

# Functional Coefficient Regression Models for Non-linear Time Series: A Polynomial Spline Approach

JIANHUA Z. HUANG

*University of Pennsylvania*

HAIPENG SHEN

*University of North Carolina at Chapel Hill*

**ABSTRACT.** We propose a global smoothing method based on polynomial splines for the estimation of functional coefficient regression models for non-linear time series. Consistency and rate of convergence results are given to support the proposed estimation method. Methods for automatic selection of the threshold variable and significant variables (or lags) are discussed. The estimated model is used to produce multi-step-ahead forecasts, including interval forecasts and density forecasts. The methodology is illustrated by simulations and two real data examples.

*Key words:* forecasting, functional autoregressive model, non-parametric regression, threshold autoregressive model, varying coefficient model

## 1. Introduction

For many real time series data, non-linear models are more appropriate than linear models for accurately describing the dynamic of the series and making multi-step-ahead forecasts (see e.g. Tong, 1990; Franses & van Dijk, 2000). Recently, non-parametric regression techniques have found important applications in non-linear time series analysis (Tjøstheim & Auestad, 1994; Tong, 1995; Härdle *et al.*, 1997). Although the non-parametric approach is appealing, its application usually requires an unrealistically large sample size when more than two lagged variables (or exogenous variables) are involved in the model (the so-called ‘curse of dimensionality’). To overcome the curse of dimensionality, it is necessary to impose some structure to the non-parametric models.

A useful structured non-parametric model that still allows appreciable flexibility is the functional coefficient regression model described as follows. Let  $\{Y_t, \mathbf{X}_t, U_t\}_{-\infty}^{\infty}$  be jointly strictly stationary processes with  $\mathbf{X}_t = (X_{t1}, \dots, X_{td})$  taking values in  $\mathbb{R}^d$  and  $U_t$  in  $\mathbb{R}$ . Let  $E(Y_t^2) < \infty$ . The multivariate regression function is defined as

$$f(\mathbf{x}, u) = E(Y_t | \mathbf{X}_t = \mathbf{x}, U_t = u).$$

In a pure time series context, both  $\mathbf{X}_t$  and  $U_t$  consist of some lagged values of  $Y_t$ . The functional coefficient regression model requires that the regression function has the form

$$f(\mathbf{x}, u) = \sum_{j=1}^d a_j(u)x_j, \tag{1}$$

where  $a_j(\cdot)$ s are measurable functions from  $\mathbb{R}$  to  $\mathbb{R}$  and  $\mathbf{x} = (x_1, \dots, x_d)^T$ . As  $U_t \in \mathbb{R}$ , only one-dimensional smoothing is needed in estimating the model (1).

The functional coefficient regression model extends several familiar non-linear time series models such as the exponential autoregressive (EXPAR) model of Haggan & Ozaki (1981) and Ozaki (1982), threshold autoregressive (TAR) model of Tong (1990), and functional

autoregressive (FAR) model of Chen & Tsay (1993); see Cai *et al.* (2000) for more discussion. We borrow terminology from TAR and call  $U_t$  the threshold variable. The formulation adopted in this paper allows both  $U_t$  and  $X_t$  in (1) to contain exogenous variables. Functional coefficient models (or varying-coefficient models) have been paid much attention recently, but most of the work has focused on i.i.d. data or longitudinal data (Hastie & Tibshirani, 1993; Hoover *et al.*, 1998; Wu *et al.*, 1998; Fan & Zhang, 1999, 2000; Chiang *et al.*, 2001; Huang *et al.*, 2002).

The local polynomial method (Fan & Gijbels, 1996) has been previously applied to the functional coefficient time series regression models. Chen & Tsay (1993) proposed an iterative algorithm in the spirit of local constant fitting to estimate the coefficient functions. Cai *et al.* (2000) and Chen & Liu (2001) used the local linear method for estimation, where the same smoothing parameter (bandwidth) was employed for all coefficient functions. The focus of these papers has been on the theoretical and descriptive aspects of the model and on the estimation of coefficient functions and hypothesis testing. The important issue of multi-step-ahead forecasting using functional coefficient models has not been well-studied. For example, Cai *et al.* (2000) only considered one-step-ahead forecasting carefully. Moreover, utilization of a single smoothing parameter in the local linear method could be inadequate if the coefficient functions have different smoothness.

In this paper, we propose a global smoothing method based on polynomial splines for the estimation of functional coefficient regression models for non-linear time series. Different coefficient functions are allowed to have different smoothing parameters. We establish consistency and rate of convergence results to give support to the proposed estimation method. Methods for automatic selection of the threshold variable and significant variables (or lags) are discussed. Moreover, we provide a method to produce multi-step-ahead forecasts, including point forecasts, interval forecasts, and density forecasts, using the estimated model.

There have been many applications of the local polynomial method to non-linear time series modelling in the literature. In addition to the references mentioned above, we list Truong & Stone (1992), Truong (1994), Tjøstheim & Auestad (1994), Yang *et al.* (1999), Cai & Masry (2000) and Tschernig & Yang (2000), to name just a few. We demonstrate in this paper that global smoothing provides an attractive alternative to local smoothing in non-linear time series analysis. The attraction of spline-based global smoothing is that it is closely related to parametric models and thus standard methods for parametric models can be extended to non-parametric settings. However, it is not clear from the literature why the spline method should work as a flexible non-parametric method for non-linear time series. In particular, the asymptotic theory for parametric models does not apply. There has been substantial recent development of asymptotic theory for the spline method for i.i.d. data (see e.g. Stone, 1994; Huang, 1998, 2001). The consistency and rates of convergence of the spline estimators developed in this paper justifies that the spline method really works in a time series context.

One appealing feature of the spline method proposed in this paper is that it yields a fitted model with a parsimonious explicit expression. This turns out to be an advantage over the existing local polynomial method. We can simulate a time series from the fitted spline models and thereby conveniently produce multi-step-ahead forecasts based on the simulated data. As a contrast, direct implementation of the simulation-based forecasting method using the local polynomial smoothing can be computationally expensive and extra care needs to be taken in order to relieve the computational burden (see section 3).

The rest of the paper is organized as follows. Section 2 introduces the proposed spline-based global smoothing method, discusses the consistency and rates of convergence of the spline estimates, and provides some implementation details, such as knot placement, knot number selection, and determination of the threshold variable and significant variables. Section 3

proposes a method for multi-step-ahead forecasting using the fitted functional coefficient model. Some results of a simulation study are reported in section 4. Two real data examples, US GNP and Dutch guilder–US dollar exchange rate time series, are used in sections 5 and 6, respectively, to illustrate the proposed method and potential usefulness of the functional coefficient model. Some concluding remarks are given in section 7. All technical proofs are relegated to the appendix.

**2. Spline estimation**

In this section, we describe our estimation method using polynomial splines. Our method involves approximating the coefficient functions  $a_j(\cdot)$ s by polynomial splines. Consistency and rates of convergence of the spline estimators are developed. Some implementation details are also discussed.

*2.1. Identification of the coefficient functions*

We first discuss the identifiability of the coefficient functions in model (1). We say that the coefficient functions in the functional coefficient model (1) are identifiable if  $f(x, u) = \sum_{j=1}^d a_j^{(1)}(u)x_j \equiv \sum_{j=1}^d a_j^{(2)}(u)x_j$  implies that  $a_j^{(1)}(u) = a_j^{(2)}(u)$  for a.e.  $u, j = 1, \dots, d$ .

We assume that  $E(\mathbf{X}_t \mathbf{X}_t^T | U_t = u)$  is positive definite for a.e.  $u$ . Under this assumption, the coefficient functions in model (1) are identifiable. To see why, denote  $\mathbf{a}(u) = (a_1(u), \dots, a_d(u))^T$ . Then

$$E \left[ \left\{ \sum_{j=1}^d a_j(U_t) X_{tj} \right\}^2 \middle| U_t = u \right] = \mathbf{a}^T(u) E(\mathbf{X}_t \mathbf{X}_t^T | U_t = u) \mathbf{a}(u). \tag{2}$$

If  $\sum_{j=1}^d a_j(u)x_j \equiv 0$ , then  $E[\{\sum_j a_j(U_t) X_{tj}\}^2] = 0$ , and thus  $E[\{\sum_j a_j(U_t) X_{tj}\}^2 | U_t = u] = 0$  for a.e.  $u$ . Therefore, it follows from (2) and the positive definiteness of  $E(\mathbf{X}_t \mathbf{X}_t^T | U_t = u)$  that  $a_j(u) = 0$  a.e.  $u, j = 1, \dots, d$ .

*2.2. Spline approximation and least squares*

Polynomial splines are piecewise polynomials with the polynomial pieces joining together smoothly at a set of interior knot points. To be precise, a (polynomial) spline of degree  $l \geq 0$  on an interval  $\mathcal{U}$  with knot sequence  $\xi_0 < \xi_1 < \dots < \xi_{M+1}$ , where  $\xi_0$  and  $\xi_{M+1}$  are the two end points of  $\mathcal{U}$ , is a function that is a polynomial of degree  $l$  on each of the intervals  $[\xi_m, \xi_{m+1})$ ,  $0 \leq m \leq M - 1$ , and  $[\xi_M, \xi_{M+1}]$ , and globally has  $l - 1$  continuous derivatives for  $l \geq 1$ . A piecewise constant function, linear spline, quadratic spline and cubic spline correspond to  $l = 0, 1, 2, 3$ , respectively. The collection of spline functions of a particular degree and knot sequence forms a linear function space and it is easy to construct a convenient basis for it. For example, the space of splines with degree 3 and knot sequence  $\xi_0, \dots, \xi_{M+1}$  forms a linear space of dimension  $M + 4$ . The truncated power basis for this space is  $1, x, x^2, x^3, (x - \xi_1)_+^3, \dots, (x - \xi_M)_+^3$ . A basis with better numerical properties is the B-spline basis. See de Boor (1978) and Schumaker (1981) for a comprehensive account of spline functions.

The success of the proposed method relies on the good approximation properties of polynomial splines. Suppose that in (1) the coefficient function  $a_j, j = 1, \dots, d$ , is smooth. Then it can be approximated well by a spline function  $a_j^*$  in the sense that  $\sup_{u \in \mathcal{U}} |a_j^*(u) - a_j(u)| \rightarrow 0$  as the number of knots of the spline tends to infinity (de Boor, 1978; Schumaker, 1981). Thus, there is a set of basis functions  $B_{js}(\cdot)$  (e.g. B-splines) and constants  $\beta_{js}^*, s = 1, \dots, K_j$ , such that

$$a_j(u) \approx a_j^*(u) = \sum_{s=1}^{K_j} \beta_{js}^* B_{js}(u). \tag{3}$$

Then, we can approximate (1) by

$$f(x, u) \approx \sum_{j=1}^d \left\{ \sum_{s=1}^{K_j} \beta_{js}^* B_{js}(u) \right\} x_j,$$

and estimate the  $\beta_{js}^*$ s by minimizing

$$\ell(\beta) = \sum_{t=1}^n \left( Y_t - \sum_{j=1}^d \left\{ \sum_{s=1}^{K_j} \beta_{js} B_{js}(U_t) \right\} X_{tj} \right)^2, \tag{4}$$

with respect to  $\beta$ , where  $\beta = (\beta_1^T, \dots, \beta_d^T)^T$  and  $\beta_j = (\beta_{j1}, \dots, \beta_{jK_j})^T$ . Assume that (4) can be uniquely minimized and denote its minimizer by  $\hat{\beta} = (\hat{\beta}_1^T, \dots, \hat{\beta}_d^T)^T$ , with  $\hat{\beta}_j = (\hat{\beta}_{j1}, \dots, \hat{\beta}_{jK_j})^T$  for  $j = 1, \dots, d$ . Then  $a_j$  is estimated by  $\hat{a}_j(u) = \sum_{s=1}^{K_j} \hat{\beta}_{js} B_{js}(u)$  for  $j = 1, \dots, d$ . We refer to  $\hat{a}_j(\cdot)$ s as the least squares spline estimates.

The idea of using basis expansions can be applied more generally to other basis systems for function approximation such as polynomial bases and Fourier bases. We focus in this paper on B-splines because of the good approximation properties of splines and the good numerical properties of the B-spline basis. When B-splines are used, the number of terms  $K_j$  in the approximation (3) depends on the number of knots and the order of the B-splines. We discuss in section 2.4.2 how to select  $K_j$ , or the number of knots, using the data. Note that different  $K_j$ s are allowed for different  $a_j$ s. This provides flexibility when different  $a_j$ s have different smoothness.

### 2.3. Consistency and rates of convergence

To provide some theoretical support of the proposed method, we establish in this section the consistency and convergence rates of the spline estimates. For the data generating process, we assume that

$$Y_t = \sum_{j=1}^d a_j(U_t) X_{tj} + \epsilon_t, \quad t = 1, \dots, n$$

where  $\epsilon_t$  is independent of  $U_{t'}, X_{t'j}$ ,  $j = 1, \dots, d$ ,  $t' \leq t$ , and  $\epsilon_{t'}, t' < t$ ,  $E(\epsilon_t) = 0$ , and  $\text{var}(\epsilon_t) \leq C$  for some constant  $C$  (the noise errors can be heteroscedastic). When  $U_t$  and  $\mathbf{X}_t = (X_{t1}, \dots, X_{td})$  consist of lagged values of  $Y_t$ , this is the FAR model considered by Chen & Tsay (1993).

We focus on the performance of the spline estimates on a compact interval  $\mathbb{C}$ . Let  $\|a\|_2 = \{\int_{\mathbb{C}} a^2(t) dt\}^{1/2}$  be the  $L_2$ -norm of a square integrable function  $a(\cdot)$  on  $\mathbb{C}$ . We say that an estimate  $\hat{a}_j$  is consistent in estimating  $a_j$  if  $\lim_n \|\hat{a}_j - a_j\|_2 = 0$  in probability.

For clarity in presentation, we now represent the spline estimate in a function space notation. Let  $\mathbb{G}_j$  be a space of polynomial splines on  $\mathbb{C}$  with a fixed degree and knots having bounded mesh ratio (i.e., the ratios of the differences between consecutive knots are bounded away from zero and infinity uniformly in  $n$ ). Then the spline estimates  $\hat{a}_{js}$  are given by

$$\{\hat{a}_j, j = 1, \dots, d\} = \arg \min_{g_j \in \mathbb{G}_j, j=1, \dots, d} \sum_{t=1}^n \left\{ Y_t - \sum_{j=1}^d g_j(U_t) X_{tj} \right\}^2 I(U_t \in \mathbb{C}).$$

This is essentially the same as (4) but in a function space notation (assuming that  $B_{js}$ ,  $s = 1, \dots, K_j$ , is a basis of  $\mathbb{G}_j$ ). Here, we employ a weighting function in the least squares criterion to screen off extreme observations, following a common practice in non-parametric time series (see e.g. Tjøstheim & Auestad, 1994).

Let  $K_n = \max_{1 \leq j \leq d} K_j$ . Set  $\rho_{n,j} = \max_{g \in \mathbb{G}_j} \|g - a_j\|_2$  and  $\rho_n = \max_{1 \leq j \leq d} \rho_{n,j} = \inf_{g \in \mathbb{G}_j} \|g - a_j\|_2$ .

**Theorem 1**

Suppose conditions (i)–(v) in the appendix hold. Then

$$\|\hat{a}_j - a_j\|_2^2 = O_P\left(\frac{K_n}{n} + \rho_n^2\right), \quad j = 1, \dots, d.$$

In particular, if  $\rho_n = o(1)$ , then  $\hat{a}_j$  is consistent in estimating  $a_j$ , that is,  $\|\hat{a}_j - a_j\|_2 = o_P(1)$ ,  $j = 1, \dots, d$ .

The rate of convergence in this theorem is in parallel with that for i.i.d. data (Stone *et al.*, 1997; Huang, 1998, 2001). Here  $K_n$  measures the sizes of the estimation spaces  $\mathbb{G}_j$ s. The quantity  $\rho_n$  measures the size of the approximation error, and its magnitude is determined by the smoothness of  $a_j$ s and the dimension of the spline spaces  $\mathbb{G}_j$ s. For example, if  $a_j$ ,  $j = 1, \dots, d$ , have bounded second derivatives, then  $\rho_n = O(K_n^{-2})$  (see theorem 7.2 in Chapter 7 of DeVore & Lorentz, 1993). In this case, the rate of convergence of  $\hat{a}_j$  to  $a_j$  is  $K_n/n + K_n^{-4}$ . In particular, if  $K_n$  increases with the sample size and is of the same order as  $n^{1/5}$  (precisely,  $K_n/n^{1/5}$  is bounded away from zero and infinity), then the rate of convergence is  $n^{-4/5}$ .

Note that in this paper we do not model the form of heteroscedasticity. Our estimate can be viewed as a quasi-likelihood estimate and the consistency and rate of convergence results hold even in the presence of heteroscedastic errors. (However, the multi-step-ahead forecasting method discussed in section 3 does require homoscedastic errors.) In principle, one could model the conditional variance function  $\sigma^2(\mathbf{x}, u) = \text{var}(Y_t | \mathbf{X}_t = \mathbf{x}, U_t = u)$  non-parametrically and improve the efficiency. Implementation of this idea is, however, beyond the scope of this paper.

2.4. Implementation

In this section we discuss some implementation details of the proposed spline method, including knot placement, data-driven knot number selection, and automatic selection of the threshold variable and significant variables (or lags).

2.4.1. Knot positions

For simplicity, we consider only two ways to place the knots. For a given number of knots, the knots can be placed evenly such that the distances between any two adjacent knots are the same (referred to as equally spaced knots). Alternatively, the knots can be placed at the sample quantiles of the threshold variable so that there are about the same number of observed values of the threshold variables between any two adjacent knots (referred to as quantile knots). The number of knots is selected using the data (see section 2.4.2).

When the threshold variable consists of lagged values of the time series, we observe in simulations that the results based on quantile knots wiggle much more (and thus less satisfactory) than those based on equally spaced knots. That is because the observed values of the threshold variable (which are the values of the time series itself) do not spread evenly and are very sparse near the boundaries, so that the model selection criteria tend to choose too many knots in the middle of the data range if quantile knots are used. We recommend using equally spaced knots.

2.4.2. Selection of the number of knots

The numbers of knots serve as the smoothing parameters, playing the same role as bandwidths in the local linear method. Although a subjective smoothing parameter may be determined by

examining the estimated curves or plots of the residuals, an automatic procedure for selecting  $K_j$  from the data is of practical interest. The numbers of knots are chosen to minimize an appropriate criterion function. We will consider four criterion functions: AIC (Akaike, 1974),  $AIC_C$  (Hurvich & Tsai, 1989), BIC (Schwarz, 1978) and modified cross-validation (MCV), (Cai *et al.*, 2000). Let  $n$  denote the number of terms on the right of (4),  $p = \sum_j K_j$  be the number of parameters to be estimated, and  $RSS = \ell(\hat{\beta})$  be the residual sum of squares in minimizing (4). The first three criteria are defined as

$$AIC = \log\left(\frac{RSS}{n}\right) + \frac{2p}{n}, \quad AIC_C = AIC + \frac{2(p+1)(p+2)}{n(n-p-2)}, \quad BIC = \log\left(\frac{RSS}{n}\right) + \log(n) \times \frac{p}{n}.$$

The MCV criterion can be regarded as a modified multi-fold cross-validation criterion that is attentive to the structure of time series data. Let  $m$  and  $Q$  be two given positive integers and  $n > mQ$ . Use  $Q$  subseries of lengths  $n - qm$  ( $q = 1, \dots, Q$ ) to estimate the unknown coefficient functions  $a_j$  and to compute the one-step forecasting errors of the next section of the time series of length  $m$  based on the estimated models. The MCV criterion function is  $AMS = \sum_{q=1}^Q AMS_q$ , where for  $q = 1, \dots, Q$ ,

$$AMS_q = \frac{1}{m} \sum_{t=n-qm+1}^{n-qm+m} \left( Y_t - \sum_{j=1}^d \left\{ \sum_{s=1}^{K_j} \hat{\beta}_{js}^{(q)} B_{js}(U_t) \right\} X_{tj} \right)^2,$$

and  $\{\hat{\beta}_{js}^{(q)}\}$  are computed from the sample  $\{(Y_t, U_t, X_t), 1 \leq t \leq n - qm\}$ . It is suggested to use  $m = [0.1n]$  and  $Q = 4$  in Cai *et al.* (2000).

In our simulation studies (some results will be reported in section 4), we find that AIC and  $AIC_C$  behave very similarly and both perform better than BIC and MCV. We shall use AIC in our real data examples.

### 2.4.3. Selection of the threshold variable and significant variables

It is important to choose the number of terms in (1) and an appropriate threshold variable  $U$  in applying functional coefficient regression models. Knowledge on physical background of the data may help. When no prior information is available, a data-driven method is desirable.

To fix ideas, consider the functional coefficient autoregressive models

$$Y_t = a_1(Y_{t-i_0})Y_{t-i_1} + \dots + a_d(Y_{t-i_0})Y_{t-i_d} + \epsilon_t, \quad i_0 > 0, \quad 0 < i_1 < \dots < i_d.$$

Extension to include exogenous variables is straightforward. This is a special case of (1) with  $U_t = Y_{t-i_0}$  and  $X_t = (X_{t1}, \dots, X_{td})^T = (Y_{t-i_1}, Y_{t-i_2}, \dots, Y_{t-i_d})^T$ . We assume  $\max(i_0, i_1, \dots, i_d) \leq p_{\max}$  for some constant  $p_{\max} > 0$ . We call  $i_0$  the threshold lag and  $i_1, \dots, i_d$  the significant lags. To select the threshold lag and significant lags, we minimize a criterion such as AIC ( $AIC_C$ , BIC or MCV can also be used). A stepwise procedure is employed to expedite computation.

Specifically, consider the following class of candidate models

$$Y_t = \sum_{j \in S_d} a_j(Y_{t-d})Y_{t-j} + \epsilon_t, \quad 1 \leq d \leq p_{\max}, \quad S_d \subset \{1, \dots, p_{\max}\}. \tag{5}$$

For a given candidate threshold lag  $d$ , we decide on an optimal subset  $S_d^*$  of significant lags by stepwise addition followed by stepwise deletion. In the addition stage, we add one significant lag at a time, choosing among all candidate lags not yet selected in the model by minimizing the mean square error (MSE); the addition process stops if the number of lags selected equals a pre-specified number  $q_{\max} \leq p_{\max}$ . In the deletion stage, we delete one lag at a time from the collection of lags selected in the addition stage, by also minimizing the MSE until there is no lagged variable left in the model. After the stepwise addition and deletion, we get a sequence of

subsets of lag indices, and the one which minimizes AIC is chosen as  $S_d^*$ . The final model is determined by the pair  $\{d, S_d^*\}$  that produces the smallest AIC. If  $p_{\max}$  is not too big, we can let  $q_{\max} = p_{\max}$ .

### 3. Forecasting

Forecasting is an important objective in time series analysis. Constructing multi-step-ahead forecasts for non-linear time series models is considerably more difficult than for linear models, and exact analytical solutions are not available. We propose a simulation-based method to make multi-step-ahead forecasts using the functional coefficient models. One advantage of the simulation-based method is that it automatically provides interval forecasts and density forecasts in addition to point forecasts.

Specifically, consider an observed time series  $\{y_1, \dots, y_T\}$  whose dynamic follows a functional coefficient model with homoscedastic errors. Suppose we want to forecast  $y_{T+k}$  for some lead time  $k \geq 1$ . For the mean squared prediction error criterion, the optimal predictor  $\hat{y}_{T+k}$  is given by the conditional mean

$$\hat{y}_{T+k} = E(y_{T+k} | \mathcal{F}_T),$$

where  $\mathcal{F}_T$  is the given information set up to time  $T$ . Using the fitted model recursively, we generate  $M$  series  $\{y_{T+1,m}, \dots, y_{T+k,m}\}$ ,  $1 \leq m \leq M$ , that follow the observed series  $y_1, \dots, y_T$ . The error term in the model can be generated by random sampling the residuals from the fitted model with replacement (this is essentially the idea of bootstrapping, see Efron & Tibshirani, 1993). The  $k$ -step-ahead point forecast is then given by the sample mean of the simulated values  $\{y_{T+k,1}, \dots, y_{T+k,M}\}$ . Before the resampling, it is advisable to perform residual diagnostics to ensure that there is no serial correlation and strong evidence of heteroscedasticity among the residuals.

The simulated series can also be used to construct interval forecasts and density forecasts. To be specific, for  $0 < \beta < 1$ , let  $y_{T+k}^\beta$  denote the  $100 \times \beta\%$  sample quantile of  $\{y_{T+k,1}, \dots, y_{T+k,M}\}$ . The  $k$ -step-ahead forecast interval with confidence level  $1 - \alpha$  is given by  $[y_{T+k}^{1-\alpha/2}, y_{T+k}^{\alpha/2}]$ . The histogram of the simulated data  $\{y_{T+k,1}, \dots, y_{T+k,M}\}$  gives a preliminary  $k$ -step-ahead density forecast; alternatively, these simulated data can be smoothed using any density estimation procedure to produce a smooth forecast density.

The forecasting method described above can in principle be used with any non-parametric estimates as long as one can generate random samples from the fitted model. Note that the spline estimates are specified by several parameters in basis expansions. The explicit expression of the spline estimate makes it convenient to generate random samples from the fitted model. Therefore, we propose to implement the above forecasting method using the spline estimates to make multi-step-ahead forecasts.

However, direct implementation with the local linear smoothing method can be computationally expensive. The local linear smoothing method does not offer a parsimonious explicit expression of the fitted model. It requires refitting at every point where the fitted coefficient functions need to be evaluated. This could cause some computational problem as a large number of time series need to be simulated in the implementation of simulation-based forecasting method (5000 series are simulated in our implementation with the spline method). Fast computation of the local linear estimator using binning has been proposed in the literature (see e.g., Fan & Gijbels, 1996). It is possible to adapt the binning technique to yield fast implementation of the simulation-based forecasting method based on the local linear smoothing. However, there is an issue of choosing bin size and how to quantify the impact of errors caused by binning, especially the cumulative effect of such errors in the multi-step-forecasting context.

Further development of the binning method in the time series context would be interesting but beyond the scope of this paper.

In actual implementation of the simulation-based forecasting method, we have applied a certain truncation rule when the value of the threshold variable to be plugged into the model is outside the range of its historical values. If this happens at the beginning of the simulated path, we replace the value of the threshold variable by its closest historical value; if this happens later in the path, we discard the whole simulated series. The reason for doing this is that the spline estimates of the coefficient functions outside the range of the historical data are not reliable. The percentage of discarded series out of the 5000 simulated series is about 10% in analysing the US GNP data in section 5 and none in analysing the Dutch guilder exchange rate data in section 6.

**4. Simulations**

Monte Carlo simulations have been conducted to evaluate the finite sample performance of the proposed spline estimator and to compare the four criteria in section 2.4.2 for selecting the number of knots in spline estimation. We find that the spline estimates with AIC or AIC<sub>C</sub> selected number of knots work the best. We present in this section results from one simulated example.

We need a criterion to gauge the performance of estimators. For an estimator  $\hat{a}_j(\cdot)$  of  $a_j(\cdot)$ , define the square-root of average squared error (RASE) as

$$RASE_j = \left[ n_{\text{grid}}^{-1} \sum_{k=1}^{n_{\text{grid}}} \{ \hat{a}_j(u_k) - a_j(u_k) \}^2 \right]^{1/2}, \tag{6}$$

where  $\{u_k, k = 1, \dots, n_{\text{grid}}\}$  are grid points chosen to be equally spaced in a certain interval within the range of data. In order to make the RASE values comparable from simulation to simulation, we need to choose a common interval to hold the grid points. However, because of the time series nature of the data, the range of the data may vary substantially from simulation to simulation, so some care needs to be taken. For the example below, the interval to hold the grid points in (6) is specified as follows: the left boundary is chosen to be the maximum of the 2.5 percentiles (100 of them) of the 100 data sets, and the right boundary is the minimum of the 97.5 percentiles of the data sets. Note that our interval for holding the grid points is different from that used in Cai *et al.* (2000).

*Example 1.* Consider an EXPAR model (Haggan & Ozaki, 1981; Ozaki, 1982; Cai *et al.*, 2000):

$$Y_t = a_1(Y_{t-1})Y_{t-1} + a_2(Y_{t-1})Y_{t-2} + \epsilon_t, \tag{7}$$

where  $a_1(u) = 0.138 + (0.316 + 0.982u)e^{-3.89u^2}$ ,  $a_2(u) = -0.437 - (0.659 + 1.260u)e^{-3.89u^2}$  and  $\{\epsilon_t\}$  are i.i.d.  $N(0, 0.2^2)$ . In each simulation run, a time series of length 400 is drawn from this model. The simulation is replicated 100 times.

For our spline estimators, we used quadratic splines with knots equally spaced between two boundary knots (the spline functions are defined using extrapolation outside the boundary knots). As the range of the data varies substantially among the simulation runs, the boundary knots were not placed at the data range. Instead, the two boundary knots were placed correspondingly at the 0.5 and 99.5 percentiles of the threshold variable  $Y_{t-1}$  for a given data set, resulting in relatively stable boundary knot positions among the simulation runs. Moreover,

Table 1. Mean (SE) of RASEs for the spline estimates of  $a_1$  and  $a_2$  with different criteria for selecting the number of knots

Function	AIC	AIC <sub>C</sub>	BIC	MCV
$a_1$	0.077 (0.0021)	0.077 (0.0021)	0.086 (0.0021)	0.098 (0.0028)
$a_2$	0.072 (0.0019)	0.072 (0.0018)	0.080 (0.0021)	0.080 (0.0026)

such placement of boundary knots can also help to enhance the numerical stability when a large number of knots is used. For each simulated data set, the optimal number of knots was selected in the range from 2 to 10 using the three criteria in section 2.4.2. Two boundary knots were included when counting the knot number. As mentioned above, different optimal knot numbers were allowed for coefficient functions  $a_1(\cdot)$  and  $a_2(\cdot)$  in order to adapt to possible different smoothness.

The summary statistics of RASEs of  $a_1(\cdot)$  and  $a_2(\cdot)$  for the spline fits with AIC, AIC<sub>C</sub>, BIC and MCV selected knot numbers are reported in Table 1. The number of grid points in calculating RASE is 240. The performances of AIC and AIC<sub>C</sub> are similar and are better than BIC and MCV. For the MCV criterion, we used  $m = 40$  and  $Q = 4$  as in Cai *et al.* (2000), which gave overall the largest mean RASEs. We also observed that MCV is sensitive to the choice of  $m$  and  $Q$ . Other choices of  $m$  and  $Q$  were tried and did not change our conclusion. Graphical tools have also been used to evaluate the fits. We observed that, while the spline fit with AIC or AIC<sub>C</sub> selected knot number can give almost unbiased estimates, BIC and MCV yield estimates with bigger biases around the modes of the unknown functions. The variances of the estimates are comparable among the methods with MCV having a slightly larger variance.

We have also applied the spline method to other simulated examples and results have been compared with those of the local linear method suggested in Cai *et al.* (2000) (data not shown). We found that the local linear method usually gave estimates with larger bias than those given by the spline method, especially around the modes; moreover, the spline fits are usually smoother than the local linear fits, while the local linear fits tend to show spurious features more often. We also found that, when the coefficient functions have different smoothness, the proposed spline method outperforms the local linear estimator that uses only a single smoothing parameter, which demonstrates the benefit of using different smoothing parameters for different coefficient functions in the proposed method.

## 5. Application to US GNP

We consider the series  $\{z_t\}$  of quarterly US real GNP (in 1982 dollars) from the first quarter of 1947 to the first quarter of 1991, a total of 177 observations. This empirical example is of considerable interest in its own right, following the papers of Tiao & Tsay (1994) and Potter (1995). The data are obtained from the Citibase database and are seasonally adjusted. Prior to analysis, logarithms and first differences of the data were taken and the result was multiplied by 100. Precisely, we use for our analysis the transformed data  $y_t = 100 \log(z_t/z_{t-1})$ , resulting in 176 data points in the series. The goal is to model the dynamic behaviour of US GNP and make out-of-sample predictions. A traditional and simple modelling approach is to fit a linear autoregressive (AR) model to the series  $\{y_t\}$ . However, evidence of non-linearity of US GNP is provided and non-linear models are proposed in the econometrics literature (see e.g. Potter, 1995). Here we would like to illustrate the proposed method by fitting a functional coefficient (FC) model and performing multi-step-ahead (out-of-sample) forecasts based on the fitted model. We observe strong evidence of the forecasting superiority of the non-linear FC model

over linear AR models, which suggests that there exists non-linearity in US GNP that cannot be captured by linear autoregressions.

In order to compare the out-of-sample forecasting performance of different models, we split the series  $\{y_t\}$  into two subseries. We take the subseries  $(y_1, \dots, y_{164})$  as the training set to choose the best AR model and the best FC model. Then we calculate the postsample multi-step-ahead forecast errors on the last 12 observations  $(y_{165}, \dots, y_{176})$ .

Using AIC to select the order of linear AR models, an AR(3) model is chosen and the (least squares) fitted model is  $\hat{Y}_t = 0.508 + 0.342Y_{t-1} + 0.178Y_{t-2} - 0.148Y_{t-3}$ . Standard diagnostics reveal that this model fits the data well.

For the functional coefficient model, we apply the method in section 2.4.3 to select the threshold lag and significant lags. Consider the class of models (5) with  $p_{\max} = 4$ . We select a model specified by the pair  $\{d, S_d\}$  that minimizes AIC in a stepwise procedure. When fitting any of the models in (5) using our spline method, we use quadratic splines with knots equally spaced between two boundary knots that are placed at the 1 and 99 per cent sample quantiles of the threshold variable. For computational simplicity, the same fixed number of knots is used for all coefficient functions in the model selection process. Different numbers of knots are tried and the results are quite stable. For the number of knots being 2–5, the following model is always selected:

$$Y_t = a_1(Y_{t-2})Y_{t-1} + a_2(Y_{t-2})Y_{t-2} + \epsilon_t. \tag{8}$$

The AIC values for the selected models are reported in Table 2. The smallest AIC is achieved when the number of knots equals 3. Diagnostics also suggest the model is a valid model.

The above AR and FC models (with three knots) fitted using the data  $\{y_1, \dots, y_{164}\}$  are employed to make out-of-sample forecasts for one- to 12-steps ahead. To implement the simulation-based forecasting method for the FC model (section 3), 5000 series are simulated using the fitted model (i.e.  $M = 5000$ ); each series has length 12 and starts from the last few observations in the training data. To simulate the error terms in the FC model, we sample with replacement from the centred residuals. (Residual diagnostics do not show evidence of heteroscedasticity.) We have also tried to use non-centred residuals to simulate the error terms and obtained similar results. The (point) forecasts are compared with the actual observed values and the absolute prediction error (APE) is computed. Table 3 shows the relative APE of forecasts for the AR(3) model and the FC model, using the AR(3) model as a benchmark.

Table 2. Results on model selection for US GNP: functional coefficient models

No. of knots	AIC	$d$	$S_d$
2	0.1137	2	1, 2
3	0.0821	2	1, 2
4	0.1059	2	1, 2
5	0.0993	2	1, 2
6	0.1211	1	1, 4

Table 3. Absolute prediction error (APE) of multi-step-ahead forecasts: one series. The first row is the APE of the AR(3) model, the second and third rows are, respectively, relative APEs of FC and ANN models, using the AR(3) model as benchmark

Forecast horizon	1	2	3	4	5	6	7	8	9	10	11	12
AR(3) APE	0.177	0.208	0.134	0.113	0.388	0.372	0.721	0.388	0.702	0.451	1.208	1.459
FC/AR(3)	0.821	1.240	0.969	1.542	0.643	0.565	0.711	0.468	0.697	0.510	0.801	0.844
ANN/AR(3)	0.104	1.423	1.044	0.605	1.046	0.989	1.016	1.018	1.040	1.113	1.074	1.131

It is clear that the FC model outperforms the AR(3) model in all except two steps (step 2 and step 4).

Following a suggestion by one referee, we also fitted an artificial neural network (ANN) model (see Franses & van Dijk, 2000, Chapter 5) to the data. The ANN model, which uses lag 1 and lag 2 variables as inputs, has a single hidden layer with four hidden units and allows a direct link from the inputs to the output. The AIC criterion was used to select the number of hidden units (from 1 to 4) and significant lags (from 1 to 4) in the model. A stepwise procedure similar to what is described in section 2.4.3 was implemented to select the lags. Using the bootstrap method, we computed one- to 12-steps ahead forecasts from the ANN model. The results are summarized in Table 3. It appears that the ANN model was outperformed by the FC model in 10 of the 12 steps.

As explained in section 3, we can use the 5000 simulated series to generate postsample multi-step-ahead interval forecasts and density forecasts. Figure 1 plots the 95 per cent forecast intervals along with the actual growth rates for the 12 quarters in the postsample period. We see that all the observed values are in the corresponding forecast intervals. Multi-step density forecasts are given in Fig. 2. We can also make probabilistic statements using the simulated series. For example, we can answer questions like, ‘What’s the chance of a positive growth rate in the next quarter?’ in five quarters?’ Table 4 gives the estimated probability for a positive growth rate for each of the next 12 quarters.

In the above comparison the forecast performance is evaluated only on one series. Following Tiao & Tsay (1994), we now apply a rolling procedure to assess the average forecast performance of the linear AR model and the non-linear FC model over many series. To be specific, consider 60 subseries  $(y_1, \dots, y_T)$ ,  $T = 105, \dots, 164$ . For each subseries, we fit an AR model and an FC model with the order of the AR model and the threshold lag and significant lags of the FC model being selected using the data. The AIC is used as the model selection criterion. When fitting the FC model, we use three equally spaced knots with the boundary knots at the 1 per cent and 99 per cent sample quantiles of the threshold variable. The result of the model selection is quite stable; AR(3) or AR(4) is always selected for all the subseries, and

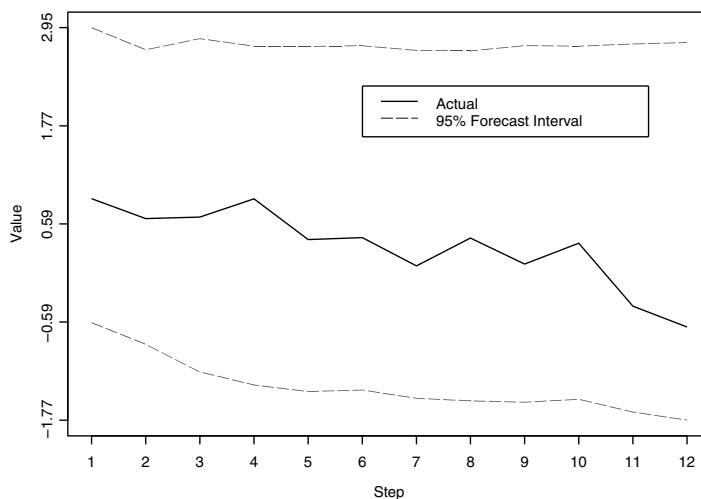


Fig. 1. One- to 12-steps ahead interval forecasts for US GNP. The solid line is the actual growth rate of the last 12 quarters, the dotted lines are the lower and upper limits of the bootstrapped 95 per cent multi-step-ahead forecast intervals obtained from the fitted FC model.

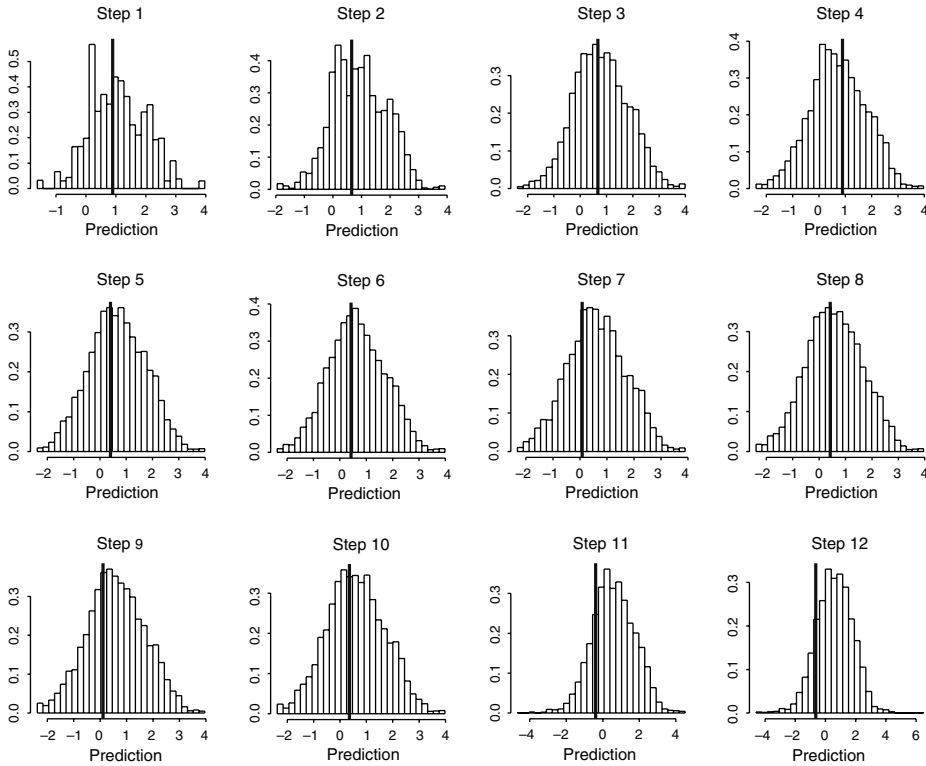


Fig. 2. One- to 12-steps density forecasts for US GNP. The vertical highlighted lines are the real growth rates; the histograms are the multi-step-ahead density forecasts based on the fitted FC model.

Table 4. Estimated probability for a positive growth rate

Period	1	2	3	4	5	6	7	8	9	10	11	12
Probability	0.89	0.84	0.78	0.76	0.73	0.72	0.72	0.70	0.71	0.70	0.69	0.69

the FC model always chooses  $y_{t-2}$  as the threshold variable and  $y_{t-1}, y_{t-2}$  as the significant variables.

Then, for each subseries, multi-step-ahead postsample forecasts up to 12 steps are computed based on the selected models. We use 5000 simulated series when implementing the forecasting procedure given in section 3. Mean squared prediction error (MSPE) is calculated for each forecasting step. Table 5 shows the relative MSPE of forecasts for the AR model and the FC model, using the AR model as a benchmark. We find gains from using the non-linear FC model at short forecasting horizons. At 2-, 3-, 4- and 5-steps ahead, the gains over the AR

Table 5. Mean squared prediction error (MSPE) of multi-step forecasts: 60 series. The first row is the MSPE of the AR model, the second and third rows are, respectively, relative MSPEs of FC and ANN models, using the AR model as benchmark

Forecast horizon	1	2	3	4	5	6	7	8	9	10	11	12
AR MSPE	1.064	1.181	1.254	1.229	1.224	1.145	1.045	0.891	0.896	0.890	0.905	0.934
FC/AR	0.999	0.869	0.902	0.913	0.938	0.955	1.001	1.024	1.013	1.005	0.991	0.963
ANN/AR	1.114	0.970	0.976	0.966	1.000	1.046	0.775	1.138	1.228	1.230	1.555	2.866

model are 13.1, 9.8, 8.7 and 6.2 per cent, respectively. At longer horizons there is no clear difference between the FC and AR models. A similar empirical forecasting exercise has been done on the same data set in the literature using the self-exciting threshold autoregressive (SETAR) model. Clements & Smith (1997) observe that the SETAR model yields inferior forecasts to the AR model for horizons up to four-steps ahead. At 1-, 2-, 3- and 4-steps ahead, the ratios of MSPE of SETAR to that of AR are approximately 1.02, 1.01, 1.07 and 1.02 (see Fig. 2 of Clements & Smith, 1997, p. 473). Table 5 also reports the results of a similar rolling forecasting exercise for ANN models. (The ANN models are selected for each subseries using AIC and a stepwise procedure similar to that described in section 2.4.3. In calculating the MSPEs for the ANN models, we have deleted three series where the ANN models produce very high MSPEs.) It appears that the FC models outperform the ANN models in 11 of the 12 steps.

**6. Application to Dutch guilder–US dollar exchange rate**

As the second example, we consider the Dutch guilder–US dollar exchange rate, expressed as the number of units of Dutch guilder per US dollar. The sample period runs from 2 January 1980 to 31 December 1997. We analyse the series on a weekly basis, using observations recorded on Wednesdays. (There are 10 observations on Wednesdays in the sample period that are not available. In these cases we use the observation from the following Thursday, and if the Thursday observation is also unavailable, we use the observation from the preceding Tuesday.) This series was used extensively in Franses & van Dijk (2000) to illustrate applications of non-linear time series models.

As an illustration of our method, we consider the following functional coefficient model:

$$Y_t = a_1(V_{t-1})Y_{t-1} + a_2(V_{t-1})Y_{t-2} + \epsilon_t, \tag{9}$$

where  $Y_t = 100 \log (P_t/P_{t-1})$  ( $P_t$  denotes the exchange rate at time period  $t$ ) is the logarithmic return measured in percentage, and  $V_{t-1} = \sum_{j=1}^4 |Y_{t-j}|$  is the average absolute return over the last four weeks, which can be considered as a measure of volatility. This model is similar to the two-regime smooth transition AR (STAR) model considered in Chapter 3 of Franses & van Dijk (2000), except that in the STAR model, the coefficient functions have a particular parametric form. We fitted model (9) to the data for the years 1980–9 using the proposed spline method. Quadratic splines with three equally spaced knots were used where the boundary knots were placed at the 1 and 99 per cent sample quantiles of the threshold variable  $V_{t-1}$ . Using the method described in section 3, we computed 1- to 5-step forecasts from the fitted model for the years 1990–7. The ratios of the MSPE and median SPE (MedSPE) criteria, relative to an AR(2) model which is used as the benchmark linear model, are given in Table 6. We also present results for the STAR model (Franses & van Dijk, 2000, Table 3.12) and an ANN model in Table 6. The ANN model, which uses  $Y_{t-1}$ ,  $Y_{t-2}$  and  $V_{t-1}$  as input variables, has one hidden layer with five hidden units and allows a direct link from the inputs to the

Table 6. Forecast evaluation of models for weekly returns on the Dutch guilder exchange rate. Reported are the relative MSPEs and MedSPEs of FC, STAR and ANN models, using an AR(2) model as benchmark

Forecast horizon	MSPE					MedSPE				
	1	2	3	4	5	1	2	3	4	5
FC/AR	1.034	1.012	0.997	0.997	1.000	1.024	1.062	0.988	0.975	0.978
STAR/AR	1.032	1.025	1.023	1.009	0.999	0.985	1.099	1.056	1.090	1.043
ANN/AR	1.535	1.276	1.107	1.037	0.999	1.183	1.115	1.105	1.091	1.025

Table 7. Density forecast evaluation of models for weekly returns on the Dutch guilder exchange rate. Reported are the  $p$ -values for five tests. For Berkowitz tests, 'Indep.' tests for independence, and 'Joint' for zero mean, unit variance and independence. For Ljung–Box test, 'Orig.' means test performed on the original  $z_t^*$  series and 'Abs.' means test on the absolute values of  $z_t^*$

	Kolmogorov–Smirnov	Berkowitz		Ljung–Box	
		Indep.	Joint	Orig.	Abs.
AR	0.016	0.881	0.550	0.168	$1.84 \times 10^{-4}$
FC	0.058	0.955	0.784	0.118	0.057
ANN	0.181	0.588	1.000	0.130	$3.46 \times 10^{-4}$

output. The number of hidden units was selected by AIC. The bootstrap method was also used to generate forecasts for the STAR and ANN models. It is seen from the table that the FC model in general outperforms the STAR model and the ANN model. Comparing with the benchmark linear model in terms of MSPE and MedSPE, however, there is not much to be gained in terms of out-of-sample forecasting by using the non-linear models in this example.

We also compared the one-step-ahead density forecasts for the years 1990–7 based on the AR, FC and ANN models fitted using the data from the years 1980 to 1989. The approach we used to evaluate the density forecasts is based on Dawid (1984) and Diebold *et al.* (1998) and was applied in, for example, Clements & Smith (2000). Let  $\{z_t\}_{t=T+1}^{T+n}$  denote the probability integral transforms of the actual realizations  $\{y_t\}$  of the variables over the forecast period with respect to the forecast densities  $p_t(y_t)$ , that is,  $z_t = \int_{-\infty}^{y_t} p_t(u) du$ ,  $t = T + 1, \dots, T + n$ . When the forecast density corresponds to the true predictive density (given by the data generating process), then the sequence of variates  $\{z_t\}_{t=T+1}^{T+n}$  is i.i.d.  $U(0, 1)$ . Hence the idea is to evaluate the forecast density by assessing whether the  $\{z_t\}$  series departs from the i.i.d. uniform distribution. We used the Kolmogorov–Smirnov (KS) test for the i.i.d. uniform hypothesis. Let  $z_t^*$  denote the inverse normal transformation of  $z_t$ . Then, the  $z_t$ s, which are i.i.d.  $U(0, 1)$  under the null, become i.i.d. standard normal variates. Berkowitz (2001) argues that more powerful tools can be applied to testing a null of i.i.d.  $N(0, 1)$  compared with one of i.i.d. uniformity and proposes a one-degree of freedom test of independence against a first-order autoregressive structure, as well as a three-degrees of freedom test of zero-mean, unit variance and independence. These two tests were also implemented in our forecast comparison. We also conducted the Ljung–Box test of serial correlation with 25 lags for the  $\{z_t^*\}$  series and the absolute values of  $\{z_t^*\}$ . The results are summarized in Table 7. Neither of Berkowitz's tests are statistically significant for all models. At the 5 per cent significance level, the KS test reveals departure of  $\{z_t\}$  from i.i.d.  $U(0, 1)$  for the AR model, while the Ljung–Box test on the absolute values of  $\{z_t^*\}$  found serial correlation for the AR and ANN models. The results suggest a comparatively good performance of density forecasts based on the FC model, at least in terms of the statistical tests considered above.

## 7. Concluding remarks

In this paper we propose a global smoothing method using spline functions for estimation of functional coefficient models for non-linear time series. Splines with equally spaced knots are used and the methodology works well in simulations and real data examples. It would be interesting to consider free-knot splines in the current context, i.e. letting the data choose the knot positions as in multivariate adaptive regression splines (Friedman, 1991; Lewis & Stevens, 1991). See Stone *et al.* (1997) and Hansen & Kooperberg (2002) for recent reviews of the free-knot methodology. Some asymptotic theory for free-knot splines for i.i.d. data has been

developed in Stone & Huang (2002). We have focused on the case where the threshold variable  $U_t$  is one-dimensional. When  $U_t$  is multi-dimensional, the proposed method is conceptually applicable, where tensor product splines or general multivariate splines can be used to approximate the coefficient functions  $a_{jt}$ . However, actual implementation may require substantial further development, and models with  $U_t$  having large dimensionality are often not practically useful due to the ‘curse of dimensionality’.

Our basis approximation method can be used to handle vector time series by replacing the objective function (4) by a generalized variance of the vector-valued residuals (we thank one referee for pointing this out to us). Specifically, suppose there are  $L$  time series generated according to

$$Y_{it} = \sum_{j=1}^{d_i} a_{ij}(U_{it})X_{itj} + \epsilon_{it}, \quad t = 1, \dots, n, \quad i = 1, \dots, L,$$

where the  $a_{ij}(\cdot)$ s are unknown functions and the  $\epsilon_{it}$ s are mean 0 errors. The  $U_{it}$ s and  $X_{itj}$ s could include lagged values from all series and exogenous predictors. To fit this model, we could approximate each  $a_{ij}(\cdot)$  by a basis expansion  $a_{ij}(u) \approx \sum_{s=1}^{L_{ij}} \beta_{ijs} B_{ijs}(u) = \mathbf{B}_{ij}^T(u) \boldsymbol{\beta}_{ij}$ , where  $\mathbf{B}_{ij} = (B_{ij1}, \dots, B_{ijL_{ij}})^T$  and  $\boldsymbol{\beta}_{ij} = (\beta_{ij1}, \dots, \beta_{ijL_{ij}})^T$ . Denote  $\boldsymbol{\beta}_i = (\boldsymbol{\beta}_{i1}^T, \dots, \boldsymbol{\beta}_{id_i}^T)^T$ ,  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_L^T)^T$ ,  $\mathbf{U}_{it} = (X_{it1} \mathbf{B}_{i1}^T(U_{it}), \dots, X_{itd_i} \mathbf{B}_{id_i}^T(U_{it}))^T$  and  $\mathbf{U}_t = (\mathbf{U}_{t1}^T, \dots, \mathbf{U}_{tL}^T)^T$ . We then minimize with respect to  $\boldsymbol{\beta}$  the generalized variance

$$\ell(\boldsymbol{\beta}) = \det \left[ \frac{1}{n} \sum_{t=1}^n (\mathbf{Y}_t - \mathbf{U}_t \boldsymbol{\beta})(\mathbf{Y}_t - \mathbf{U}_t \boldsymbol{\beta})^T \right]$$

and thus get an estimate of  $a_{ij}(\cdot)$  through the basis expansions. We shall explore the practical applications of this multivariate extension of our method further in our future research. It would be useful to point out that our basis approximation method is generally applicable to other structured non-parametric models for multivariate time series such as additive models, where unknown non-linear transformations of predictors enter the right-hand side of the model equation in an additive manner.

**Acknowledgements**

We wish to thank the Editor, an Associate Editor, two referees, and Paul Shaman for their helpful comments. We also wish to thank Zongwu Cai for providing us his S-PLUS codes for implementing the local linear method of Cai *et al.* (2000).

**References**

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19**, 716–723.  
 Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *J. Bus. Econ. Statist.* **19**, 465–474.  
 de Boor, C. (1978). *A practical guide to splines*. Springer, New York.  
 Bosq, D. (1998). *Nonparametric statistics for stochastic processes: estimation and prediction*, 2nd edn. Springer-Verlag, Berlin.  
 Cai, Z. & Masry, E. (2000). Nonparametric estimation of additive nonlinear ARX time series: local linear fitting and projections. *Econ. Theory* **16**, 465–501.  
 Cai, Z., Fan, J. & Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *J. Amer. Statist. Assoc.* **95**, 941–956.  
 Chen, R. & Liu, L.-M. (2001). Functional coefficient autoregressive models: estimation and tests of hypotheses. *J. Time Ser. Anal.* **22**, 151–173.

- Chen, R. & Tsay, R. S. (1993). Functional coefficient autoregressive models. *J. Amer. Statist. Assoc.* **88**, 298–308.
- Chiang, C.-T., Rice, J. & Wu, C. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *J. Amer. Statist. Assoc.* **96**, 605–619.
- Clements, M. P. & Smith, J. (1997). The performance of alternative forecasting methods for SETAR models. *Int. J. Forecasting* **13**, 463–475.
- Clements, M. P. & Smith, J. (2000). Evaluating the forecast densities of linear and non-linear models: applications to output growth and unemployment. *J. Forecasting* **19**, 255–276.
- Dawid, A. P. (1984). Statistical theory: the prequential approach. *J. Roy. Statist. Soc. Ser. A* **147**, 278–292.
- DeVore R. A. & Lorentz, G. G. (1993). *Constructive approximation*. Springer-Verlag, Berlin.
- Diebold, F. X., Gunther, T. A. & Tay, A. S. (1998). Evaluating density forecasts: with applications to financial risk management. *Int. Econ. Rev.* **39** 863–883.
- Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall, New York.
- Fan, J. & Gijbels, I. (1996). *Local polynomial modeling and its applications*. Chapman & Hall, New York.
- Fan, J. & Zhang, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27**, 1491–1518.
- Fan, J. & Zhang, J.-T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *J. Roy. Statist. Soc. Ser. B* **62**, 303–322.
- Franses, P. H. & van Dijk, D. (2000). *Nonlinear time series models in empirical finance*. Cambridge University Press, Cambridge.
- Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19**, 1–141.
- Hansen, M. H. & Kooperberg, C. (2002). Spline adaptation in extended linear models (with discussion). *Statist. Sci.* **17**, 2–51.
- Hastie, T. J. & Tibshirani, R. J. (1993). Varying-coefficient models (with discussion). *J. Roy. Statist. Soc. Ser. B* **55**, 757–796.
- Haggan, V. & Ozaki, T. (1981). Modeling nonlinear vibrations using an amplitude-dependent autoregressive time series model. *Biometrika* **68**, 189–196.
- Härdle, W., Lütkepohl, H. & Chen, R. (1997). A review of nonparametric time series analysis. *Int. Statist. Rev.* **65**, 49–72.
- Hoover, D. R., Rice, J. A., Wu, C. O. & Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* **85**, 809–822.
- Huang, J. Z. (1998). Projection estimation in multiple regression with applications to functional ANOVA models. *Ann. Statist.* **26**, 242–272.
- Huang, J. Z. (2001). Concave extended linear modeling: a theoretical synthesis. *Statist. Sinica* **11**, 173–197.
- Huang, J. Z., Wu, C. O. & Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika* **89**, 111–128.
- Hurvich, C. M. & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.
- Lewis, P. A. W. & Stevens, J. G. (1991). Nonlinear modeling of time series using multivariate adaptive regression splines (MARS). *J. Amer. Statist. Assoc.* **86**, 864–877.
- Ozaki, T. (1982). The statistical analysis of perturbed limit cycle processes using non-linear time series models. *J. Time Ser. Anal.* **3**, 29–41.
- Potter, S. M. (1995). A nonlinear approach to US GNP. *J. Appl. Econ.* **10**, 109–125.
- Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461–464.
- Schumaker, L. L. (1981). *Spline functions: basic theory*. Wiley, New York.
- Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22**, 18–184.
- Stone, C. J. & Huang, J. Z. (2002). Free knot splines in concave extended linear modeling. *J. Statist. Plann. Inference* **108**, 219–253.
- Stone, C. J., Hansen, M., Kooperberg, C. & Truong, Y. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *Ann. Statist.* **25**, 1371–1470.
- Tiao, G. C. & Tsay, R. S. (1994). Some advances in non-linear and adaptive modelling in time-series. *J. Forecasting* **13**, 109–131.
- Tjøstheim, D. & Auestad, B. H. (1994). Nonparametric identification of nonlinear time series: projections. *J. Amer. Statist. Assoc.* **89**, 1398–1409.
- Tong, H. (1990). *Nonlinear time series: a dynamical system approach*. Oxford University Press, Oxford.
- Tong, H. (1995). A personal overview of non-linear time series analysis from a chaos perspective (with discussion). *Scand. J. Statist.* **22**, 399–445.

Truong, Y. (1994). Nonparametric time series regression. *Ann. Inst. Statist. Math.* **46**, 279–293.  
 Truong, Y. & Stone, C. J. (1992). Nonparametric function estimation involving time series. *Ann. Statist.* **20**, 77–97.  
 Tschernig, R. & Yang, L. J. (2000). Nonparametric lag selection for time series. *J. Time Ser. Anal.* **21**, 457–487.  
 Yang, L. J., Härdle, W. & Nielsen, J. P. (1999). Nonparametric autoregression with multiplicative volatility and additive mean. *J. Time Ser. Anal.* **20**, 579–604.  
 Wu, C. O., Chiang, C.-T. & 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.

Received February 2003, in final form February 2004

Jianhua Huang, The Wharton School, Statistics Department, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340, USA.  
 E-mail: jianhua@wharton.upenn.edu

**Appendix**

In the appendix, we give the proof of theorem 1. For two sequences of positive numbers  $a_n$  and  $b_n$ , we write  $a_n \lesssim b_n$  if  $a_n/b_n$  is uniformly bounded and  $a_n \asymp b_n$  if  $a_n \lesssim b_n$  and  $b_n \lesssim a_n$ .

We impose the following conditions:

- (i) The marginal density of  $U_t$  is bounded away from zero and infinity uniformly on  $\mathbb{C}$ .
- (ii) The eigenvalues of  $E(X_t X_t^T | U_t = u)$  are uniformly bounded away from zero and infinity for all  $u \in \mathbb{C}$ .
- (iii)  $K_n \asymp n^r, 0 < r < 1$ .
- (iv) The process  $\{Y_t, X_t, U_t\}_{-\infty}^{\infty}$  is jointly strictly stationary. The  $\alpha$ -mixing coefficient  $\alpha(t)$  of  $\{X_t, U_t, Y_t\}$  satisfies  $\alpha(t) \leq Ct^{-\alpha}$  for  $\alpha > (5/2)r/(1-r)$ .
- (v) For some sufficient large  $m > 0, E(|X_{tj}|^m) < \infty, j = 1, \dots, d$ .

Condition (ii) is necessary for identification of the coefficient functions in our functional coefficient model. Other conditions are commonly used in the literature.

From the function space notation in section 2.3, we see that the spline estimates  $\hat{a}_j$  are uniquely determined by the function spaces  $\mathbb{G}_j$ . Different sets of basis functions can be used to span the spaces  $\mathbb{G}_j$  and thus give the same estimates  $\hat{a}_j$ . We employ the B-spline basis in our proofs for convenience, but the results do not depend on the choice of basis.

For  $j = 1, \dots, d$ , set  $B_{js} = K_j^{1/2} N_{js}, s = 1, \dots, K_j$ , where  $N_{js}$  are B-splines as defined in DeVore & Lorentz (1993, Chapter 5). There are positive constants  $M_1$  and  $M_2$  such that

$$M_1 |\beta_j|^2 \leq \int \left\{ \sum_s \beta_{js} B_{js}(u) \right\}^2 du \leq M_2 |\beta_j|^2, \tag{10}$$

where  $\beta_j = (\beta_{j1}, \dots, \beta_{jK_j})^T$  (see theorem 4.2 of Chapter 5 of DeVore & Lorentz, 1993).

We need the following two lemmas whose proofs are given after the proof of theorem 1.

**Lemma 1**

$$\sup_{a_j \in \mathbb{G}_j, j=1, \dots, d} \left| \frac{\frac{1}{n} \sum_t \left\{ \sum_j a_j(U_t) X_{tj} \right\}^2}{E \left\{ \sum_j a_j(U_t) X_{tj} \right\}^2} - 1 \right| = o_P(1).$$

Let  $\otimes$  be an  $(n \times \sum_j K_j)$ -matrix with the  $t$ -th row being a vector with entries  $B_{js}(U_t) X_{tj}, s = 1, \dots, K_j, j = 1, \dots, d$ .

**Lemma 2**

There is an interval  $[M_1, M_2]$  with  $0 < M_1 < M_2$  such that

$$P\left\{ \text{all the eigenvalues of } \frac{1}{n} \mathbb{X}^T \mathbb{X} \text{ fall in } [M_1, M_2] \right\} \rightarrow 1.$$

*Proof of theorem 1.* Denote  $\mathbb{Y} = (Y_1, \dots, Y_n)^T$ . Then  $\hat{\boldsymbol{\beta}} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Y}$ . Note that  $\hat{a}_j(u) = \sum_s \hat{\beta}_{js} B_{js}(u)$ , where  $\hat{\beta}_{js}$  are components of  $\hat{\boldsymbol{\beta}}$ . Set  $\tilde{\mathbb{Y}} = (\tilde{Y}_1, \dots, \tilde{Y}_n)^T$  with  $\tilde{Y}_t = \sum_j a_j(U_t) X_{tj}$  and define  $\tilde{\boldsymbol{\beta}} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \tilde{\mathbb{Y}}$ . Write  $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^T, \dots, \tilde{\boldsymbol{\beta}}_d^T)^T$ ,  $\tilde{\boldsymbol{\beta}}_j = (\tilde{\beta}_{j1}, \dots, \tilde{\beta}_{jK_j})^T$ ,  $j = 1, \dots, d$ . Let  $\tilde{a}_j(u) = \sum_s \tilde{\beta}_{js} B_{js}(u)$ . In the following we evaluate first the magnitude of  $\|\hat{a}_j - \tilde{a}_j\|_2$  and then that of  $\|\hat{a}_j - a_j\|_2$ .

Denote  $\mathbb{E} = (\epsilon_1, \dots, \epsilon_n)^T$ . Then  $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{E}$ . As  $\epsilon_t$  has mean 0 and is independent of  $U_t, X_{t'j}, j = 1, \dots, d, t' \leq t$ , and  $\epsilon_{t'}, t' < t$ , by conditioning we obtain

$$E\{B_{js}(U_t) X_{tj} \epsilon_t B_{js}(U_{t'}) X_{t'j} \epsilon_{t'}\} = 0, \quad t' \neq t.$$

Hence

$$E(\mathbb{E}^T \mathbb{X} \mathbb{X}^T \mathbb{E}) = E\left[ \sum_j \sum_s \left\{ \sum_t B_{js}(U_t) X_{tj} \epsilon_t \right\}^2 \right] = \sum_j \sum_s \sum_t E\{B_{js}^2(U_t) X_{tj}^2 \epsilon_t^2\} \approx n \sum_j K_j,$$

Therefore,  $\mathbb{E}^T \mathbb{X} \mathbb{X}^T \mathbb{E} = O_P(n \sum_j K_j)$ . Using lemma 2, we have that

$$\|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|^2 = \mathbb{E}^T \mathbb{X} (\mathbb{X}^T \mathbb{X})^{-1} (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{E} \approx \frac{1}{n^2} \mathbb{E}^T \mathbb{X} \mathbb{X}^T \mathbb{E} = O_P\left(\frac{\sum_j K_j}{n}\right),$$

which together with (10) yields

$$\sum_j \|\hat{a}_j - \tilde{a}_j\|_2^2 \asymp \|\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}\|^2 = O_P\left(\frac{\sum_j K_j}{n}\right).$$

Let  $a_j^* \in \mathbb{G}_j$  be such that  $\|a_j^* - a_j\|_2 = \inf_{g \in \mathbb{G}_j} \|g - a_j\|_2 = \rho_{nj}$ . Write  $a_j^*(u) = \sum_s \beta_{js}^* B_{js}(u)$ . Let  $\boldsymbol{\beta}^*$  be a  $\sum_j K_j$ -dimensional vector with entries  $\beta_{js}^*$ ,  $s = 1, \dots, K_j, j = 1, \dots, d$ . By (10) and lemma 2, with probability tending to 1,

$$\sum_j \|\tilde{a}_j - a_j^*\|_2^2 \asymp \|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2 \asymp \frac{1}{n} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \mathbb{X}^T \mathbb{X} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*).$$

As  $\mathbb{X} \tilde{\boldsymbol{\beta}} = \mathbb{X} (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \tilde{\mathbb{Y}}$  is an orthogonal projection,

$$\frac{1}{n} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T \mathbb{X}^T \mathbb{X} (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \leq \frac{1}{n} |\tilde{\mathbb{Y}} - \mathbb{X} \boldsymbol{\beta}^*|^2 = \frac{1}{n} \sum_t \left[ \sum_j \{a_j(U_t) - a_j^*(U_t)\} X_{tj} \right]^2.$$

However, by conditions (i) and (ii),

$$E\left[ \sum_j \{a_j(U_t) - a_j^*(U_t)\} X_{tj} \right]^2 \asymp \sum_j \|a_j - a_j^*\|_2^2,$$

and thus

$$\frac{1}{n} \sum_t \left[ \sum_j \{a_j(U_t) - a_j^*(U_t)\} X_{tj} \right]^2 = O_P\left(\sum_j \|a_j - a_j^*\|_2^2\right).$$

Consequently,

$$\sum_j \|\tilde{a}_j - a_j^*\|_2^2 = O_P\left(\sum_j \|a_j - a_j^*\|_2^2\right) = O_P\left(\sum_j \rho_{nj}^2\right).$$

The proof of theorem 1 is complete.

*Proof of lemma 1.* For a stationary time series  $Z_1, \dots, Z_n, \dots$ , denote  $E_n(Z) = (1/n)\sum_t Z_t$  and  $E(Z) = E(Z_t)$ . For  $a_j \in \mathbb{G}_j$  and  $a_{j'} \in \mathbb{G}_{j'}$ ,  $1 \leq j, j' \leq d$ , write  $a_j = \sum_s \beta_{js} B_{js}$  and  $a_{j'} = \sum_{s'} \beta_{j's'} B_{j's'}$ . Fix  $\eta > 0$ . If

$$|(E_n - E)\{B_{js}(U)B_{j's'}(U)X_j X_{j'}\}| \leq \eta, \quad s = 1, \dots, K_j, \quad s' = 1, \dots, K_{j'}, \tag{11}$$

then

$$\begin{aligned} |(E_n - E)\{a_j(U)a_{j'}(U)X_j X_{j'}\}| &= \left| \sum_s \sum_{s'} \beta_{js} \beta_{j's'} (E_n - E)\{B_{js}(U)B_{j's'}(U)X_j X_{j'}\} \right| \\ &\leq \eta \sum_s \sum_{s'} |\beta_{js}| |\beta_{j's'}| I_{s,s'}, \end{aligned}$$

where  $I_{s,s'}$  equals 1 if the supports of  $B_{js}$  and  $B_{j's'}$  overlap and zero otherwise. By the properties of the B-splines, there is a constant  $C_1$  such that  $\sum_s I_{s,s'} \leq C_1$  and  $\sum_{s'} I_{s,s'} \leq C_1$ . It follows from the Cauchy–Schwarz inequality and (10) that

$$\begin{aligned} \sum_s \sum_{s'} |\beta_{js}| |\beta_{j's'}| I_{s,s'} &\leq \sum_s |\beta_{js}| \left\{ \sum_{s'} \beta_{j's'}^2 I_{s,s'} \right\}^{1/2} C_1^{1/2} \\ &\leq \left\{ \sum_s \beta_{js}^2 \right\}^{1/2} \left\{ \sum_s \sum_{s'} \beta_{j's'}^2 I_{s,s'} \right\}^{1/2} C_1^{1/2} \\ &\leq C_1 \left\{ \sum_s \beta_{js}^2 \right\}^{1/2} \left\{ \sum_{s'} \beta_{j's'}^2 \right\}^{1/2} \leq C_2 \|a_j\|_2 \|a_{j'}\|_2. \end{aligned}$$

Thus, (11) implies that  $|(E_n - E)\{a_j(U)a_{j'}(U)X_j X_{j'}\}| \leq \eta C_2 \|a_j\|_2 \|a_{j'}\|_2$ . Consequently,

$$\begin{aligned} &= P\left\{ \sup_{a_j \in \mathbb{G}_j, a_{j'} \in \mathbb{G}_{j'}} \frac{|(E_n - E)\{a_j(U)a_{j'}(U)X_j X_{j'}\}|}{\|a_j\|_2 \|a_{j'}\|_2} > \eta \right\} \\ &\leq \sum_s \sum_{s'} I_{s,s'} P\left\{ |(E_n - E)\{B_{js}(U)B_{j's'}(U)X_j X_{j'}\}| > \frac{\eta}{C_2} \right\}. \end{aligned}$$

Let  $\tilde{X}_{ij} = X_{ij} I(|X_{ij}| \leq n^\delta)$  for some  $\delta > 0$  and define  $\tilde{X}_{ij}$  similarly. Note that

$$P\{X_{ij} \neq \tilde{X}_{ij} \text{ for some } t = 1, \dots, n\} \leq \sum_t P(|X_{ij}| > n^\delta) \leq \frac{E|X_{ij}|^m}{n^{m\delta-1}} \rightarrow 0,$$

provided  $m > \delta^{-1}$ . Note also that  $\sum_s \sum_{s'} I_{s,s'} \lesssim K_n \approx n^r$ . Thus

$$\ll \lesssim n^r \max_{s,s'} P\left\{ |(E_n - E)\{B_{js}(U)B_{j's'}(U)\tilde{X}_j \tilde{X}_{j'}\}| > \frac{\eta}{C_2} \right\}.$$

Applying theorem 1.4 of Bosq (1998) (with  $q = n^\gamma$  for  $0 < \gamma < 1$ ), we obtain that, for  $k \geq 3$ ,

$$\ll \lesssim n^r \{n^{1-\gamma} \exp(-n^{\gamma-(r+4\delta)}) + n^{\gamma+k(r+2\delta)/(2k+1)-2\alpha(1-\gamma)k/(2k+1)}\},$$

where  $\alpha$  is the parameter in the upper bound of the  $\alpha$ -mixing coefficients [see condition (iii)]. As  $\lim_{\delta \rightarrow 0, \gamma \rightarrow r, k \rightarrow \infty} \{r + \gamma + k(r + 2\delta)/(2k + 1)\} / \{2(1 - \gamma)k/(2k + 1)\} = (5/2)r/(1 - r)$ , for  $\alpha > (5/2)r/(1 - r)$ , we can always find  $\delta > 0$ ,  $\gamma > 0$ , and  $k \geq 3$  satisfying  $1 > \gamma > r + 4\delta$  and  $r + \gamma + k(r + 2\delta)/(2k + 1) - 2\alpha(1 - \gamma)k/(2k + 1) < 0$ . Hence,  $\mathbb{I} \rightarrow 0$  for such a choice of  $\delta$ ,  $\gamma$ , and  $k$ .

Observe that, if  $|(E_n - E)(a_j(U) a_{j'}(U) X_j X_{j'})| \leq \eta \|a_j\|_2 \|a_{j'}\|_2$ ,  $j, j' = 1, \dots, d$ , then

$$\left| (E_n - E) \left\{ \sum_j a_j(U) X_j \right\}^2 \right| \leq \eta \sum_j \sum_{j'} \|a_j\|_2 \|a_{j'}\|_2 = \eta \left\{ \sum_j \|a_j\|_2 \right\}^2 \approx \eta \sum_j \|a_j\|_2^2.$$

Therefore, the desired result follows from the fact that  $\mathbb{I} \rightarrow 0$ .

*Proof of lemma 2.* Let  $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_d^T)^T$ ,  $\beta_j = (\beta_{j1}, \dots, \beta_{jk})$ ,  $j = 1, \dots, d$ . It follows from lemma 1 that

$$\frac{1}{n} \boldsymbol{\beta}^T \otimes \mathbb{I} \otimes \boldsymbol{\beta} = \frac{1}{n} \sum_t \left\{ \sum_j a_j(U_t; \boldsymbol{\beta}) X_{tj} \right\}^2 \asymp E \left\{ \sum_j a_j(U; \boldsymbol{\beta}) X_{tj} \right\}^2,$$

where  $a_j(u; \boldsymbol{\beta}) = \sum_s \beta_{js} B_{js}(u)$ . By (2), conditions (i) and (ii), and (10),

$$E \left\{ \sum_j a_j(U; \boldsymbol{\beta}) X_{tj} \right\}^2 \asymp \sum_j \|a_j(\cdot; \boldsymbol{\beta})\|_2^2 \asymp |\boldsymbol{\beta}|^2.$$

Thus,  $(1/n) \boldsymbol{\beta}^T \otimes \mathbb{I} \otimes \boldsymbol{\beta} \asymp |\boldsymbol{\beta}|^2$  holds uniformly for all  $\boldsymbol{\beta}$ , which yields the desired result.