where $g_0(\cdot), \ldots, g_{d-1}(\cdot)$ are uniquely determined as follows:

$$g_0(z) = F(0, \ldots, 0, z/\alpha_d), \tag{A.6}$$

$$g_j(z) = F_j - g_0(z), \qquad j = 1, \ldots, d-1, \tag{A.7}$$

where F_j denotes the value of F at $x_j = 1$, $x_d = (z - \alpha_j)/\alpha_d$ and $x_k = 0$ for all the other ks.

Proof. Note that

$$x_d = \left\{\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x} - \sum_{j=1}^{d-1}\alpha_j x_j\right\}\bigg/\alpha_d.$$

Define

$$\begin{aligned} g_0(z) &= f_0(z) + \frac{1}{\alpha_d}f_d(z)z, \\ g_j(z) &= f_j(z) - \frac{\alpha_j}{\alpha_d} \qquad \text{for } j = 1, \ldots, d-1. \end{aligned}$$

It is easy to see that equation (A.5) follows immediately. Letting $x_1 = \ldots = x_{d-1} = 0$ and $x_d = z/\alpha_d$ in equation (A.5), we obtain equation (A.6). Letting $x_j = 1$, $x_d = (z - \alpha_j)/\alpha_d$ and $x_k = 0$ for all the other ks, we obtain equation (A.7). The proof is completed.

References

Bickel, P. J. (1975) One-step Huber estimates in linear models. *J. Am. Statist. Ass.*, **70**, 428–433.

Cai, Z., Fan, J. and Li, R. (2000) Efficient estimation and inferences for varying-coefficient models. *J. Am. Statist. Ass.*, **95**, 888–902.

Cai, Z., Fan, J. and Yao, Q. (2000) Functional-coefficient regression models for nonlinear time series models. *J. Am. Statist. Ass.*, **95**, 941–956.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997) Generalized partially linear single-index models. *J. Am. Statist. Ass.*, **92**, 477–489.

Carroll, R. J., Ruppert, D. and Welsh, A. H. (1998) Local estimating equations. *J. Am. Statist. Ass.*, **93**, 214–227.

Chen, R. and Tsay, R. S. (1993) Functional-coefficient autoregressive models. *J. Am. Statist. Ass.*, **88**, 298–308.

Cleveland, W. S., Grosse, E. and Shyu, W. M. (1992) Local regression models. In *Statistical Models in S* (eds J. M. Chambers and T. J. Hastie), pp. 309–376. Pacific Grove: Wadsworth and Brooks.

Craven, P. and Wahba, G. (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377–403.

Elton, C. S. (1942) *Voles, Mice and Lemmings*. Oxford: Clarendon.

Errington, P. L. (1963) *Muskrat Populations*. Ames: Iowa State University Press.

Fan, J. and Chen, J. (1999) One-step local quasi-likelihood estimation. *J. R. Statist. Soc.* B, **61**, 927–943.

Fan, J. and Gijbels, I. (1996) *Local Polynomial Modeling and Its Applications*. London: Chapman and Hall.

Fan, J., Zhang, C. and Zhang, J. (2001) Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.*, **29**, 153–193.

Fan, J. and Zhang, J.-T. (2000) Two-step estimation of functional linear models with applications to longitudinal data. *J. R. Statist. Soc.* B, **62**, 303–322.

Fan, J. and Zhang, W. (1999) Statistical estimation in varying-coefficient models. *Ann. Statist.*, **27**, 1491–1518.

Fan, J. and Zhang, W. (2000) Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scand. J. Statist.*, **27**, 715–731.

Gallant, A. R., Hsieh, D. A. and Tauchen, G. E. (1991) On fitting a recalcitrant series: the pound/dollar exchange rate, 1974-1983. In *Nonparametric and Semiparametric Methods in Econometrics and Statistics* (eds W. A. Barnett, J. Powell and G. E. Tauchen), pp. 199–240. Cambridge: Cambridge University Press.

Härdle, W., Hall, P. and Ichimura, H. (1993) Optimal smoothing in single-index models. *Ann. Statist.*, **21**, 157–178.

Hastie, T. and Tibshirani, R. (1993) Varying-coefficient models (with discussion). *J. R. Statist. Soc.* B, **55**, 757–796.

Hong, Y. and Lee, T.-H. (1999) Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Rev. Econ. Statist.*, to be published.

Hoover, D. R., Rice, J. A., Wu, C. O. and Yang, L.-P. (1998) Nonparametric smoothing estimates of time-varying-coefficient models with longitudinal data. *Biometrika*, **85**, 809–822.

Hurvich, C. M., Simonoff, J. S. and Tsai, C.-L. (1998) Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. R. Statist. Soc.* B, **60**, 271–293.

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

Kauermann, G. and Tutz, G. (1999) On model diagnostics using varying-coefficient models. *Biometrika*, **86**, 119–128.

LeBaron, B. (1997) Technical trading rule and regime shifts in foreign exchange. In *Advances in Trading Rules* (eds E. Acar and S. Satchell), pp. 5–40. Oxford: Butterworth–Heinemann.

LeBaron, B. (1999) Technical trading rule profitability and foreign exchange intervention. *J. Int. Econ.*, **49**, 125–143.

May, R. M. (1981) Models for two interacting populations. In *Theoretical Ecology* (ed. R. M. May), pp. 78–104. Oxford: Blackwell.

Newey, W. K. and Stoker, T. M. (1993) Efficiency of weighted average derivative estimators and index models. *Econometrica*, **61**, 1199–1223.

Nicholls, D. F. and Quinn, B. G. (1982) Random coefficient autoregressive models: an introduction. *Lect. Notes Statist.*, **11**.

Ramsay, J. O. and Silverman, B. W. (1997) *The Analysis of Functional Data*. New York: Springer.

Ruppert, D. (1997) Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Am. Statist. Ass.*, **92**, 1049–1062.

Samarov, A. M. (1993) Exploring regression structure using nonparametric functional estimation. *J. Am. Statist. Ass.*, **88**, 836–847.

Seifert, B. and Gasser, Th. (1996) Finite-sample variance of local polynomial: analysis and solutions. *J. Am. Statist. Ass.*, **91**, 267–275.

Simonoff, J. S. and Tsai, C. L. (1999) Semiparametric and additive model selection using an improved Akaike information criterion. *Comput. Graph. Statist.*, **8**, 22–40.

Stenseth, N. C., Falck, W., Bjørnstad, O. N. and Krebs, C. J. (1997) Population regulation in snowshoe hare and Canadian lynx; asymmetric food web configurations between hare and lynx. *Proc. Natn. Acad. Sci. USA*, **94**, 5147–5152.

Stenseth, N. C., Falck, W., Chan, K. S., Bjørnstad, O. N., Tong, H., O'Donoghue, M., Boonstra, R., Boutin, S., Krebs, C. J. and Yoccoz, N. G. (1998) From patterns to processes: phase- and density-dependencies in Canadian lynx cycle. *Proc. Natn. Acad. Sci. Wash.*, **95**, 15430–15435.

Wahba, G. (1977) A survey of some smoothing problems and the method of generalized cross-validation for solving them. In *Applications of Statistics* (ed. P. R. Krisnaiah), pp. 507–523. Amsterdam: North-Holland.

Wu, C. O., Chiang, C.-T. and Hoover, D. R. (1998) Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *J. Am. Statist. Ass.*, **93**, 1388–1402.

Xia, Y. and Li, W. K. (1999a) On the estimation and testing of functional-coefficient linear models. *Statist. Sin.*, **9**, 735–757.

Xia, Y. and Li, W. K. (1999b) On single-index coefficient regression models. *J. Am. Statist. Ass.*, **94**, 1275–1285.

Yao, Q., Tong, H., Finkenstädt, B. and Stenseth, N. C. (2000) Common structure in panels of short time series. *Proc. R. Soc. Lond.* B, **267**, 2459–2467.

Zhang, W. and Lee, S. Y. (1999) On local polynomial fitting of varying-coefficient models. To be published.

Zhang, W. and Lee, S. Y. (2000) Variable bandwidth selection in varying-coefficient models. *J. Multiv. Anal.*, **74**, 116–134.