ADAPTIVE VARYING-COEFFICIENT LINEAR MODELS

by

Jianqing Fan[1]
Department of Statistics, University of North Carolina, Chapel Hill

Qiwei Yao[2]
Department of Statistics, London School of Economics and Political Science

Zongwu Cai
Department of Mathematics, University of North Carolina, Charlotte

Contents:

Abstract
1. Introduction
2. Overview of varying-coefficient modesl
3. Adaptive varying-coefficient linear models
4. Varying-coefficient linear models with two indices
5. Numerical properties

Appendix: Proof of Theorem 1
References
Figure 7(a)

The Suntory Centre
Suntory and Toyota International Centres
for Economics and Related Disciplines
London School of Economics and Political Science

Discussion Paper
No. EM/00/388
April 2000

Houghton Street
London WC2A 2AE
Tel.: 020 7405 7686

---

[1] Supported partially by NSF grant DMS-9803200.
[2] Supported partially by EPSRC Grant L16385 and BBSRC/EPSRC Grant 96/MMI09785.
We thank Professors N C Stenseth and A R Gallant for making available Canadian mink-muskrat data and pound/dollar exchange data analysed in Section 4.2.

Abstract

Varying-coefficient linear models arise from multivariate nonparametric regression, nonlinear time series modelling and forecasting, functional data analysis, longitudinal data analysis, and others. It has been a common practice to assume that the vary-coefficients are functions of a given variable which is often called an *index.* A frequently asked question is which variable should be used as the index. In this paper, we explore the class of the varying-coefficient linear models in which the index is unknown and is estimated as a linear combination of regression and/or other variables. This will enlarge the modelling capacity substantially. We search for the index such that the derived varying-coefficient model provides the best approximation to the underlying unknown multi-dimensional regression function in the least square sense. The search is implemented through the newly proposed hybrid backfitting algorithm. The core of the algorithm is the alternative iteration between estimating the index through a one-step scheme and estimating coefficient functions through a one-dimensional local linear smoothing. The generalised cross-validation method for choosing bandwidth is efficiently incorporated into the algorithm. The locally significant variables are selected in terms of the combined use of *t*-statistic and Akaike information criterion. We further extend the algorithm for the models with two indices. Simulation shows that the proposed methodology has appreciable flexibility to model complex multivariate nonlinear 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; generalised cross-validation; local linear regression; local significant variable selection; one-step estimation; smoothing index; varying-coefficient linear models.

JEL No.: C22

1 Introduction

During the recent years, with increasing computing power it has become commonplace to access and to attempt to analyze data of unprecedented size and complexity. With these changes has come an increasing demand for the development of computationally intensive methodologies which are designed to identify complicated data structures at not too excessive computing cost. Data-analytic techniques developed from statistical prospective views have been proved powerful for exploiting hidden structures in high-dimensional data. Witness of this includes, among others, additive modeling (Breiman and Friedman, 1985; Hastie and Tibshirani, 1990), low-dimensional interaction modeling (Friedman, 1991; Gu and Wahba, 1993; Stone *et al.*, 1996), multiple-index models (Friedman and Stuetzle, 1991; Härdle and Stoker, 1989; Li, 1991), partially linear models (Wahba, 1984; Green and Silverman, 1994), varying-coefficient linear models (Cleveland *et al.*, 1992; Hastie and Tibshirani, 1993), and their hybrids (Carroll *et al.*, 1997; Fan Härdle and Mammen, 1998). Those models are designed to attenuate the so-called 'curse of dimensionality' problem by exploring low-dimensional structures, although different models explore different aspects of high-dimensional data and incorporate different prior knowledge. The aim of the exercises is to reduce possible modeling bias and to let data select a model which describes themselves well. Depending on each particular data set, some methods perform better and are more appropriate to use than others, but none of them is uniformly superior. They together provide useful statistical toolkits for exploring hidden structures in high-dimensional data. For general knowledge of nonparametric and semi-parametric modeling techniques, we refer to the books by Hastie and Tibshirani (1990), Wahba (1990), Green and Silverman (1994), Wand and Jones (1995), Fan and Gijbels (1996) and Simonoff (1996).

Suppose we are interested in estimating 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

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

where $\boldsymbol{\beta} \in \Re^d$ is an unknown direction, $\mathbf{x} = (x_1, \ldots, x_d)^T$, $x_0 = 1$, and coefficients $g_0(\cdot), \ldots, g_d(\cdot)$ are unknown functions. We choose the direction $\boldsymbol{\beta}$ and coefficient functions $\{g_j(\cdot)\}$ such that $E\{G(\mathbf{X}) - g(\mathbf{X})\}^2$ obtains its minimum. The appeal of this model is that once $\boldsymbol{\beta}$ is known, we can directly estimate $g_j(\cdot)'$s by the standard one-dimensional kernel regression localized around $\boldsymbol{\beta}^T\mathbf{x}$. 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. The model (1.1) appears linear in each coordinates of $\mathbf{x}$ when the index $\boldsymbol{\beta}^T\mathbf{x}$ is fixed. It may include additional quadratic and cross-product terms of x_j's (or more generally any given functions of x_j's) as 'new' components of $\mathbf{x}$. Hence it is in fact considerably flexible to cater to complex multivariate nonlinear structure.

We develop an efficient back-fitting algorithm to estimate $g(\cdot)$. The virtue of the algorithm

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

The form of the model (1.1) is not new. It was proposed in Ichimura (1993). Recently, Xia and Li (1999a) extended the idea and the results of Härdle, Hall and Ichimura (1993) from the single-index model to the adaptive varying-coefficient model (1.1). Their basic idea is to estimate the coefficient functions with a given bandwidth and a direction $\boldsymbol{\beta}$, and then choose the bandwidth and the direction by the cross-validation. Based on the assumption that the bandwidth is of the order $O(n^{-1/5})$ and the direction $\boldsymbol{\beta}$ is within an $O_p(n^{-1/2})$ consistent neighborhood of the true value, they obtained some interesting theoretical results. However, the approach suffers from the heavy computational expenses. This somehow explains why most previous work focused on the case when the direction $\boldsymbol{\beta}$ is given and is parallel to one of coordinates. See §2 for an overview. The new approach in this paper differs from those in the literature in three key aspects: (a) only one-dimensional smoother is used in estimation, (b) the index coefficient $\boldsymbol{\beta}$ is estimated by data and (c) within a local region around $\boldsymbol{\beta}^T\mathbf{x}$, we select significant variables x_j's to avoid overfitting. Aspect (b) is different from Härdle, Hall and Ichimura (1993) and Xia and Li (1999a) since we estimate the coefficient functions and the direction simultaneously; no cross-validation is needed. This idea is similar *in spirit* to that of Caroll *et al.* (1997) who showed that a semiparametric efficient estimator of the direction $\boldsymbol{\beta}$ can be obtained via this approach. Further we provide a theorem (i.e. Theorem 1(ii) in §3 below) on the model identification problem of the form (1.1), which has not been addressed before.

The application of varying-coefficient models is diverse; ranging over generalized linear models, nonlinear time series, functional data analysis, longitudinal data analysis, and other interdisciplinary areas. While these problems are inner related, they are not often referred to each other. In §2, we will give an overview on the current state-of-art of the varying-coefficient models in practice.

The rest of the paper is organized as follows. §3 deals with the adaptive varying-coefficient model (1.1). The extension to the adaptive varying-coefficient models to the case with two indices is outlined in §4. The numerical results of three simulated examples are reported in §5.1, which demonstrate that the proposed methodology is capable to capture complex nonlinear structure with moderate sample sizes, and further the required computation typically takes less than a minute on a Pentium II 350MHz PC. The methodology is further illustrated in §5.2 through Canadian mink-

muskrat data in 1925-1944 and the pound/dollar exchange rates in 1974-1983. All the technical proofs are relegated in the Appendix.

2 Overview of varying-coefficient models

Varying coefficient models have been successfully applied to multi-dimensional nonparametric regression, generalized linear models, nonlinear time series models, longitudinal and functional data analysis, interest rate modeling in finance, international conflict study in political sciences and others. The basic idea is to approximate a unknown multi-dimensional regression function by a (conditionally) linear model with the coefficients being functions of a covariate called index. Most of the work to date assumes that the index is given. The adaptive varying-coefficient models allow data to choose the index automatically. This section presents an overview on the recent development of the varying-coefficient models.

2.1 Varying coefficient models

The varying-coefficient models were introduced by Cleveland, Grosse and Shyu (1992) in the extension of local regression techniques from one-dimensional to multi-dimensional setting. Suppose that we are given a random sample $\{(U_i, \mathbf{X}_i, Y_i); 1 \leq i \leq n\}$, where Y_i is the response variable and $(U_i, \mathbf{X}_i)$ are covariates. The local polynomial regression essentially fits the conditional linear model

$$Y_i = \sum_{j=0}^{d} g_j(U_i) X_{ij} + \varepsilon_i, \tag{2.1}$$

where X_{ij} is the j-th component of $\mathbf{X}_i$, $X_{i0} \equiv 1$, and ε_i has conditional mean zero and conditional variance $\sigma^2(U_i)$, given U_i and $\mathbf{X}_i$. The coefficient functions $\{g_j(\cdot)\}$ are assumed to be smooth. An extension of the local regression technique was given by Hastie and Tibshirani (1993) via introducing kernel weights. Let $K(\cdot)$ be a kernel function on $\Re$ and $h = h_n$ be a bandwidth. Set $K_h(\cdot) = h^{-1}K(\cdot/h)$. For a given u_0 and x close to u_0, it follows a Taylor expansion that

$$g_j(x) \approx g_j(u_0) + g_j'(u_0)(x - u_0) \equiv b_j + c_j(x - u_0). \tag{2.2}$$

Here, the only local linear approximation is used for the sake of simplicity. It can be easily generalized to the local polynomial regression (Fan and Gijbels, 1996). Thus, for those observations where U_i's are around u_0, the data follow an approximation linear model:

$$Y_i \approx \sum_{j=0}^{d} \{b_j + c_j(U_i - u_0)\} X_{ij} + \varepsilon_i.$$

The local parameters can be estimated via a weighted local regression, namely

$$\widehat{g}_j(u_0) = \widehat{b}_j, \quad j = 0, \ldots, d, \tag{2.3}$$

where $\{\widehat{b}_j, \widehat{c}_j\}$ is the least-squares solution which minimizes

$$\sum_{i=1}^{n}\Big[Y_i - \sum_{j=0}^{d}\{b_j + c_j(U_i - u_0)\}X_{ij}\Big]^2 K_h(U_i - u_0). \tag{2.4}$$

The conditional bias and variance of the estimators were derived in Caroll, Ruppert and Welsh (1998) and Fan and Zhang (2000a). As expected, the bias depends only on local approximation error and is of order $O(h_n^2)$, and the variance is of order $O(1/(nh))$ and depends only on the effective number of local data points, the local (conditional) variance and local correlation matrix of the covariates $\mathbf{X}$. The asymptotic normality of the estimators and data-driven bandwidth selection procedure were presented in Zhang and Lee (1999, 2000). Furthermore, the distribution of the maximum discrepancy between the estimated coefficients and true coefficients was discussed by Xia and Li (1999b) and Fan and Zhang (2000b). The confidence bands and hypothesis testing problems were also discussed therein.

Complementary to the local regression technique is the smoothing splines method. Hastie and Tibshirani (1993) proposed a smoothing spline estimator derived via minimizing

$$\sum_{i=1}^{n}\Big\{Y_i - \sum_{j=0}^{d} g_j(U_i)X_{ij}\Big\}^2 + \sum_{j=0}^{d}\lambda_j\|g_j''\|_2^2, \tag{2.5}$$

where $\{\lambda_j\}$ are positive regularization parameters. As an initial attempt, one usually chooses $\lambda_j = \lambda$ for all j. Note that the local regression solves many (usually in the order of 100) weighted regression problems (2.4), while the smoothing spline method solves one large parametric problem (number of parameters is in the order of nd).

The local regression estimator (2.3) assumes implicitly that the coefficient functions $\{g_j(\cdot)\}$ admit a similar degree of smoothness so that they can be equally well approximated in a local neighborhood (see (2.2)). When the functions $\{g_j(\cdot)\}$ have different degrees of smoothness, it is shown in Fan and Zhang (2000a) that the local regression estimator (2.3) is suboptimal under their asymptotic formulation. The intuition is clear: a smooth component asks for a large bandwidth to reduce the variance, while a rough component requires a small bandwidth to reduce the bias. This problem cannot be overcome by, for example, simply using a large bandwidth to estimating a smooth component only; see Fan and Zhang (2000a). While the asymptotic properties for the smoothing spline estimator (2.3) are not easy to derive, we expect that smoothing splines would suffer from the same problem even when $\{\lambda_j\}$ are appropriately specified. However the drawback can be removed by using a two-step procedure proposed in Fan and Zhang (2000a). The basic idea is to get an initial estimator for $\{\widehat{g}_j(\cdot)\}$ using a small bandwidth h_0. The bandwidth h_0 is so small that the biases in estimation of $\{\widehat{g}_j(\cdot)\}$ are negligible. Then, compute the partial residuals

$$\widehat{Y}_{i,j} = Y_i - \sum_{k\neq j}\widehat{g}_k(U_i)X_{ik}$$

and apply the local linear regression technique to the pseudo univariate varying-coefficient model

$$\widehat{Y}_{i,j} = g_j(U_i)X_{ij} + \varepsilon_i$$

using bandwidth h_j to estimate $g_j(\cdot)$. The advantage of this is two folds: the bandwidth h_j can now be selected purposely for estimating $g_j(\cdot)$ only and univariate bandwidth selection techniques can be applied.

When the model (2.1) is misspecified, the above local fitting techniques intend to find the best linear function at each given $U = u_0$ to approximate the regression function $E(Y|U = u, \mathbf{X})$. Similarly, the smoothing spline (2.5) finds the best varying-coefficient function to approximate the regression surface $E(Y|U, \mathbf{X})$.

In nonparametric modeling, we are constantly challenged by the question whether a simpler parametric model fits the data adequately or not. For example, we may ask if the coefficients in the model (2.1) are all constant. This amounts to testing the parametric hypothesis

$$H_0 : g_j(\cdot) = \beta_j, \quad j = 0, \ldots, d,$$

against nonparametric alternative (2.1). We can also ask whether the covariates X_1 and X_2 are significant. This is equivalent to testing

$$H_0 : g_1(\cdot) = 0 \quad \text{and} \quad g_2(\cdot) = 0.$$

In this case, both null and alternative hypotheses are nonparametric. While these questions arise frequently in practice, they are poorly understood. The conventional approach uses the discrepancy measures such as the distances between estimated functions under null and alternative hypotheses. See, for example, Bickel and Rosenblatt (1973), Härdle and Mammen (1993) and Hart (1997). Fan, Zhang and Zhang (1999) argued that these methods were not as fundamental as the likelihood ratio based statistics. Generalized likelihood ratio tests are proposed there for various nonparametric testing problems and the Wilks phenomenon and optimality properties are unveiled. The basic idea of the generalized likelihood ratio tests is to find good estimators under the null and full models and then substitute them into the likelihood function to obtain a likelihood ratio statistic. A fundamental property of the derived test is that the asymptotic null distribution is independent of nuisance functions and is χ^2-distributed. This allows us to use either the asymptotic null distribution or bootstrap methods to determine the p-values of the tests. See also Cai, Fan and Li (2000) for bootstrap estimation of null distributions and empirical power calculations.

2.2 Generalized varying-coefficient models

Varying coefficient models can be readily extended to the context of the generalized linear models. This allows us to model a transform of the regression function by a varying-coefficient model

$$\ell\{E(Y|U, \mathbf{X})\} = \sum_{j=0}^{d} g_j(U)\, X_j$$

with a given link function $\ell(\cdot)$, where $X_0 = 1$. The unknown coefficient functions can be estimated by the local maximum likelihood approach. Namely, the local sum of squares in (2.4) is replaced

by the local likelihood or the local quasi-likelihood (Cai, Fan and Li, 2000). This approach can be viewed as a specific case of the local estimation equation method of Carroll, Ruppert and Welsh (1998). The spline method can also be applied in this context (Hastie and Tibshirani, 1993).

Carroll, Ruppert and Welsh (1998) derived the asymptotic expressions for conditional mean and variance for the local equation estimators. The results can be extended to the generalized varying-coefficient models with some additional work. Cai, Fan and Li (2000) established the asymptotic normality of the local maximum likelihood estimator. They also proposed a fast implementation algorithm based on a one-step local maximum likelihood estimator. The basic idea is to compute genuine local MLEs at a few well-separate grid points and then to use them as initial values for the local MLEs at their nearest grid points via one-step Newton-Raphson iteration. The estimates at all grid points are obtained by repeating the above exercise in which a newly defined estimate is treated as an initial estimate at its next grid point. Cai, Fan and Li (2000) showed that this estimator shares the same asymptotic behavior as the genuine local likelihood estimator. Kauermann and Tutz (1999) proposed a graphical technique to diagnose the discrepancy between a parametric model and a varying-coefficient model. Cai (1999) used a two-step procedure to deal with the situation where the coefficient functions $\{g_j(\cdot)\}$ admit different degrees of smoothness. The testing procedure and estimation method in Cai, Fan and Li (2000) have been successfully applied by Cederman and Penubarti (1999) to the study of international relation conflict in political sciences.

2.3 Nonlinear time series

Varying-coefficient models have been elegantly applied to modeling and forecasting time series data (Nicholls and Quinn, 1982; Chen and Tsay, 1993). They are natural extensions of the thresholded autoregression models of Tong (1990). Let $\{X_t\}$ be a time series. The varying-coefficient model is of form

$$X_t = g_0(X_{t-p}) + \sum_{j=1}^{d} g_j(X_{t-p})X_{t-j} + \varepsilon_t \tag{2.6}$$

for some given lags d and p. The geometric ergodicity of this model was studied by Chen and Tsay (1993), who also proposed a nearest neighborhood type of estimator. The local linear regression estimation (2.4) applies readily to this autoregressive setting. The asymptotic normality of such an estimator has been established in Cai, Fan and Yao (1998). They also proposed a generalized pseudo-likelihood test for testing linear autoregressive models or thresholded models against model (2.6). The procedure is basically the same as the generalized likelihood ratio statistic for the independent data, but now adapts to the time series setting. A bootstrap method is used to estimate the asymptotic null distribution. The testing procedure and estimation method have been successfully applied by Hong and Lee (1999) to the inference and forecast of exchange rates and by Cai and Tiwari (1999) to an environmental study.

2.4 Analysis of longitudinal and functional data

In many applications, observations for different individuals are collected over a period of time. The number of observations for different individuals may be different and so is the time when the data are recorded. This type of data is termed as longitudinal data. Often, interest lies in studying the association between the covariates and the response variable. To this end, a linear model is often employed:

$$Y_i(t_{ij}) = \beta_0 + \mathbf{X}_i(t_{ij})^T\boldsymbol{\beta} + \varepsilon_i(t_{ij}), \tag{2.7}$$

where $(\mathbf{X}_i(t_{ij}), Y_i(t_{ij}))$ is the observed datum for the ith individual at time t_{ij} and $\varepsilon_i(t_{ij})$ is the stochastic noise. The key difference from cross-sectional data is that the error process $\{\varepsilon_i(t_{ij})\}$ within subject i is correlated. See, for example, Diggle, Liang and Zeger (1994) and Hand and Crowder (1996).

Despite of its success in many applications, the model (2.7) does not allow the association to vary over time, even though the covariates and the response variable change over time and environment. To account for this, Zeger and Diggle (1994) and Moyeed and Diggle (1994) proposed a semiparametric model which allows the intercept β_0 to vary over time, but not the other coefficients. To facilitate the genuine variation of the association over time, Brunback and Rice (1998) and Hoover *et al.* (1998) proposed to use the varying-coefficient model

$$Y_i(t_{ij}) = \beta_0(t_{ij}) + \mathbf{X}_i(t_{ij})^T\boldsymbol{\beta}(t_{ij}) + \varepsilon_i(t_{ij}), \tag{2.8}$$

where the coefficient functions are assumed to be smooth functions of time. This is a specific case of the functional linear model discussed in Ramsay and Silverman (1997) in the context of functional data analysis. When there is no covariate, the model (2.8) was studied by Hart and Wehrly (1986, 1993) for repeated measurements and by Rice and Silverman (1991) for functional data. There the mean regression was estimated by the kernel and smoothing spline methods. A 'deleting one-subject each time' cross-validation was proposed in Rice and Silverman (1991) for choosing smoothing parameters.

The coefficients in the model (2.8) can be estimated by the kernel and smoothing spline methods (Brumback and Rice, 1998; Hoover *et al.*, 1998). The basic idea is the same as those outlined in §2.1. Brumback and Rice (1998) pointed out that intensive computation is required for using smoothing splines because one has to invert blindly a matrix of the order of the total number of data points (i.e. sum of the number of repeated measurements for each individual). Fan and Zhang (2000) proposed a two-step method to overcome this drawback. The basic idea is related to the two-step method outlined in §2.1, but now adapts to longitudinal data setting. For each distinct data time point t_j, collect the subjects having observations at time t_j (or more generally around t_j) and fit the linear model (2.7) for those data points. This gives us the initial estimated coefficients at time t_j. In the second step, instead of smoothing on the partial residuals, the initial estimated coefficients are smoothed directly. They reported that this method was more efficient (in terms of computation)

than smoothing splines and more flexible and efficient than the conventional kernel method. The asymptotic bias and variance of kernel method was studied by Hoover *et al.* (1998). Furthermore, Wu, Chiang and Hoover (1998) proposed approaches for constructing confidence regions based on the kernel method.

3 Adaptive varying-coefficient linear models

3.1 Approximation and identifiability

Since $G(\mathbf{x}) = E(Y \mid \mathbf{X} = \mathbf{x})$ 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 $g(\cdot)$ with finite second moment. Therefore, the search for the LS approximation $g(\cdot)$ of $G(\cdot)$, as defined in (1.1), is equivalent to the search for such a $g(\cdot)$ that $E\{Y - g(\mathbf{X})\}^2$ obtains its minimum. Theorem 1(i) below indicates that there always exists such a $g(\cdot)$ under some mild conditions. Obviously, if $G(\mathbf{x})$ is in the form of the RHS of (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 (3.2) below). In fact, the model (1.1) is an over-parametrized form in the sense that one of $\{g_j(\cdot)\}$ can be represented in terms of the others. Theorem 1(ii) confirms that once the direction $\boldsymbol{\beta}$ is specified, the function $g(\cdot)$ has a representation with at most d (instead of $d+1$) $g_j(\cdot)$-functions. Furthermore, those $g_j(\cdot)$-functions are identifiable.

Theorem 1. (i) Assume that the distribution function of $(\mathbf{X}, Y)$ is continuous, and $E\{Y^2 + ||\mathbf{X}||^2\} < \infty$. Then, there exists a $g(\cdot)$ defined by (1.1) for which

$$E\{Y - g(\mathbf{X})\}^2 = \inf_{\boldsymbol{\alpha}} \inf_{f_0, \ldots, f_d} E\left\{ Y - \sum_{j=0}^{d} f_j(\boldsymbol{\alpha}^T \mathbf{X}) X_j \right\}^2, \tag{3.1}$$

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

(ii) 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 that

$$g(\mathbf{x}) = \boldsymbol{\alpha}^T \mathbf{x} \boldsymbol{\beta}^T \mathbf{x} + \boldsymbol{\gamma}^T \mathbf{x} + c, \tag{3.2}$$

where $\boldsymbol{\alpha}$, $\boldsymbol{\gamma} \in \Re^d$, $c \in \Re$ are constants, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are not parallel with each other. Furthermore, once $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$ is given and $\beta_d \neq 0$, we may let $g_d(\cdot) \equiv 0$. Consequently, all the other $g_j(\cdot)'$s 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 RHS of (1.1), there may exist more than one $g(\mathbf{x})'$s, being of the form of (1.1), for which (3.1) 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 of the form of varying-coefficient linear model (1.1). However, (3.1) holds for both $g(x_1,x_2) = 1.25x_1^2$, and $1.25x_2^2$.

Without loss of the generality, we always assume from now on that in the 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 LS approximation of $g(\cdot)$ which cannot be expressed in the form of (3.2).

3.2 Estimation

Suppose that $\{(\mathbf{X}_t,\, Y_t);\, 1 \le t \le n\}$ are observations from a strictly stationary process, and $(\mathbf{X}_t,\, Y_t)$ has the same marginal distribution as $(\mathbf{X},\, Y)$. Of interest is to estimate the surface $g(\cdot)$ defined by (1.1) and (3.1). It is clear from (3.1) that we need to search for the minimizers of $\{f_j(\cdot)\}$ for any given direction $\boldsymbol{\alpha}$ and then find the direction at which the mean squared error (MSE) is minimized. A genuine search is almost always intractable in practice. We adapt a back-fitting algorithm which has been demonstrated to be efficient for solving such a computationally intensive optimization problem.

We assume that $\beta_d \neq 0$. It follows from Theorem 1(ii) that we only search for an approximation in the form

$$g(\mathbf{x}) = \sum_{j=0}^{d-1} g_j(\boldsymbol{\beta}^T\mathbf{x})x_j, \tag{3.3}$$

since the term $g_d(\boldsymbol{\beta}^T\mathbf{x})x_d$ can be expressed as a linear combination of terms in (3.3). Our task can be formally split into two parts — estimation of functions $g_j(\cdot)'$s with $\boldsymbol{\beta}$ given and estimation of the index coefficient $\boldsymbol{\beta}$ with given functions $\{g_j(\cdot)\}$. We also discuss how to choose the smoothing parameter h, and how to apply backward deletion to choose locally significant variables. The algorithm for practical implementation will be summarized at the end of this section.

3.2.1 Local linear estimators for $g_j(\cdot)'$s with given $\boldsymbol{\beta}$

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

$$g(\mathbf{X}) = \arg\min_{f\in\mathcal{F}(\boldsymbol{\beta})} E\left[\{Y - f(\mathbf{X})\}^2 \,\middle|\, \boldsymbol{\beta}^T\mathbf{X}\right], \tag{3.4}$$

where

$$\mathcal{F}(\boldsymbol{\beta}) = \left\{ f(\mathbf{x}) = \sum_{j=0}^{d-1} f_j(\boldsymbol{\beta}^T\mathbf{x})x_j \,\middle|\, f_0(\cdot),\, \ldots,\, f_{d-1}(\cdot) \text{ measurable, and } E\{f(\mathbf{X})\}^2 < \infty \right\}. \tag{3.5}$$

The least-squares property in (3.4) suggests the estimators $\widehat{g}_j(z) = \widehat{b}_j,\ j = 0,\, \ldots,\, d-1$, where $\left(\widehat{b}_0,\, \ldots,\, \widehat{b}_{d-1}\right)$ is the minimizer of the sum of weighted squares

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

where $w(\cdot)$ is a bounded weight function with a bounded support, which is introduced to control the boundary effect. Since $\boldsymbol{\beta}^T\mathbf{X}$ is observable when $\boldsymbol{\beta}$ is given, the estimation of $g_j(\cdot)'$s by minimizing the above sum of squares can be viewed as an extension of standard kernel regression estimation. In fact, by imposing a specified structure on the form of $g(\cdot)$, we are able to transfer the estimation of a multivariate function into the estimation of several univariate functions. Therefore, only one-dimensional kernel smoothing is involved.

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

$$\sum_{t=1}^{n}\left[Y_t - \sum_{j=0}^{d-1}\left\{b_j + c_j\left(\boldsymbol{\beta}^T\mathbf{X}_t - z\right)\right\} X_{tj}\right]^2 K_h(\boldsymbol{\beta}^T\mathbf{X}_t - z)\, w(\boldsymbol{\beta}^T\mathbf{X}_t) \tag{3.6}$$

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

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

It follows from the least squares theory that

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{\Sigma}(z)\mathcal{X}^T(z)\mathcal{W}(z)\mathcal{Y}, \quad \text{and} \quad \boldsymbol{\Sigma}(z) = \{\mathcal{X}^T(z)\mathcal{W}(z)\mathcal{X}(z)\}^{-1}, \tag{3.7}$$

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

3.2.2 Search for β-direction with $g_j(\cdot)'$s fixed

The minimization property in (3.1) suggests that we should search for $\boldsymbol{\beta}$ for which the function

$$R(\beta) = \frac{1}{n}\sum_{t=1}^{n}\left\{Y_t - \sum_{j=0}^{d-1} g_j(\boldsymbol{\beta}^T\mathbf{X}_t)X_{tj}\right\}^2 w(\boldsymbol{\beta}^T\mathbf{X}_t) \tag{3.8}$$

obtains its minimum. In fact, we will use the estimators of $\{g_j(\cdot)\}$ which cannot be estimated with reasonable accuracy at the tails. Obviously, a genuine exhaustive search will be a forbidden task even for moderate d. A simple way out is to employ one-step estimation scheme (see, for example, Bickel, 1975). The proposed method is in the spirit of one-step Newton-Raphson estimation. We anticipate that the derived estimator performs well if the initial value is reasonably good (see Fan and Chen, 1997). We outline the procedure below.

Suppose that $\widehat{\boldsymbol{\beta}}$ is the minimizer of (3.8). Then $\dot{R}\left(\widehat{\boldsymbol{\beta}}\right) = 0$, where $\dot{R}(\cdot)$ denotes the derivative of $R(\cdot)$. For any $\boldsymbol{\beta}^{(0)}$ close to $\widehat{\boldsymbol{\beta}}$, we have the approximation

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

where $\ddot{R}(\cdot)$ is the Hessian matrix of $R(\cdot)$. The above observation leads us to define the one-step iterative estimate for $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} - \left\{ \ddot{R}\left(\boldsymbol{\beta}^{(0)}\right) \right\}^{-1} \dot{R}\left(\boldsymbol{\beta}^{(0)}\right), \tag{3.9}$$

where $\boldsymbol{\beta}^{(0)}$ is the initial value. We re-scale $\boldsymbol{\beta}^{(1)}$ such that it has unit norm whose first non-vanishing element is positive. We need to evaluate all the first two partial derivatives of $R(\cdot)$. It is easy to see from (3.8) that

$$\begin{aligned}
\dot{R}(\boldsymbol{\beta}) &= -\frac{2}{n}\sum_{t=1}^{n}\left\{ Y_t - \sum_{j=0}^{d-1} g_j(\boldsymbol{\beta}^T\mathbf{X}_t)X_{tj} \right\}\left\{ \sum_{j=0}^{d-1} \dot{g}_j(\boldsymbol{\beta}^T\mathbf{X}_t)X_{tj} \right\}\mathbf{X}_t w(\boldsymbol{\beta}^T\mathbf{X}_t), \\
\ddot{R}(\boldsymbol{\beta}) &= \frac{2}{n}\sum_{t=1}^{n}\left\{ \sum_{j=0}^{d-1} \dot{g}_j(\boldsymbol{\beta}^T\mathbf{X}_t)X_{tj} \right\}^2 \mathbf{X}_t\mathbf{X}_t^T w(\boldsymbol{\beta}^T\mathbf{X}_t) \\
&\quad -\frac{2}{n}\sum_{t=1}^{n}\left\{ Y_t - \sum_{j=0}^{d-1} g_j(\boldsymbol{\beta}^T\mathbf{X}_t)X_{tj} \right\}\left\{ \sum_{j=0}^{d-1} \ddot{g}_j(\boldsymbol{\beta}^T\mathbf{X}_t)X_{tj} \right\}\mathbf{X}_t\mathbf{X}_t^T w(\boldsymbol{\beta}^T\mathbf{X}_t).
\end{aligned} \tag{3.10}$$

Note that in the above derivation, we assume that the derivative of the weight function $w(\cdot)$ is 0 for the sake of simplicity. In practice, we usually let $w(\cdot)$ be an indicator function. Further, $\boldsymbol{\beta}$ in $w(\boldsymbol{\beta}^T X_t)$ is fixed at the value of its previous iteration.

In practical implementation, the matrix $\ddot{R}(\cdot)$ could be singular or nearly so. A common technique to deal with this problem is the ridge regression (Seift and Gasser, 1996). For this purpose, we propose using the estimator (3.9) with $\ddot{R}$ replaced by $\ddot{R}_r$, which is defined by the RHS of (3.10) with $\mathbf{X}_t\mathbf{X}_t^T$ replaced by $\mathbf{X}_t\mathbf{X}_t^T + q_n\,\mathbf{I}_d$ for some positive ridge parameter q_n.

Now we briefly mention two alternative methods to estimating $\boldsymbol{\beta}$, although we don't expect that they are as efficient as the above method. The first one is based on 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 by the Monte Carol or quasi-Monte Carol methods (Fang and Wang, 1995) and then computing $R(\boldsymbol{\beta})$. Stop the algorithm if the minimum fails to decrease significantly in every 100 new draws (say). The second approach is to adapt the average derivative method of Neway and Stoker (1993) and Samarov (1993). Under the model (1.1), the direction $\boldsymbol{\beta}$ is parallel to the expected difference between gradient vector of the regression surface and $(g_1(\boldsymbol{\beta}^T\mathbf{x}), \ldots, g_{d-1}(\boldsymbol{\beta}^T\mathbf{x}), 0)^T$ and hence can be estimated by the average derivative method via iteration.

3.2.3 Bandwidth selection

We apply the generalized cross-validation (GCV) method, proposed by Wahba (1977) and Craven and Wahba(1979), to choose bandwidth h in estimation of $\{g_j(\cdot)\}$. The criterion can be described as follows. For given $\boldsymbol{\beta}$, let $\widehat{Y}_t = \sum_{j=0}^{d-1} \widehat{g}_j(\boldsymbol{\beta}^T\mathbf{X}_t)X_{tj}$. It is easy to see that all those predicted values are in fact the linear combinations of $\mathcal{Y} = (Y_1, \ldots, Y_n)^T$ with coefficients depending on $\{\mathbf{X}_t\}$ only. Namely, we can write

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

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

$$\mathrm{GCV}(h) \equiv \frac{1}{n\{1 - n^{-1}\mathrm{tr}(\mathbf{H}(h))\}^2} \sum_{t=1}^{n} \{Y_t - \widehat{Y}_t\}^2 w(\boldsymbol{\beta}^T \mathbf{X}_t), \tag{3.11}$$

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

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

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

where $\mathbf{0}$ denotes the $d \times 1$ vector with all components 0, and $\boldsymbol{\Sigma}(\cdot)$ is defined as in (3.7). The coefficient in front of Y_i on the RHS of the above expression is

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

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

3.2.4 Choosing locally significant variables

As we discussed before, the model (3.3) can be over-parametrized. Thus, it is necessary to select significant variables for each given z after the 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 its t-value, which in the meanwhile yields a new and reduced model. We select the best model according to the AIC. We opt for this rule because of its computational efficiency and simplicity.

We start with the full model

$$g(\mathbf{x}) = \sum_{j=0}^{d-1} g_j(\boldsymbol{\beta}^T \mathbf{x}) x_j. \tag{3.12}$$

For fixed $\boldsymbol{\beta}^T \mathbf{X} = z$, (3.12) could be viewed as a (local) linear regression model. The least squares estimator $\widehat{\boldsymbol{\theta}} \equiv \widehat{\boldsymbol{\theta}}(z)$ given in (3.7) entails

$$\mathrm{RSS}_d(z) = \sum_{t=1}^{n} \left[Y_t - \sum_{j=0}^{d-1} \{\widehat{g}_j(z) + \widehat{\dot{g}}_j(z)(\boldsymbol{\beta}^T \mathbf{X}_t - z)\} X_{tj} \right]^2 K_h(\boldsymbol{\beta}^T \mathbf{X}_t - z) w(\boldsymbol{\beta}^T \mathbf{X}_t). \tag{3.13}$$

The 'degree of freedom' of $\mathrm{RSS}_d(z)$ is $m(d,z) = n_z - p(d,z)$ where $n_z = \mathrm{tr}\{\mathcal{W}(z)\}$ may be regarded as the number of observations used in the local estimation and $p(d,z) = \mathrm{tr}\{\mathbf{\Sigma}(z)\mathcal{X}^T(z)\mathcal{W}^2(z)\mathcal{X}(z)\}$ as the number of local parameters. Now we define the AIC for this model as follows

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

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

$$t_k(z) = \frac{\widehat{g}_k(z)}{\sqrt{c_k(z)\mathrm{RSS}(z)/m(d,z)}} \quad \text{and} \quad t_{d+k} = \frac{\widehat{\dot{g}}_k(z)}{\sqrt{c_{d+k}(z)\mathrm{RSS}(z)/m(d,z)}}$$

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

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

Let j be the minimizer of $T_k^2(z)$ over $0 \le k < d$, we delete x_j from the full model (3.12). This leads to a model with $(d-1)$ 'linear terms'. Repeating the above process, we may define $\mathrm{AIC}_l(z)$ for all $1 \le l \le d$. The selected model should have $k-1$ 'linear terms' x_j's such that $\mathrm{AIC}_k = \min_{1\le l\le d} \mathrm{AIC}_l(z)$.

3.3 Implementation

Now we outline the algorithm as follows.

Step 1: Standardize the data set $\{\mathbf{X}_t\}$ such that it has sample mean 0 and the sample variance and covariance matrix $\mathbf{I}_d$. Specify an initial value of $\boldsymbol{\beta}$, say, the coefficient of the (global) linear fitting.

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 (3.8) differ insignificantly.

(a) For a given direction $\boldsymbol{\beta}$, we estimate the functions $g_j(\cdot)'$s by (2.8).

(b) For given $g_j(\cdot)'$s, we search direction $\boldsymbol{\beta}$ using the algorithms described in §2.2.2.

Step 3: For $k = 1, \ldots, q$, calculate $\mathrm{GCV}(h_k)$ with $\boldsymbol{\beta}$ equal its estimated value, where $\mathrm{GCV}(\cdot)$ is defined as in §2.2.3. Let $\widehat{a}_1$ and $\widehat{a}_2$ be the minimizer of $\sum_{k=1}^q \{\mathrm{GCV}(h_k) - a_0 - a_1 h_k^4 - a_2/(nh_k)\}^2$. Define the bandwidth $\widehat{h} = (\widehat{a}_2/(4\,n\,\widehat{a}_1))^{1/5}$, if $\widehat{a}_1$ and $\widehat{a}_2$ are positive; $\widehat{h} = \mathrm{argmin}_{h_k} GCV(h_k)$, otherwise.

Step 4: For $h = \widehat{h}$ selected in Step 3, repeat (a) and (b) in Step 2 until two successive values of $R(\boldsymbol{\beta})$ differ insignificantly.

Step 5: For $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ selected from Step 4, we regard (3.6) with each fixed z as a least squares problem for a linear regression model, and apply the stepwise deletion method described in §2.2.4 to select significant variables X_{tj}'s at a fixed point z.

Some additional remarks are now in order.

Remark 2. (i) The standardization in Step 1 also ensures that the sample mean of $\{\boldsymbol{\beta}^T\mathbf{X}_t\}$ is 0 and the sample variance is 1 for any unit vector $\boldsymbol{\beta}$. This effectively re-write the model (3.3) as

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

where $\widehat{\mu}$ and $\widehat{\Sigma}$ are the sample mean and sample variance, respectively. In the numerical examples in §5, we report $\widehat{\Sigma}^{-1/2}\widehat{\boldsymbol{\beta}}/||\widehat{\Sigma}^{-1/2}\widehat{\boldsymbol{\beta}}||$ as the estimated value of $\boldsymbol{\beta}$ defined in (3.3).

(ii) We may choose the weight function $w(z) = I(|z| \le 2+\delta)$ for some small $\delta \ge 0$. We estimate the functions $g_j(\cdot)'$s in Step 3 on 101 regular grids in the interval $[-1.5,\ 1.5]$ first, and then estimate the values of the functions elsewhere by linear interpolation. This will significantly reduce the computational time, especially when the sample size n is large. Finally (in Step 4), we estimate $g_j(\cdot)'$s on the interval $[-2,\ 2]$.

(iii) We use the Epanechnikov kernel in our calculation. To estimate the bandwidth $\widehat{h}$, we let $q = 15$ and $h_k = 0.2 \times 1.2^{k-1}$ in Step 3. In the case that at least one of $\widehat{a}_1$ and $\widehat{a}_2$ are non-positive, we let $\widehat{h}$ equal to the minimizer of GCV(h) over h_k for $k = 1, \ldots, q$. Note that the data have been standardized in Step 1. The selected values of bandwidth practically covers the range of 0.2 to 2.57 times of the standard deviation of the data. (If we use Gaussian kernel, we may select the range of the bandwidth between 0.1 and 1.5 times of the standard deviation.)

(iv) Note that all the estimators for $g_j(\cdot)'$s are fixed in the search for $\boldsymbol{\beta}$ in Step 2(b). To further stabilize the search, we smooth the estimates of $g_j(\cdot)'$s using a simple moving-average technique: replace an estimate on a grid point by a weighted average on its 5 nearest neighbors with weights $\{1/2,\ 1/6,\ 1/6,\ 1/12,\ 1/12\}$. The edge points should be adjusted accordingly.

(v) In the application of the one-step iterative algorithm to search for $\boldsymbol{\beta}$, we estimate the derivatives of $g_j(\cdot)'$s based on their adjusted estimates on the grid points, smoothed by a moving-average described in (iv) above. For example, we define

$$\widehat{\dot{g}}_j(z) = \{\widehat{g}_j(z_1) - \widehat{g}_j(z_2)\}\,/(z_1 - z_2), \quad j = 0, \ldots, d, \tag{3.14}$$

and

$$\widehat{\ddot{g}}_j(z) = \{\widehat{g}_j(z_1) - 2\widehat{g}_j(z_2) + \widehat{g}_j(z_3)\}\,/(z_1 - z_2)^2, \quad j = 0, \ldots, d, \tag{3.15}$$

where $z_1 > z_2 > z_3$ are three nearest neighbors of z among the 101 regular grid points (see (ii) above), and $\widehat{g}_j(z_k)$ denote the adjusted estimate at z_k. We recommend to iterate equation (3.9) a few times (instead of just once) to speed up the convergence, because a reasonably good initial value is required to ensure the good performance of the iterative estimator (see Fan and Chen, 1997).

4 Varying-coefficient linear models with two indices

A natural extension of the method discussed in the previous section is to use varying-coefficient functions with more than one indices. In this section, we consider the models with two indices but one of them is known. We assume knowing one index in order to keep computation practically feasible.

To simplify the notation, 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 (3.1)), by a function of the form

$$g(\mathbf{x}, v) = \sum_{j=0}^{d-1} g_j(\boldsymbol{\beta}^T \mathbf{x}, v) x_j, \tag{4.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^T$ is a $d \times 1$ unknown unit vector. Similar to Theorem 1(ii), it can be proved that under some mild conditions on $g(\mathbf{x}, v)$, the expression on the RHS of (4.1) is unique if the first non-zero β_k is positive and $\beta_d \neq 0$. Let $\{(\mathbf{X}_t, V_t, Y_t); 1 \leq t \leq n\}$ be observations from a strictly stationary process, and $(\mathbf{X}_t, V_t, Y_t)$ has the same distribution as $(\mathbf{X}, V, Y)$.

The estimation for $g(\mathbf{x}, v)$ can be carried out in the similar manner as in one index case (see §3.3). We outline the algorithm below briefly.

Step 1: Standardize the data set $\{\mathbf{X}_t\}$ such that it has sample mean 0 and the sample variance and 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 (4.2) differ by insignificantly.

(a) For a given direction $\boldsymbol{\beta}$, we estimate the functions $g_j(\cdot, \cdot)'$s in terms of local linear regression.

(b) For given $g_j(\cdot, \cdot)'$s, we search direction $\boldsymbol{\beta}$ using a one-step iteration algorithms.

Step 3: For $k = 1, \ldots, q$, calculate GCV(h_k) with $\boldsymbol{\beta}$ equal its estimated value, where GCV($\cdot$) is as defined in §3.2.3. Let $\widehat{a}_1$ and $\widehat{a}_2$ be the minimizer of $\sum_{k=1}^{q} \{\text{GCV}(h_k) - a_0 - a_1 h_k^4 - a_2/(nh_k^2)\}^2$. Define the optimal bandwidth $\widehat{h} \equiv (\widehat{a}_2/(2\, n\, \widehat{a}_1))^{1/6}$.

Step 4: For $h = \widehat{h}$ selected in Step 3, repeat (a) and (b) in Step 2 until two successive values of $R(\boldsymbol{\beta})$ differ by a small amount.

Step 5: For $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$ selected from Step 4, select local significant variables for each given point (z, v).

Remark 3. (i) In Step 2(a) above, The local linear regression estimation leads to the problem of minimizing the weighted sum of squares

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

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 the simplicity of implementation. The derived estimators are $\widehat{g}_j(z,v) = \widehat{a}_j$, $\widehat{\dot{g}}_j(z,v) = \widehat{b}_j$ and $\widehat{\dot{g}}_{j,v}(z,v) = \widehat{c}_j$ for $j = 0, \ldots, d-1$, where $\dot{g}_j(z,v) = \partial g_j(z,v)/\partial z$ and $\dot{g}_{j,v}(z,v) = \partial g_j(z,v)/\partial v$.

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

$$R(\boldsymbol{\beta}) = \frac{1}{n}\sum_{t=1}^{n}\left\{Y_t - \sum_{j=0}^{d-1} g_j(\boldsymbol{\beta}^T\mathbf{X}_t, V_t)X_{tj}\right\}^2 w(\boldsymbol{\beta}^T\mathbf{X}_t, V_t). \tag{4.2}$$

A one-step iterative algorithm may be constructed for the purpose in the similar manner as in the case with one index only; see §3.2.2 above. The required estimates for the second derivatives of $g_j(z,v)$ may be obtained via a partially local quadratic regression.

(iii) In Step 3, the estimated $g(\mathbf{x}, v)$ is linear in the variable $\{Y_t\}$ (for a given $\boldsymbol{\beta}$). Thus, the generalized cross-validation method outlined in §3.2.3 continues to apply.

(iv) Further, locally around the given indices $\boldsymbol{\beta}^T\mathbf{x}$ and v, model (4.1) is approximately a linear model. Thus, the local variable selection technique outlined in §3.2.4 is still applicable in Step 5 above.

5 Numerical properties

We always use the Epanechnikov kernel and its productive form for the bivariate kernel, in our calculation. We always use the one-step iterative algorithm described in §3.2.2 to estimate the index $\boldsymbol{\beta}$. In fact, we iterate ridge version of equation (3.9) two to four times to speed up the convergence. We stop the search in Step 2 when either the two successive values of $R(\boldsymbol{\beta})$ differ less than 0.001, or the number of replications of (a) and (b) in Step 2 exceeds 30. We set initially the ridge parameter $q_n = 0.001\, n^{-1/2}$ and keep doubling its value until the $\ddot{R}_r(\cdot)$ is no longer *ill-conditioned* with respect to the precision of computers.

5.1 Simulation

We demonstrate the finite sample performance of the varying-coefficient model with one index through Examples 1 and 2, and with two indices in Example 3. Examples 1 and 3 are nonlinear regression models with independent observations while Example 2 is a nonlinear time series model.

We use absolute inner product $|\boldsymbol{\beta}^T\widehat{\boldsymbol{\beta}}|$ to measure the goodness of the estimated direction $\widehat{\boldsymbol{\beta}}$. Their inner product represents the cosine of the angles between the two directions. For both Examples 1 and 2 below, we evaluate the performance of the estimator in terms of the mean absolute deviation error

$$\mathcal{E}_{\text{MAD}} = \frac{1}{101\, d}\sum_{j=0}^{d-1}\sum_{k=1}^{101}|\widehat{g}_j(z_k) - g_j(z_k)|, \tag{5.1}$$

where z_k, $k = 1, \ldots, 101$ are the regular grid points on [-2, 2] after the standardization. For example 3, $\mathcal{E}_{\rm MAD}$ is calculated on the observed values instead of regular grid points as in the above expression.

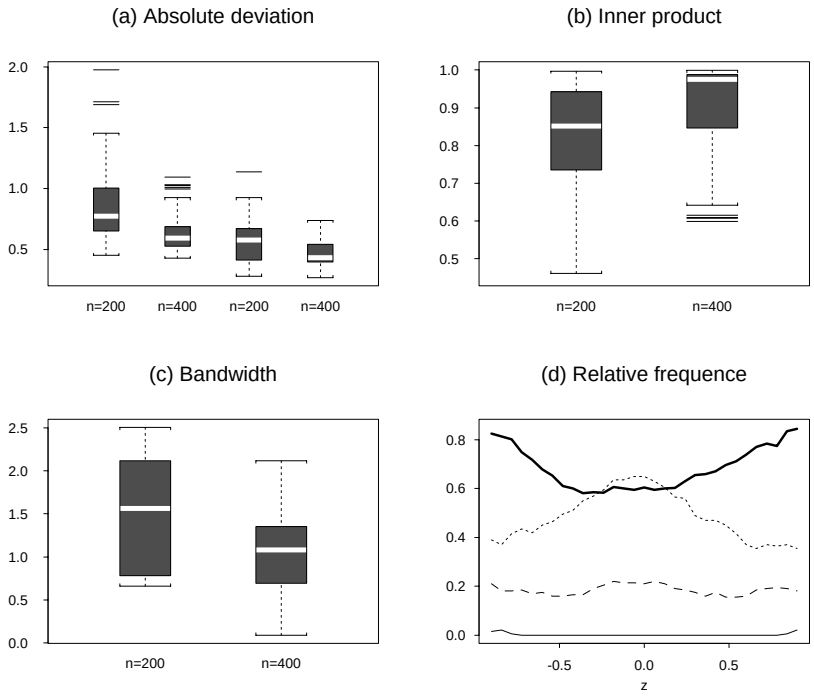


Figure 1: *Simulation results for Example 1. (a) The boxplots of the mean absolute deviation error* $\mathcal{E}_{\rm MAD}$. *The two panels on the left are based on the estimated* $\boldsymbol{\beta}$, *and the two panels on the right are based on the true* $\boldsymbol{\beta}$. *(b) The boxplots of the absolute inner product* $|\boldsymbol{\beta}^T\widehat{\boldsymbol{\beta}}|$. *(c) The boxplots of selected bandwidths. (d) The plots of the relative frequencies of deletion of locally insignificant terms at* z *against* z: *thin solid line — for the intercept; dotted line — for* X_{t1}, *thick solid line — for* X_{t2}, *and dashed line — for* X_{t3}.

Example 1. Let us consider the regression model

$$\begin{aligned} Y_t &= 3\exp\{-Z_t^2\} + 0.8 Z_t X_{t1} + 1.5\sin(\pi Z_t) X_{t3} + \varepsilon_t, \\ &\text{with } \; Z_t = \frac{1}{3}(X_{t1} + 2X_{t2} + 2X_{t4}), \end{aligned} \tag{5.2}$$

where $\mathbf{X}_t \equiv (X_{t1}, \ldots, X_{t4})^T$, for $t \geq 1$, are independent random vector uniformly distributed on $[-1, 1]^4$, and $\{\varepsilon_t\}$ is a sequence of independent standard normal random variables. It is easy to see that the regression function in the above model is in the form of (3.3) with $d = 4$, $\boldsymbol{\beta} = \frac{1}{3}(1, 2, 0, 2)^T$, and the coefficient functions

$$g_0(z) = 3e^{-z^2}, \quad g_1(z) = 0.8z, \quad g_2(z) \equiv 0, \;\text{ and }\; g_3(z) = 1.5\sin(\pi z).$$

We now apply the algorithm described in §3.2.5 to estimate parameters in this model. We conduct two simulations with sample size 200 and 400 respectively, each with 200 replications. The CPU time for each replication with sample size 400 is under 70 seconds in average in a Sun Ultra-1

143MHz Workstation, and is about 18 seconds in a Pentium II 350MHz PC (Linux). The results are summarized in Fig. 1. Fig. 1(a) displays the boxplots of the mean absolute deviation errors. We also plot the mean absolute deviation errors obtained 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 $\widehat{\boldsymbol{\beta}}$ derived from the one-step iterative algorithm is close to the true $\boldsymbol{\beta}$ with high frequencies in the simulation replications. The average iteration time in search for $\boldsymbol{\beta}$ is 14.43 for $n = 400$ and 18.25 for $n = 200$. Most outliers in Fig. 1(a) and Fig. 1(b) correspond to the cases where the search for $\boldsymbol{\beta}$ did not converge within 30 iterations, which is the upper limit set in the simulation. Fig. 1(c) indicates that the proposed bandwidth selector described in §2.2.3 seems quite stable. We also applied the method in §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 of the standard deviations of $\boldsymbol{\beta}^T\mathbf{X}$. The relative frequencies of deletion are depicted in Fig. 1(d). There is overwhelming evidence to include the 'intercept' $g_0(z) = 3e^{-z^2}$ 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 to keep the term X_{t3} in the model. Note that the term X_{t2} is less significant, the magnitude of its 'coefficient' $g_1(z) = 0.8z$ being smaller than those of both $g_0(z)$ and $g_3(z)$.

Fig. 2 presents a typical example of the estimated coefficient functions. The curves are plotted on the range from -1.5 to 1.5 times of the standard deviation of $\boldsymbol{\beta}^T\mathbf{X}$. The typical example was selected in such a way that the corresponding $\mathcal{E}_{\mathrm{MAD}}$ is equal to its median among the 200 replicated simulations with the sample size $n = 400$. For this example, the selected bandwidth is 0.597, $\boldsymbol{\beta}^T\widehat{\boldsymbol{\beta}} = 0.946$. For the sake of comparison, we also plot the estimated functions obtained using the true index $\boldsymbol{\beta}$. The deficiency due to unknown $\boldsymbol{\beta}$ is almost negligible once $\widehat{\boldsymbol{\beta}}$ is reasonably accurate. Note that the biases of estimators for the coefficient functions $g_0(\cdot)$, $g_1(\cdot)$ and $g_2(\cdot)$ (but not necessarily for $g(\cdot)$) are large near to boundaries. We believe that this is due to the collinearity of variables $X_1, \cdots, X_4$ and small effective local sample size near the tails. The coefficient functions are not so easily identified locally in those areas. However, there is no evidence that this problem will distort the estimation for the target function $g(\mathbf{x})$.

Example 2. We now consider a time series model

$$Y_t = -Y_{t-2}\exp(-Y_{t-2}^2/2) + \frac{1}{1+Y_{t-2}^2}\cos(1.5Y_{t-2})Y_{t-1} + \varepsilon_t, \tag{5.3}$$

where $\{\varepsilon_t\}$ is a sequence of independent normal random variables with mean 0 and variance 0.25. If we regard $\mathbf{X}_t \equiv (Y_{t-1}, Y_{t-2})^T$ as the regressor, (5.3) is of the form of model (3.3) with $d = 2$, $\boldsymbol{\beta} = (0, 1)$, and

$$g_0(z) = -z\exp(-z^2/2), \qquad g_1(z) = \cos(1.5z)/(1+z^2).$$

To illustrate the application of our algorithm to this model, we conduct two simulations with sample size 200 and 400 respectively with 200 replications. For each replication, we predict the 50 post-sample points and compare them with the true values. One realization with sample size 400 lasts less than 15 seconds in average on a Sun Ultra-1 143MHz Workstation, and less than 4 seconds on

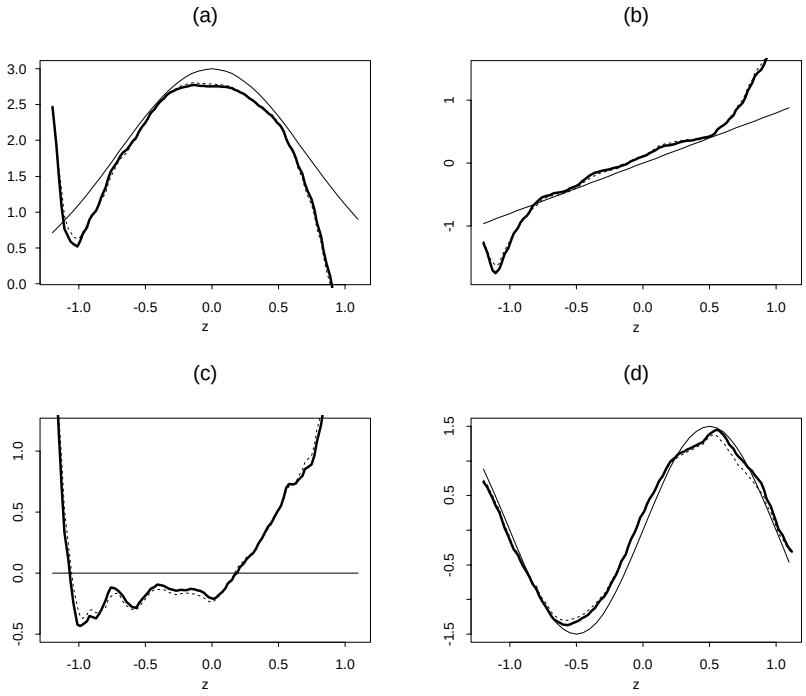


Figure 2: *Simulation results for Example 1 ($n = 400$). The plot of estimated coefficient functions (thick line), true functions (thin line), and estimated functions with true index $\boldsymbol{\beta}$ (dotted line). (a) $g_0(z) = 3e^{-z^2}$; (b) $g_1(z) = 0.8z$; (c) $g_2(z) = 0$; (d) $g_3(z) = 1.5\sin(\pi z)$.*

a Pentium II 350MHz PC. The results are summarized in Fig. 3. Fig. 3(a) displays the boxplots of the mean absolute deviation errors. For sample size $n = 400$, the mediums of $\mathcal{E}_{\mathrm{MAD}}$ with estimated and true $\boldsymbol{\beta}$ are about the same, although the distribution of $\mathcal{E}_{\mathrm{MAD}}$ with $\widehat{\boldsymbol{\beta}}$ has a long tail on the right. Fig. 3(b) shows that the estimator $\widehat{\boldsymbol{\beta}}$ derived from the one-step iterative algorithm is close to the true $\boldsymbol{\beta}$ with high frequencies in the simulation replications. The average iteration time in search for $\boldsymbol{\beta}$ is 7.80 for $n = 400$ and 17.62 for $n = 200$. In fact, the search did not converge within 30 iterations for 21 out of 200 replications with $n = 200$, and for one out of 200 replications with $n = 400$. Fig. 3(c) is the boxplot of the selected bandwidths.

We also compared prediction performance of various models in the simulation with the sample size $n = 400$. For each of 200 realizations, we predict 50 post-sample points from four different models, namely the fitted varying-coefficient models with true and estimated $\boldsymbol{\beta}$, a purely nonparametric model based on local linear regression of Y_t on (Y_{t-1}, Y_{t-2}) with the bandwidth selected by the GCV-criterion, and a linear autoregressive model with the order (≥ 2) determined by AIC. In our simulation, AIC always selected order 2 in the 200 replications. Fig. 3(d) presents the boxplots of the average absolute predictive errors. The varying-coefficient models with true and estimated $\boldsymbol{\beta}$ are the two best predictors since they specify the correct form of the true model (see Fig. 3(d)). The median of the predictive errors from the nonparametric model based on local linear regression is about the same as that from the varying-coefficient model. But the variance is much larger. The linear autoregressive model performs poorly in this example since the data are generated from a

very nonlinear model.

Fig.4 presents a typical example of the estimated coefficient functions with the sample size $n = 400$. The curves are plotted on the range from -1.5 to 1.5 times of the standard deviation of $\boldsymbol{\beta}^T\mathbf{X}$. For the case with $n = 400$, the selected bandwidth is 0.781, and $\boldsymbol{\beta}^T\widehat{\boldsymbol{\beta}} = 0.999$. (The median of $\boldsymbol{\beta}^T\widehat{\boldsymbol{\beta}}$ in the simulation of 200 replications with $n = 400$ is 0.999.)

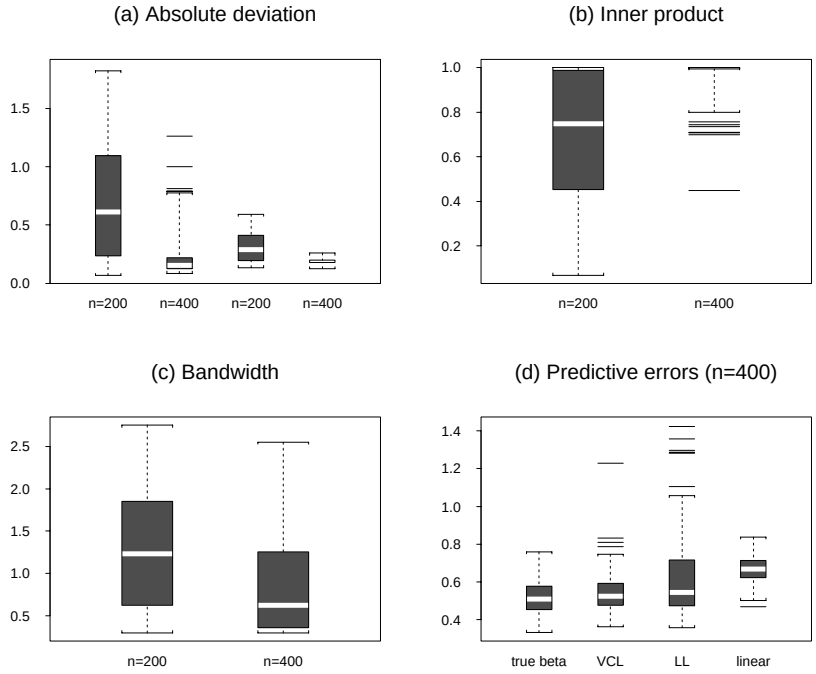


Figure 3: *Simulation results for Example 2. The boxplots of (a) the mean absolute deviation error* $\mathcal{E}_{\mathrm{MAD}}$ *(the two panels on the left are based on* $\widehat{\boldsymbol{\beta}}$*, and the two panels on the right are based on the true* $\boldsymbol{\beta}$*), (b) the absolute inner product* $|\boldsymbol{\beta}^T\widehat{\boldsymbol{\beta}}|$*, (c) the selected bandwidths, and (d) the average absolute predictive errors of the varying-coefficient models with true* $\boldsymbol{\beta}$ *and* $\widehat{\boldsymbol{\beta}}$*, nonparametric model based on local linear regression, and linear AR-model determined by AIC (from left to right).*

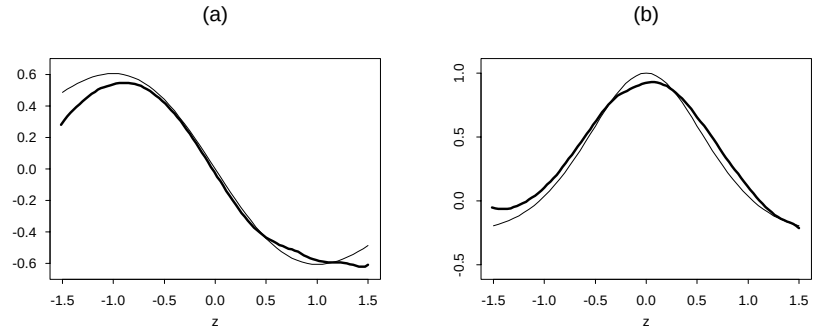


Figure 4: *Simulation results for Example 2. The plot of estimated coefficient functions (thick line), true functions (thin line). (a)* $g_0(z) = -ze^{-z^2/2}$*; (b)* $g_1(z) = \cos(1.5z)/(1+z^2)$*. The sample size* $n = 400$*.*

Example 3. We consider the regression model

$$\begin{aligned} Y_t &= 3\exp(-Z_t^2 + X_{t1}) + (Z_t + X_{t1}^2)X_{t1} - \log(Z_t^2 + X_{t1}^2)X_{t2} + 1.5\sin(\pi Z_t + X_{t1})X_{t3} + \varepsilon_t, \\ &\quad \text{with } Z_t = \frac{1}{2}(X_{t1} + X_{t2} + X_{t3} + X_{t4}), \end{aligned}$$

where $\{X_{t1}, \ldots, X_{t4}\}$ and $\{\varepsilon_t\}$ are the same as in Example 1. Obviously, the regression function in the above model is of the form (4.1) with $d = 4$, $\boldsymbol{\beta} = \frac{1}{2}(1, 1, 1, 1)^T$, $V_t = X_{t1}$ and the two-dimensional coefficient functions

$$g_0(z, v) = 3e^{-z^2+v}, \quad g_1(z, v) = z + v^2, \quad g_2(z, v) = -\log(z^2 + v^2), \quad g_3(z, v) = 1.5\sin(\pi z + v),$$

which are plotted in Fig. 5. Assuming the direction of $V_t = X_{t1}$ is given, we now apply the algorithm described in §3.4 to estimate the coefficient functions. We conduct three simulations with sample size 200, 400 and 600 respectively, each with 100 replications. The CUP time for each realization, in a Sun Ultra-10 300MHz Workstation, is about 18 seconds for $n = 200$, 1 minute and 20 seconds for $n = 400$ and 3 minutes and 10 seconds for $n = 600$. Fig. 6(a) shows that the mean absolute deviation error decreases when n increases. For the sake of comparison, we also present the mean absolute deviation error of the estimator based on true value of $\boldsymbol{\beta}$. Fig. 6(b) displays the boxplots of the absolute inner product $|\boldsymbol{\beta}^T\widehat{\boldsymbol{\beta}}|$, which indicates that the one-step iteration algorithm described in §3.2 works reasonably well. The boxplots of bandwidths selected by the GCV-method stated in §3.3 are depicted in Fig. 6(c).

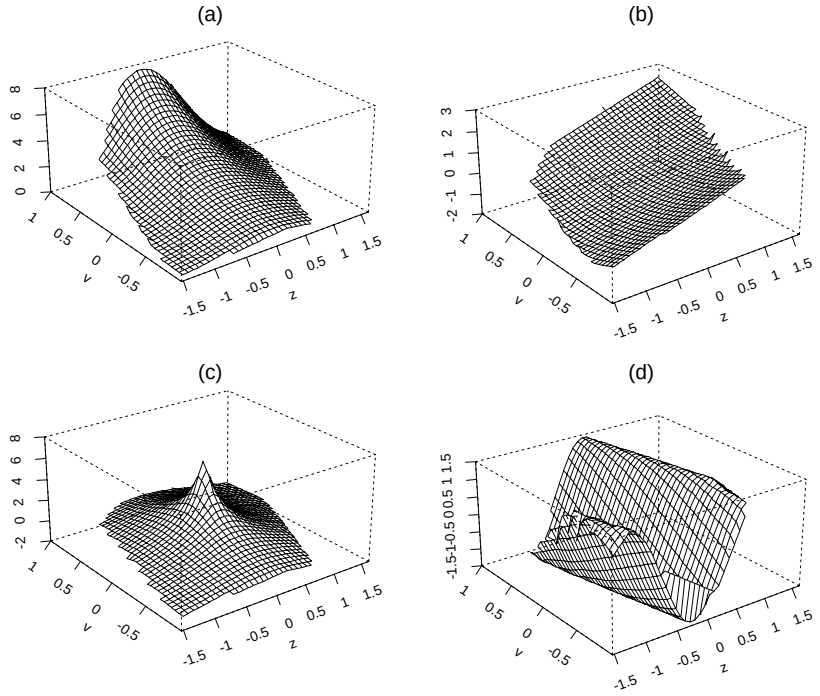


Figure 5: *The coefficient functions of Example 3. (a)* $g_0(z, v) = 3e^{-z^2+v}$, *(b)* $g_1(z, v) = z + v^2$, *(c)* $g_2(z, v) = -log(z^2 + v^2)$, *and (d)* $g_3(z, v) = 1.5\sin(\pi z + v)$.

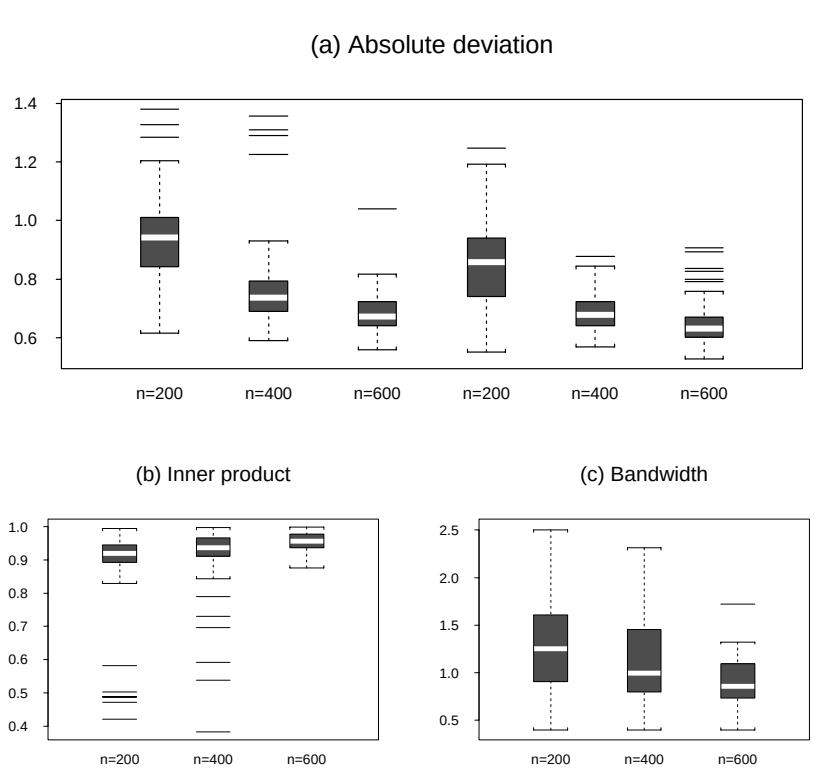


Figure 6: *The simulation results for Example 3. The boxplots of (a) the mean absolute deviation error* $\mathcal{E}_{\mathrm{MAD}}$*, (b) the absolute inner product* $|\boldsymbol{\beta}^T\widehat{\boldsymbol{\beta}}|$*, and (c) the selected bandwidths. The three panels on the left in (a) are based on the estimated* $\boldsymbol{\beta}$*, and the three panels on the right are based on the true* $\boldsymbol{\beta}$*. The three panels on the left in (a) are based on the estimated* $\boldsymbol{\beta}$*, and the three on the right are based on the true* $\boldsymbol{\beta}$*.*

5.2 Real data examples

Example 4. 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. 7 indicates the 82 posts where furs were collected. Biological evidence suggests that mink is a key predator on muskrat (Errington, 1961, 1963). Fig. 7(b) plots the time series of the mink and the muskrat (on the natural logarithmic scale) from 8 posts selected randomly among the 82 posts. Most series exhibit cycles with a period of around 10 years. There exists a clear synchrony between the fluctuations of the two species with a delay of about one or two years. Since there is a general lack of data on both prey and predator from the same area and over the same time period, this data set offers a unique opportunity for quantitative analysis aiming at a deeper understanding of the interaction between prey (*i.e.* muskrat) and predator (*i.e.* mink). As a starting point, we introduce an ecological model to describe the mink-muskrat interaction. Based on the food chain interaction model of May (1981), Stenseth *et al.* (1997) proposed a deterministic

(a) The 82 trapping posts for the mink and the muskrat in Canada

(*Put Fig. 7(a) here*)

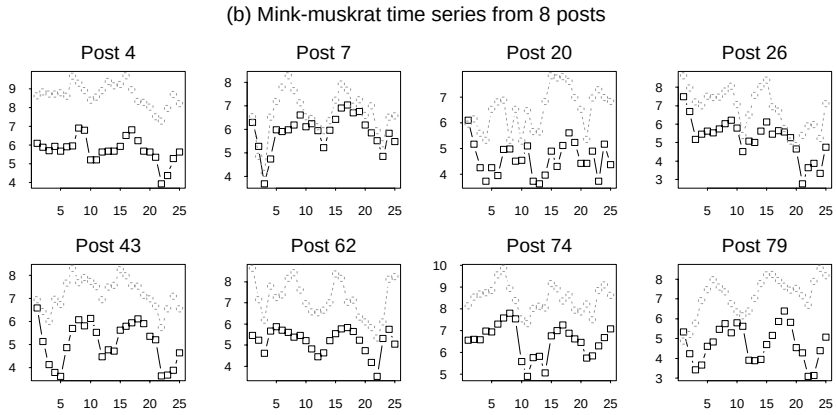


Figure 7: *(a) A map of 82 posts for the mink and the muskrat in Canada in 1925 – 1949. (b) The time series plots of the mink and the muskrat data from 8 randomly selected posts. Solid lines — mink; dashed lines — muskrats.*

model to describe the predator-prey interaction, namely

$$\begin{cases} X_{t+1} - X_t &= a_0(\theta_t) - a_1(\theta_t)X_t - a_2(\theta_t)Y_t, \\ Y_{t+1} - Y_t &= b_0(\theta_t) - b_1(\theta_t)Y_t + b_2(\theta_t)X_t, \end{cases} \tag{5.4}$$

where X_t and Y_t denote the population abundances, on a natural logarithmic scale, of muskrat and 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 nonlinear effect due to, among others, the different hunting/escaping behavior or the different reproduction rates of animals at different stages of population fluctuation (Stenseth *et al.*, 1999). Biologically speaking, $a_1(\theta_t)$ and $b_1(\theta_t)$ reflect the within species regulation whereas $a_2(\theta_t)$ and $b_2(\theta_t)$ reflect the food chain interaction between the two species, and $a_0(\theta_t)$ and $b_0(\theta_t)$ are the intrinsic rates of changes. A simple option which facilitates statistical data analysis is to use a threshold variable to define the regime effect which switches between two regimes. The model implied, with added random noise, could have the form

$$\begin{cases} X_{t+1} &= (a_{10} + a_{11}X_t + a_{12}Y_t)I(X_t \le r_1) + (a_{20} + a_{21}X_t + a_{22}Y_t)I(X_t > r_1) + \varepsilon_{1,t+1}, \\ Y_{t+1} &= (b_{10} + b_{11}Y_t + b_{12}X_t)I(X_t \le r_2) + (b_{20} + b_{21}Y_t + b_{22}X_t)I(X_t > r_2) + \varepsilon_{2,t+1}, \end{cases} \tag{5.5}$$

where we choose muskrat variable X_t as the threshold variable. It is easy to see from (5.4) that both a_{12} and a_{22} should be non-positive, and both b_{12} and b_{22} should be non-negative. The model (5.5) assumes the populations muskrat and mink are piecewise linear functions of their immediate lagged values. Note that each time series has only 25 points, which imposes intrinsic difficulties for statistical data analysis even with simple nonlinear models such as (5.5). Yao *et al.* (1998) conducted some statistical tests on the common structure for each pair among those 82 regions and

further 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. 7; the western area consisting of the 30 posts on the left in Fig. 7 (*i.e.* post 17 and those on its left); and the central area consisting of the remaining 43 posts in the middle. Yao *et al.* (1998) fitted model (5.5) to each of pooled data sets and reported some interesting and ecologically interpretable findings.

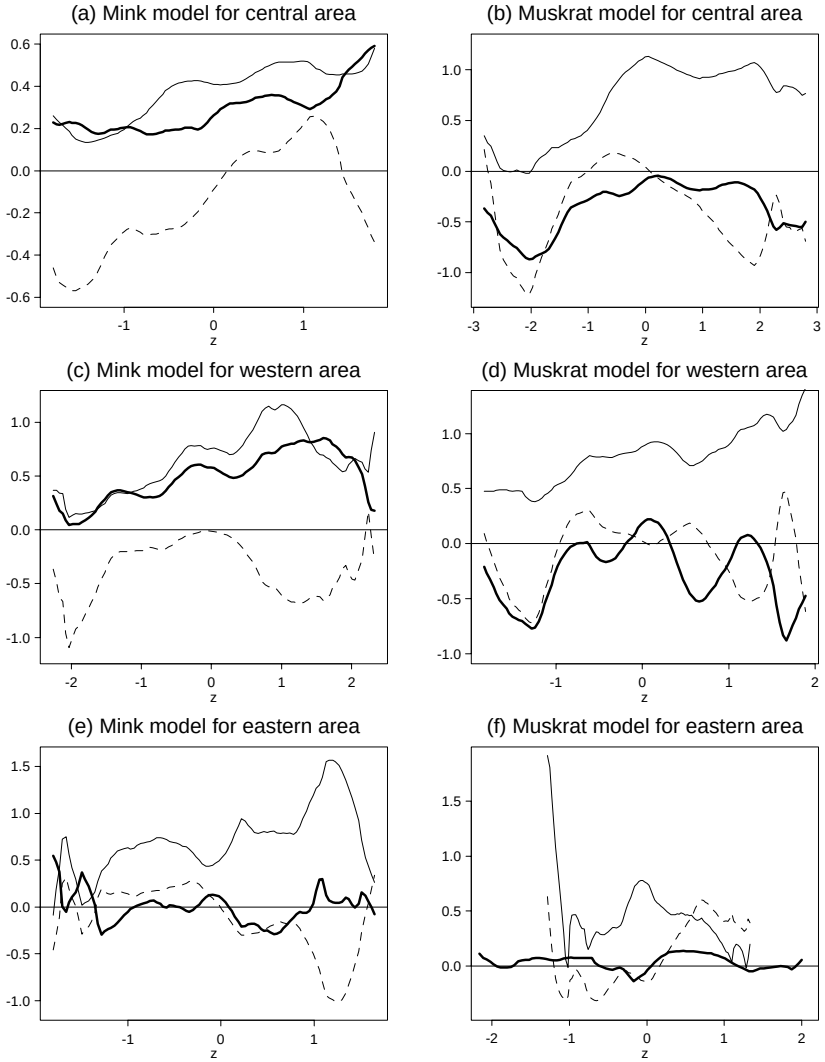


Figure 8: *Estimated coefficient functions for Canadian mink-muskrat data. (a), (c) & (d): thick solid lines — $g_x(\cdot)$; solid lines — $g_y(\cdot)$; dashed lines — $g_0(\cdot)$. (b), (d) & (f): thick solid lines — $f_y(\cdot)$; solid lines — $f_x(\cdot)$; dashed lines — $f_0(\cdot)$.*

Clearly, the model (5.5) simplifies the nonlinear interaction into two states (for each of muskrat or mink models) with a prescribed threshold variable X_t. Note that $X_t - X_{t-1}$ is the muskrat population growth rate. It is biologically interesting to find out which would be an appropriate 'threshold' variable to define the regime effect among X_t, $X_t - X_{t-1}$, Y_t and $Y_t - Y_{t-1}$. With the new technique proposed in this paper, we fit the pooled data for each of the three areas with the

model

$$\begin{cases} X_{t+1} &= f_0(Z_t) + f_1(Z_t)Y_{t-1} + f_2(Z_t)Y_t + f_3(Z_t)X_{t-1} + \varepsilon_{1,t+1}, \\ Y_{t+1} &= g_0(Z_t) + g_1(Z_t)Y_{t-1} + g_2(Z_t)Y_t + g_3(Z_t)X_{t-1} + \varepsilon_{2,t+1}, \end{cases} \tag{5.6}$$

where $Z_t = \beta_1 Y_{t-1} + \beta_2 Y_t + \beta_3 X_{t-1} + \beta_4 X_t$ with $\boldsymbol{\beta} \equiv (\beta_1, \beta_2, \beta_3, \beta_4)^T$ selected by data. Comparing with (5.4) and (5.5), we include further lagged values X_{t-1} and Y_{t-1} into the above model. We will apply the local variable selection technique in §2.2.4 to detect any redundant variables at 31 regular grid points over the range from -1.5 to 1.5 times of standard deviation of Z_t. To eliminate the effect of different sampling weights in different regions and for different species, we first standardized the mink series and muskrat series separately for each post. Since there are some missing values in the data from post 15, we exclude it from our analysis. The sample size for eastern, central and western areas are therefore 207, 989 and 667 respectively. We denote R_{MSE} the ratio of the mean squares errors from the fitted model over the sample variance of the variable to be fitted.

First, we use the second equation of (5.6) to model mink population dynamics in the central area. The selected $\boldsymbol{\beta}$ is $(0.424, 0.320, 0.432, 0.733)^T$, the selected bandwidth is 0.415, and $R_{MSE} = 0.449$. The local variable selection indicates that X_{t-1} is the least significant variable over all, for it is significant at only 7 out of 31 grid points; see §2.2.4. By leaving it out, we reduce to the model

$$Y_{t+1} = g_0(Z_t) + g_y(Z_t)Y_t + g_x(Z_t)X_t + \varepsilon_{2,t+1}, \tag{5.7}$$

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

$$Z_t = (0.540Y_t - 0.634Y_{t-1}) + 0.553X_t, \tag{5.8}$$

which suggests that the nonlinearity is dominated by the growth rate of mink and the population of muskrat in the previous year. The estimated coefficient functions are plotted in Fig. 8(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 indicates a natural reproduction process of mink population. Both $g_y(\cdot)$ and $g_x(\cdot)$ are approximately increasing with respect to the sum of growth rate of mink and population of muskrat; see (5.8). But the 'intercept' $g_0(\cdot)$ drops sharply after Z_t reaches a threshold around 1. This *might* be related to the fact that mink population could suffer from its over-sized growth rate due to the competition for food and living environment among mink themselves. All the terms in the model (5.7) are significant in most places; the number of significant grid points for 'intercept', Y_t and X_t are 21, 31 and 26 (out of 31 in total). The selected bandwidth is 0.597 and $R_{MSE} = 0.461$.

Fitting the first equation of (5.6) to muskrat dynamics in the central area, we obtained $\widehat{\boldsymbol{\beta}} = (0.632, 0.308, 0.210, 0.680)^T$, $\widehat{h} = 0.346$ and $R_{MSE} = 0.518$. The overall least significant variable is Y_{t-1} which is only significant in 9 out of 31 grid points. By leave it out, the model is reduced to

$$X_{t+1} = f_0(Z_t) + f_y(Z_t)Y_t + f_x(Z_t)X_t + \varepsilon_{1,t+1}, \tag{5.9}$$

where $Z_t = \beta_1 Y_t + \beta_2 X_t + \beta_3 X_{t-1}$. The results from fitting are as follows: $Z_t = 0.542Y_t + 0.720X_t + 0.435X_{t-1}$, $\widehat{h} = 0.498$ and $R_{MSE} = 0.559$. All the terms in the model (5.9) are significant at least

at 25 grid points (out of 31). The estimated coefficient functions are plotted in Fig. 8(b). The coefficient function $f_y(\cdot)$ is always negative, which reflects the fact that mink is the key predator of muskrat in this core of the boreal forest in Canada. The coefficient $f_x(\cdot)$ is positive, as well-expected.

We repeated the above exercise for pooled data in the western area, and obtained similar results. In fact, the model (5.7) appears appropriate for mink dynamics with $Z_t = 0.469Y_t + 0.723X_t + 0.507Y_{t-1}$, $R_{MSE} = 0.446$, $\widehat{h} = 0.415$, and the estimated coefficient functions plotted in Fig. 8(c). The model (5.9) appears appropriate for muskrat dynamics with $Z_t = 0.419Y_t + 0.708X_t + 0.569X_{t-1}$, $\widehat{h} = 0.415$, $R_{MSE} = 0.416$, and the estimated coefficient functions plotted in Fig. 8(d).

Finally, we fit the data in the eastern area. The results are radically different from those of the central and the west stated above. To fit the mink dynamics with the second equation of (5.6), we discovered that both X_t and X_{t-1} are significant only in small portions of the sample space. After leaving out X_{t-1}, the fitting with the model (5.7) give $Z_t = 0.173Y_t - 0.394X_t + 0.901Y_{t-1}$, $\widehat{h} = 0.597$ and $R_{MSE} = 0.681$. The local variable selection indicates that out of 31 grid points, the 'intercept', Y_t and X_t are significant at 15, 31 and 4 points respectively. There is clear auto-dependence in mink series $\{Y_t\}$ while muskrat data $\{X_t\}$ carry little information about mink. The estimated coefficients, depicted in Fig. 8(e), are consistent with the above observations. 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 the model (5.9) is not significant at all the 31 grid points tested. The estimated coefficient function $f_y(\cdot)$ is plotted as the thick curve in Fig. 8(f), which is always close to 0. We fit the data with a further simplified model

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

The results are as follows: $Z_t = 0.667X_t - 0.745X_{t-1}$, $\widehat{h} = 0.498$ and $R_{MSE} = 0.584$. The estimated coefficient functions are superimposed on Fig. 8(f). Note the different ranges of z-values are due to different Z_t's are used in the above model and the model (5.9).

In summary, we have facilitated the data analysis of the biological food chain interaction model of Stenseth *et al.* (1997) by portraying the nonlinearity through varying-coefficient functions. 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 a strong evidence of predator-prey interaction between mink and muskrat in the central and western areas. However, no evidence for such an interaction exists in the eastern area. In light of what is known in 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. It has been also observed that foxes have a much more pronounced influence on the entire system of this area (Elton, 1942).

Example 5. This example concerns the daily closing bid prices of the pound sterling in terms of US dollar from 2 January 1974 to 30 December 1983, which forms a time series of length 2510. The previous analysis of this 'particularly difficult' data set can be found in Gallant, Hsieh and Tauchen (1991) and the references within. Let X_t be the exchange rate on the t-th day. We model the

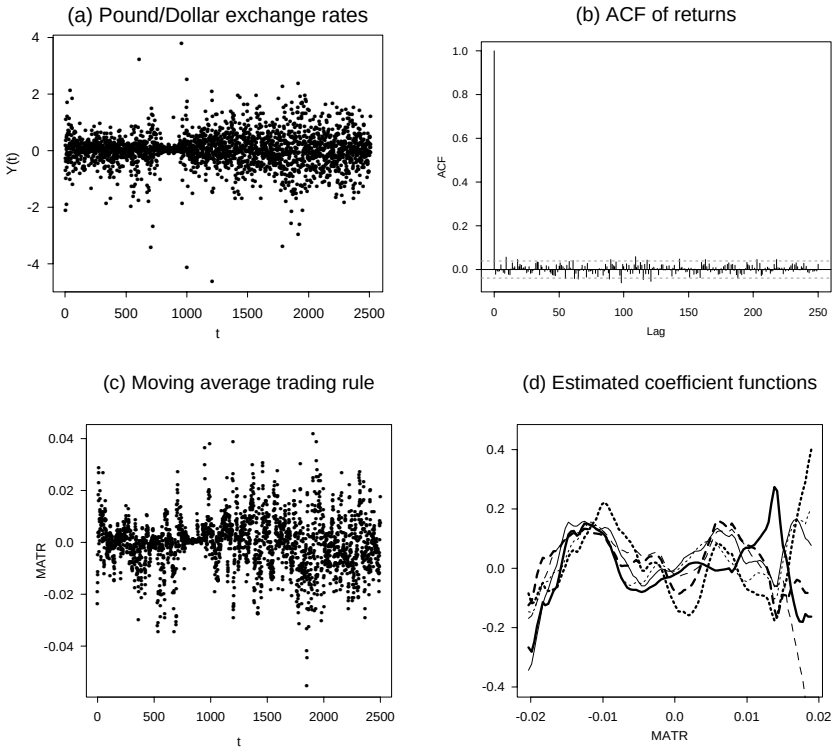


Figure 9: *(a) The plot of Pound/Dollar exchange rate return series* $\{Y_t\}$. *(b) The autoregressive function of* $\{Y_t\}$. *(c) The plot of* $\{U_t = Y_t/(\sum_{i=0}^{9} Y_{t-i}/10)\}$. *(d) The estimated coefficient functions of model (5.10) with* $Z_t = U_{t-1}$ *and* $m = 5$. *Thick solid lines –* $g_0(\cdot)$, *thick dotted lines —* $g_1(\cdot)$, *thick dashed lines —* $g_2(\cdot)$, *solid lines —* $g_3(\cdot)$, *dotted lines —* $g_4(\cdot)$, *dashed lines —* $g_5(\cdot)$.

return series $\{Y_t = 100\log(X_t/X_{t-1})\}$ plotted in Fig. 9(a) using the techniques developed in this paper. It is well-known that the classical financial theory assumes that time series $\{Y_t\}$ is typically a martingale difference process and Y_t is unpredictable. Fig. 9(b) shows that there exists almost no significant autocorrelation in the series $\{Y_t\}$.

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

$$g_0(Z_t) + \sum_{i=1}^{m} g_j(Z_t) Y_{t-i}, \tag{5.10}$$

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 technical trading rule* (MATR) in finance, and $U_{t-1}+1$ is the ratio of exchange rate at the time $t-1$ to the average rate over 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 detailed discussion of the MATR, we refer to the papers by 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 the 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 the sake of simplicity). It is more relevant than the conventional mean squared predictive errors or average absolute predictive errors for evaluating forecasting for financial data; see Hong and Lee (1999). Under this criterion, we need to predict the direction of market movement rather than its quantity. For the MATR, the mean trading return is defined as

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

Let $\widehat{Y}_t$ be defined as the estimated function given in (5.10). The mean trading return for the forecasting based on our varying-coefficient modeling is defined as

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

On the other hand, 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

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

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

Now we are ready to proceed with calculation. First, we let $m = 5$ and $L = 10$ in (5.10), *i.e.* we use one week data in the past as 'regressors' in the model and define the MATR by comparing with the average rate in last two weeks. The selected $\boldsymbol{\beta}$ is $(0.0068, 0.0077, 0.0198, 0.9998)^T$ which suggests that U_t plays an important role in the underlying nonlinear 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 financial data. The selected bandwidth is 0.24. The ratio $\mathrm{MTR}_{\mathrm{VC}}/\mathrm{MTR}_{\mathrm{ideal}} = 5.53\%$. The predictability is much lower than that of the MATR. If we include rates in last two weeks as regressors in the model (*i.e.* $m = 10$ in (5.10)), the ratio $\mathrm{MTR}_{\mathrm{VC}}/\mathrm{MTR}_{\mathrm{ideal}}$ increases to 7.26% which is still distance away from $\mathrm{MTR}_{\mathrm{MA}}/\mathrm{MTR}_{\mathrm{ideal}}$, while the ratio of the MSE of the fitted model to the sample variance of $\{Y_t\}$ 87.96%. The selected bandwidth is still 0.24, and $\widehat{\boldsymbol{\beta}} = (0.0020, 0.0052, 0.0129, 0.9999)^T$. Different subsets of regressors should be used at different places in the state space, according to our local variable selection procedure in Section 2.2.4.

The above calculations (also others not reported here) seem to suggest that U_t could be a dominated component in the selected index. This leads us to use the model (5.10) with prescribed $Z_t = U_{t-1}$, which is actually the approach adopted by Hong and Lee (1999). For $m = 5$, the fitting to the data used in estimation became worse; the ratio of the MSE of the fitted model to the sample variance of $\{Y_t\}$ is 97.39%. But it provides a better post-sample forecasting; $\mathrm{MTR}_{\mathrm{VC}}/\mathrm{MTR}_{\mathrm{ideal}}$ is 23.76%. The selected bandwidth is 0.24. The plots of estimated coefficient functions indicate

a possible under-smoothing. By increasing the bandwidth to 0.40, $\mathrm{MTR_{VC}}/\mathrm{MTR_{ideal}}$ is 31.35%. The estimated coefficient functions are plotted in Fig. 9(d). The rate of correct predictions for the direction of market movement (*i.e.* sign of Y_t) is 50% for the MATR, and 53% and 58% for the varying-coefficient model with bandwidth 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 with several financial data sets with only partial success from using varying-coefficient modeling techniques (with U_t as the prescribed index). In fact, for this particular data set, the model (5.10) with $Z_t = U_t$ and $m = 10$ gives a negative value of $\mathrm{MTR_{VC}}$. Note that the 'super-dominating' position of U_t in the selected smoothing variable $\widehat{\boldsymbol{\beta}}^T \mathbf{X}_t$ is partially due to the scaling difference between U_t and (Y_t, X_t); see also Fig. 9(a) and Fig. 9(c). In fact, if we standardize U_t, Y_t and X_t separately beforehand, the resulted $\widehat{\boldsymbol{\beta}}$ is $(0.59, -0.52, 0.07, 0.62)^T$ when $m = 5$, which is dominated by U_{t-1} and the contrast between Y_{t-1} and Y_{t-2}. ($\mathrm{MTR_{VC}}/\mathrm{MTR_{ideal}} = 1.42\%$. The ratio of 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. Finally, we remark that a different modeling approach should be adopted if our primary target is to maximize the mean trading return, which is obviously beyond the scope of this paper.

Appendix: Proof of Theorem 1

We use the same notation as in §2.

Proof of Theorem 1(i). It follows from the ordinary least-squares theory that the minimization of

$$E\left[\{Y - f(X)\}^2 \,\middle|\, \boldsymbol{\alpha}^T\mathbf{X} = z\right],$$

over the class of functions f defined by (3.5) exists. Let $f_0^*(z), \ldots, f_{d-1}^*(z)$ be the minimizer. Then

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

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

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

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

Consequently,

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

is bounded by $E(Y^2)$, and continuous on the compact set $\{\boldsymbol{\alpha} \in R^d \,|\, ||\boldsymbol{\alpha}|| = 1\}$. Hence, there exists $\boldsymbol{\beta}$ in the above set such that $R(\boldsymbol{\alpha})$ obtains its minimum at $\boldsymbol{\alpha} = \boldsymbol{\beta}$. Therefore, $g(\cdot)$ fulfilled (3.1) exists.

Theorem 1(ii) follows from the following two lemmas immediately. □

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

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

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

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

Proof. Without loss of the generality we assume $\boldsymbol{\beta} = (c, 0, \ldots, 0)^T$. Then, it follows from (A.1) that $\partial^2 F(\mathbf{x})/\partial x_j^2 = 0$ for $j = 1, \ldots, d$. Write $\boldsymbol{\alpha}^T\mathbf{x} = z$. Then from (A.2), we have that

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

For i with $\alpha_i \neq 0$, the above equation can be written as

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

This implies that

$$\ddot{f}_j(z) = \ddot{f}_i(z)\frac{\alpha_j}{\alpha_i}, \quad 1 \leq j \leq d \tag{A.3}$$

and

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

(A.3) implies that $f_j(z) = f_i(z)\alpha_j\alpha_i^{-1} + a_j z + b_j$. Substituting this into (A.2), we have

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

Therefore, we can choose $f_i(\cdot) \equiv 0$ in (A.2) for all $1 < i \leq d$ for which $\alpha_i \neq 0$, while all other $f_j(\cdot)'$s $(1 \leq j \leq d)$ are linear. Now, an application of the argument (A.4) to the newly formulated function leads to $f_0(z) = a_0 z + b_0$. Thus,

$$F(\mathbf{x}) = a_0\boldsymbol{\alpha}^T\mathbf{x} + b_0 + (a_1\boldsymbol{\alpha}^T\mathbf{x} + b_1)x_1 + \sum_{\substack{1<j\leq d \\ \alpha_j=0}} (a_j\boldsymbol{\alpha}^T\mathbf{x} + b_j)x_j.$$

Now, $\partial^2 F(\mathbf{x})/\partial x_i \partial x_j = a_j\alpha_i$ for any $\alpha_i \neq 0$, $\alpha_j = 0$ and $j \geq 2$, which should be 0 according to (A.1) since $\boldsymbol{\beta} = (c, 0, \ldots, 0)^T$. Hence, all a_j's $(j \geq 2)$ in the above expression are zero. This implies that

$$F(\mathbf{x}) = a_0\boldsymbol{\alpha}^T\mathbf{x} + b_0 + (a_i\boldsymbol{\alpha}^T\mathbf{x})x_1 + \sum_j b_j x_j = b_0 + (a_0\boldsymbol{\alpha}^T + \mathbf{b}^T)\mathbf{x} + a_i\boldsymbol{\alpha}^T\mathbf{x}\boldsymbol{\beta}^T\mathbf{x},$$

where $\mathbf{b}$ denotes a vector with the j-th component b_j for $j = 1$ and $\alpha_j = 0$, and 0 otherwise. The proof is completed. □

Lemma A.2. For any

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

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

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

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

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

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

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 k's.

Proof. Note that $x_d = \{\boldsymbol{\alpha}^T\mathbf{x} - \sum_{j=1}^{d-1} \alpha_j x_j\}/\alpha_d$. Define

$$g_0(z) = f_0(z) + \frac{1}{\alpha_d} f_d(z)z \quad \text{and} \quad g_j(z) = f_j(z) - \frac{\alpha_j}{\alpha_d} \quad \text{for } j = 1, \ldots, d-1.$$

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

Acknowledgements

We thank Professors N.C. Stenseth and A.R. Gallant for making available Canadian mink-muskrat data and pound/dollar exchange data analyzed in Section 4.2.

References

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

Bickel, P.J. and Rosenblatt, M. (1973) On some global measures of the deviation of density function estimates. *Ann. Statist.*, **1**, 1071–1095.

Breiman, L. and Friedman, J.H. (1985) Estimating optimal transformations for multiple regression and correlation (with discussion). *J. Ameri. Statist. Assoc.*, **80**, 580–619.

Brumback, B. and Rice, J.A. (1998) Smoothing spline models for the analysis of nested and crossed samples of curves (with discussion). *J. Amer. Statist. Assoc.*, **93**, 961–994.

Cai, Z. (1999) An optimal approach in varying-coefficient models. *Submitted for publication.*

Cai, Z., Fan, J. and Li, R. (2000) Efficient estimation and inferences for varying-coefficient models. *J. Ameri. Statist. Assoc.*, to appear.

Cai, Z., Fan, J. and Yao, Q. (1998) Functional-coefficient regression models for nonlinear time series models. by *J. Ameri. Statist. Assoc.*, to appear.

Cai, Z. and Tiwari, R.C. (1999) Application of local linear autoregressive model to BOD time series. *Environmetrics*, to appear.

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

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

Cederman, L.E. and Penubarti, M. (1999) Evolutionary liberalism: Exploring the dynamics of interstate conflict. *Submitted for publication.*

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

Cleveland, W.S., Grosse, E. and Shyu, W.M. (1992) Local regression models. In *Statistical Models in S* (ed. J.M. Chambers and T.J. Hastie), pp.309–376. Pacific Grove: Wadsworth & 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. *Numerische Mathematik*, **31**, 377–403.

Diggle, P.J., Liang, K.Y. and Zeger, S. L. (1994) *Analysis of Longitudinal Data.* Oxford: Oxford University Press.

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

Errington, P.L. (1961) *Muskrats and Marsh Management.* Lincoln: University of Nebraska Press.

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

Fan, J. and Chen, J. (1997) One-step local quasi-likelihood estimation. *Manuscript.*

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

Fan, J., Härdle. W.H. and Mammen, E. (1998) Direct estimation of additive and linear components for high dimensional data. *Ann. Statist.*, **26**, 943-971.

Fan, J., Zhang, C. and Zhang, J. (1999) Sieve likelihood ratio statistics and Wilks phenomenon. *Technical report*, Department of Statistics, University of California at Los Angeles.

Fan, J. and Zhang, J. (2000) Functional linear models for longitudinal data. *J. R. Statist. Soc. B*, to appear.

Fan, J. and Zhang, W. (2000a) Statistical estimation in varying-coefficient models. *Ann. Statist.*, to appear.

Fan, J. and Zhang, W. (2000b) Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scand. J. Statist.*, to appear.

Fang, K.T. and Wang, Y. (1995) *Number-theoretic Methods in Statistics*. London: Chapman and Hall.

Friedman, J.H. (1991) Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, **19**, 1–141.

Friedman, J.H. and Stuetzle, W. (1981) Projection pursuit regression. *J. Ameri. Statist. Assoc.*, **76**, 817–823.

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* (ed. W.A. Barnett, J. Powell and G.E. Tauchen), pp.199-240. Cambridge: Cambridge University Press.

Green, P.J. and Silverman, B.W. (1994) *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman and Hall.

Gu, C. and Wahba, G. (1993) Semiparametric ANOVA with tensor product thin plate splines. *J. R. Statist. Soc. B*, **55**, 353-368.

Hand, D. and Crowder, M. (1996) *Practical Longitudinal Data Analysis*. London: Chapman and Hall.

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

Härdle, W. and Mammen, E. (1993) Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, **21**, 1926–1947.

Härdle, W. and Stoker, D.W. (1989) Investigating smooth multiple regression by method of average derivatives. *J. Amer. Statist. Assoc.*, **84**, 986–955.

Hart, J.D. (1997) *Nonparametric Smoothing and Lack-of-Fit Tests*. New York: Springer-Verlag.

Hart, J.D. and Wehrly, T.E. (1986) Kernel regression estimation using repeated measurements data. *J. Amer. Statist. Assoc.*, **81**, 1080–1088.

Hart, J.D. and Wehrly, T.E. (1993) Consistency of cross-validation when the data are curves. *Stochastic Process and their Applications*, **45**, 351–361.

Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models*. London: Chapman and Hall.

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

Hong, Y. and Lee, T.-H. (1999) Inference and forecast of exchange rates via generalized spectrum and nonlinear time series models. *Manuscript*.

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.

Ichimura, H. (1993) Semiparametric least-squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics*, **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* (ed. E. Acar and S Satchell). Butterworth-Heinemann.

LeBaron, B. (1999) Technical trading rule profitability and foreign exchange intervention. *J. International Economics*, to appear.

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

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

Moyeed, R.A. and Diggle, P.J. (1994) Rates of convergence in semi-parametric modeling of longitudinal data. *Austral. J. Statist.*, **38**, 75–93.

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*, Lecture Notes in Statistics, No. 11. New York: Springer-Verlag.

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

Rice, J. A. and Silverman, B. W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc. B*, **53**, 233–243.

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

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

Seift, B. and Gasser, Th. (1996) Finite-sample variance of local polynomial: Analysis and solutions. *J. Ameri. Statist. Assoc.*, **91**, 267–275.

Silverman, B. W. (1996) Smoothed functional principal components analysis by choice of norm. *Ann. Statist.*, **24**, 1–24.

Simonoff, J.S. (1996) *Smoothing Methods in Statistics*. New York: Springer-Verlag.

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. *Proceedings of National Academy of Science, US.*, **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. (1999) ¿From patterns to processes: phase- and density-dependencies in Canadian lynx cycle. *Proceedings of National Academy of Science, Washington*, to appear.

Stone, C.J., Hansen, M.H., Kooperberg, C. and Truong, Y.K. (1997) Polynomial splines and their tensor products in extended linear modeling. *Ann. Statist.*, **25**, 1371–1470.

Tong, H. (1990) *Nonlinear Time Series: A Dynamical System Approach*. Oxford: Oxford University Press.

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), 507–523. Amsterdam, North Holland.

Wahba, G. (1984) Partial spline models for semiparametric estimation of functions of several variables. In *Statistical Analysis of Time Series*, Proceedings of the Japan U.S. Joint Seminar, Tokyo, 319–329. Tokyo: Institute of Statistical Mathematics.

Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: SIAM.

Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing*. London: Chapman and Hall.

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. Amer. Statist. Assoc.*, **93**, 1388–1402.

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

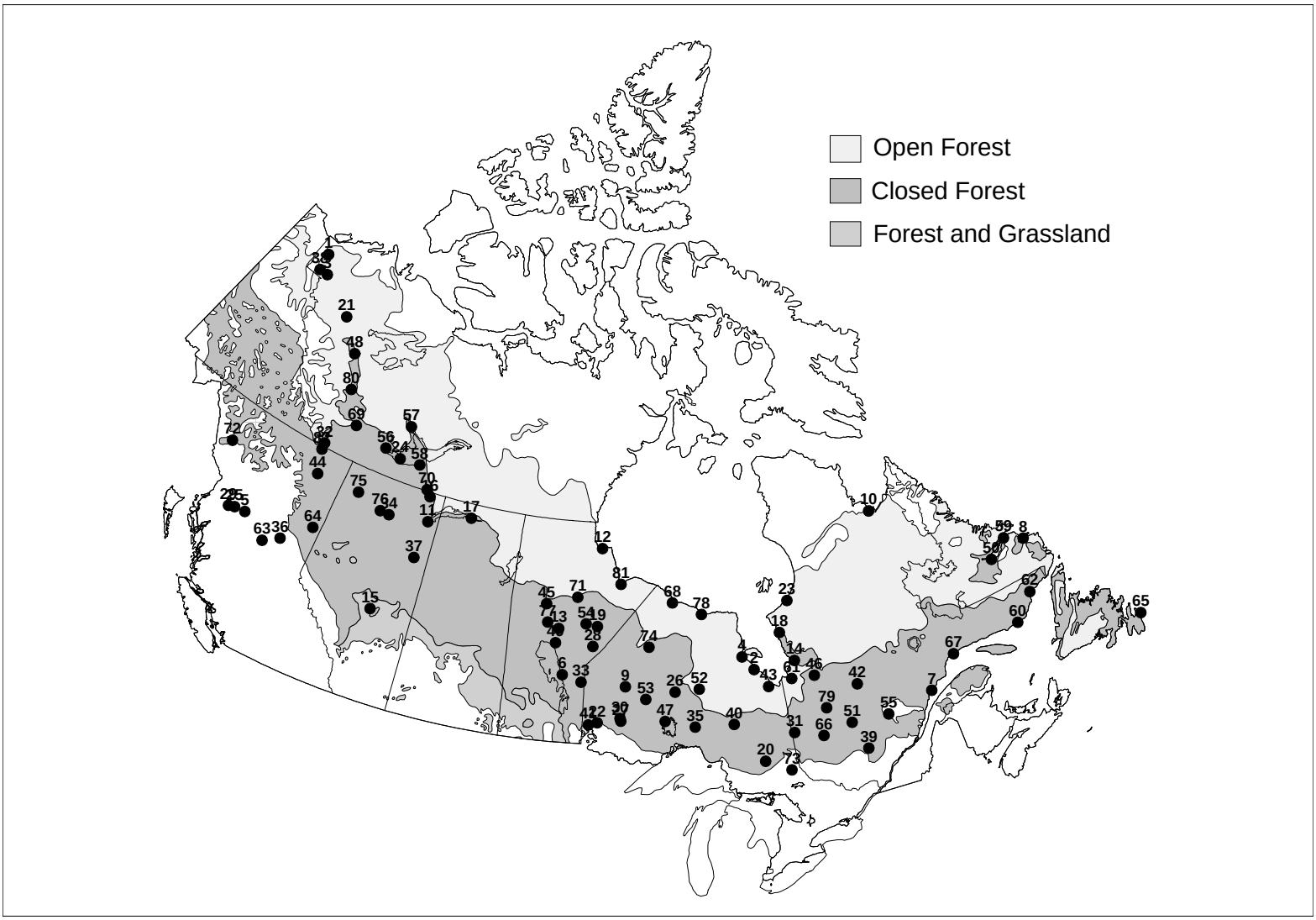
Xia, Y. and Li, W.K. (1999b) On the estimation and testing of functional-coefficient linear models. *Statistica Sinica*, **9**, 735–757.

Yao, Q., Tong, H., Finkenstädt, B. and Stenseth, N.C. (1998) Common structure in panels of short time series. *Submittedfor publication.*

Zeger, S.L. and Diggle, P.J. (1994) Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics*, **50**, 689–699.

Zhang, W. and Lee, S.Y. (1999) On local polynomial fitting of varying-coefficient models. *Submitted for publication.*

Zhang, W. and Lee, S.Y. (2000) Variable bandwidth selection in varying-coefficient models. *J. Multivariate Analysis*, to appear.


Open Forest
Closed Forest
Forest and Grassland