Spline Estimation of Functional Coefficient Regression Models for Nonlinear Time Series with Correlated Errors

Michel H. Montoril, Pedro A. Morettin, and Chang Chiann

Department of Statistics
Institute of Mathematics and Statistics
University of São Paulo, Brazil
Corresponding author: Chang Chiann, e-mail: chang@ime.usp.br

Abstract

Many data in applications exhibit nonlinear features such as nonlinearity between lagged variables and heteroscedasticity. They require nonlinear models to describe the data. Parametric nonlinear time series models provide powerful tools for analyzing nonlinear time series data when the models are correctly specified, thus the choice for the form of a parametric model is very critical. A natural alternative is to use nonparametric methods. One of the interesting nonparametric models to fit nonlinear time series is the well known functional coefficient regression (FCR) model. There are in literature some works related to this model, with different approaches of estimation (e.g., kernel estimation, spline estimation). A very common supposition of the model is related to the errors, where they are natural to be supposed independent. In this work we will study the estimation of FCR model by splines, with dependent errors. The comparison of the rate of convergence between the models with correlated and independent errors will be done. Moreover, a real application will illustrate the method by fitting a model and performing forecasts. We will compare our results with others used in literature.

Keywords: Polynomial splines, nonlinear time series, absolute prediction error.

1 Introduction

Functional coefficient regression (FCR) models have been attracted interest of many researchers, specially because of their flexibility. Such models generalize both parametric and nonparametric models. Some other examples can be found in Cai et al. (2000). Let $\{Y_t, U_t, \boldsymbol{X}_t\}$ be a jointly strictly stationary process, where U_t is a real random variable and $\boldsymbol{X}_t$ a random vector in $\mathbb{R}^d$. Let $\mathbb{E}(Y_t^2) < \infty$. Considering the multivariate regression function $m(\boldsymbol{x}, u) = \mathbb{E}(Y_t|\boldsymbol{X}_t = \boldsymbol{x}, U_t = u)$, the FCR model has the form

$$m(\boldsymbol{x}, u) = \sum_{i=1}^{d} f_j(u)x_i, \tag{1}$$

where the $f_j(\cdot)$s are measurable functions from $\mathbb{R}$ to $\mathbb{R}$ and $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$, with $\top$ denoting the transpose of a matrix or vector.

Usually, the works that study this model suppose that the errors are independent (see, for example, Cai et al., 2000; Huang and Shen, 2004). However, there are data sets that do not satisfy such assumption, as it is possible to observe in residual analysis. This kind of problem motivates us to study FCR models under perspective of correlated errors. In this work, we use polynomial splines to fit FCR model, as made by Huang and Shen (2004). Actually, our goal is to generalize the results of the authors in the sense that independence of the errors is not necessary to be supposed. We will work with the assumption of correlated errors.

This work is organized as follows. In section 2, based on ideas of Huang and Shen (2004), we generalize the estimator for the coefficient functions, in the sense that correlated errors can be considered. Consistence and rates of convergence in such circumstances are presented. In section 3 some results of a simulation study are reported and a data set (Wolf's annual sunspot number) is studied, where we compute forecasts and compare to other approach of FCR models. In section 4 we give some concluding remarks.

2 Estimation

Following the arguments of Huang and Shen (2004), for some spline basis $\{B_{js}(\cdot)\}$, it is possible to find a vector of constants $\boldsymbol{\beta}_j^* = (\beta_{j1}^*, \ldots, \beta_{jK_j}^*)^\top$, $s = 1, 2, \ldots, K_j$, such that the coefficient function of the FCR model can approximated by a spline function, i.e.,

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

Then, it is possible to approximate (1) by

$$m(\boldsymbol{x}, u) \approx \sum_{j=1}^{d} \sum_{s=1}^{K_j} \beta_{js}^* B_{js}(u) x_i. \tag{2}$$

Based on the approximation (2) it is possible to estimate the coefficients of the spline functions and, thus, estimate the coefficient functions of the model (1).Once estimated the coefficients of the spline functions, a natural estimator of the coefficient functions is

$$\hat{f}_j(u) = \sum_{s=1}^{K_j} \hat{\beta}_{js} B_{js}(u),$$

where $\hat{\boldsymbol{\beta}}_j = (\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jK_j})^\top$, $j = 1, 2, \ldots, d$, is the estimator of $\boldsymbol{\beta}_j^*$.

As we mentioned before, some data sets cannot satisfy the assumption of independence for the errors of the model. Thus, denoting by $\boldsymbol{\Sigma}$ the covariance matrix of the error vector, and supposing initially that it is known, it is possible to estimate the coefficient vector of the spline functions minimizing the weighted least squares function

$$\ell(\boldsymbol{\beta}) = (\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}), \tag{3}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_d^\top)^\top$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ and the t-th row of $\mathbb{X}$ corresponds to the vector $B_{js}(U_t)X_{tj}$, $s = 1, 2, \ldots, K_j$, $j = 1, \ldots, d$. Hence, the coefficient vector estimator corresponds to

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}^\top \boldsymbol{\Sigma}^{-1} \mathbb{X})^{-1} \mathbb{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{Y}. \tag{4}$$

When the errors are independent, $\boldsymbol{\Sigma}$ is a diagonal matrix and, since the variances are bounded away from zero and infinity, (4) has the same rate of convergence as the ordinary least squares estimator, proposed by Huang and Shen (2004). With similar assumptions used by the authors, we derived rates of convergence for distances between the estimators and the real functions. These rates of convergence are presented here as a theorem, which generalizes the one presented by Huang and Shen (2004), in the sense that we verify consistence in quadratic mean, while the authors verify consistence in probability. Moreover, we do not need suppose independence for the errors. Before to show such theorem, it is worth to present firstly the necessary assumptions and explain that the notation $\asymp$, used here, means same rate of convergence.

Assumptions.

(S0) The eigenvalues of $\boldsymbol{\Sigma}$ are bounded away from zero and infinity;

(S1) The marginal density of U_t is bounded away from zero and infinity uniformly on $\mathcal{C}$;

(S2) The eigenvalues of $\mathbb{E}(\boldsymbol{X}_t \boldsymbol{X}_t^\top | U_t = u)$ are uniformly bounded away from zero and infinity for all $u \in \mathcal{C}$;

(S3) $K_j \asymp n^r$, $0 < r < 1$, $j = 1, \ldots, d$;

(S4) The process $\{Y_t, \boldsymbol{X}_t, U_t\}_{t \in \mathbb{Z}}$ is jointly strictly stationary. The α-mixing coefficient $\alpha(t)$ of $\{Y_t, \boldsymbol{X}_t, U_t\}_{t \in \mathbb{Z}}$ satisfies $\alpha(t) \leq Ct^{-\alpha}$ for $\alpha > (2 + r)/(1 - r)$;

(S5) For some sufficient large $m > 0$, $\mathbb{E}|X_{ti}|^m < \infty$, $j = 1, \ldots, d$.

Theorem 1 *If the suppositions above hold, then*

$$\mathbb{E}\|\hat{f}_j - f_j\|_2^2 = O\left(\frac{K_n}{n} + \rho_n^2\right), \quad j = 1, \ldots, d.$$

In particular, if $\rho_n = o(1)$, then $\hat{f}_j$ é consistent in estimating f_j, i.e., $\mathbb{E}\|\hat{f}_j - f_j\|_2^2 = o(1)$, $j = 1, \ldots, d$.

The rate of convergence in theorem above is interest, but in practical situations the covariance matrix $\boldsymbol{\Sigma}$ is unknown. Thus, it become necessary to estimate this matrix, say $\hat{\boldsymbol{\Sigma}}$, and then estimate the coefficient vector $\boldsymbol{\beta}$, say

$$\tilde{\boldsymbol{\beta}} = (\mathbb{X}^\top \hat{\boldsymbol{\Sigma}}^{-1} \mathbb{X})^{-1} \mathbb{X}^\top \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{Y}. \tag{5}$$

Note that since the estimator of the covariance matrix be consistent in probability, i.e., all the eigenvalues of $\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma} - \boldsymbol{I}$ are $o_p(1)$, where $\boldsymbol{I}$ is a identity matrix, it is possible to find that

$$|\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}|^2 = o_p(1). \tag{6}$$

Thus, denoting by $\tilde{f}_j(u) = \sum_{s=1}^{K_j} \tilde{\beta}_{js} B_{js}(u)$, $j = 1, \ldots, d$, the coefficient function estimators, based on (6), another generalization of Theorem 1 from Huang and Shen (2004) is presented in proposition bellow.

Proposition 1 *If assumptions (S0) – (S5) hold, with $\hat{\boldsymbol{\Sigma}}$ consistent in probability in estimating $\boldsymbol{\Sigma}$, then*

$$\|\tilde{f}_j - f_j\|_2^2 = O_p\left(\frac{K_n}{n} + \rho_n^2\right), \quad j = 1, \ldots, d.$$

In particular, if $\rho_n = o(1)$, then $\tilde{f}_j$ is consistent in probability in estimating f_j, i.e., $\|\tilde{f}_j - f_j\|_2 = o_p(1)$, $j = 1, \ldots, d$.

Now, the problem is that $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$ are $(n \times n)$ matrices and, in general, consistency may not be achieved. It is possible, however, to suppose that $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$, i.e., the covariance matrix is a function of a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$, with p being a fixed number. Usually, one suppose that the errors of the model can be represented as an autoregressive processes. Thus, since the coefficient function estimators are consistent, the residuals are good predictors of errors and, hence, it is possible to find a consistent estimator to $\boldsymbol{\theta}$, say $\hat{\boldsymbol{\theta}}$, and hence obtain a consistent estimator for the covariance matrix, $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$.

The estimation procedure can be made iteratively, borrowing some ideas of Cochrane and Orcutt (1949). First of all, one can estimate the coefficient vector $\boldsymbol{\beta}$ acting as if the errors were independent and, then, computing the residuals. Next, an autoregressive model could be fitted to the residuals and, by the estimative of the autoregressive coefficients, the covariance matrix could be estimated. In the following, using the estimate of covariance matrix, the coefficient vector $\tilde{\boldsymbol{\beta}}$ could be computed by (5). This double stage procedure (computation of $\hat{\boldsymbol{\Sigma}}$ and $\tilde{\boldsymbol{\beta}}$) can be repeated until some convergence criterion be achieved (e.g., the residual mean square does not vary, except for a very small value that can be considered insignificant).

The estimates of the procedure above mentioned provide a good approximation to the minimum of (3). Based on this approximation, we can think in an analogous procedure, that is the one we chose to adopt in this work. One could rewrite (3) in terms of backshift operators, i.e., since we are under the assumption that the errors are autoregressive, we could be interested in the minimization of the white noise variance. In other words, denoting by $\boldsymbol{\eta}$ the vector $(\boldsymbol{\beta}^\top, \boldsymbol{\phi}^\top)^\top$ and $\mathbb{x}_t$ as the t-th row of $\mathbb{X}$, we estimate jointly the coefficients of the FCR model $\boldsymbol{\beta}$ and the autoregressive coefficients $\boldsymbol{\phi}$ minimizing numerically

$$\ell(\boldsymbol{\eta}) = \sum_{t=1}^{n} \left\{ \phi_p(L) \left(Y_t - \mathbb{x}_t^\top \boldsymbol{\beta} \right) \right\}^2, \tag{7}$$

where $\phi_p(L) = 1 - \phi_1 L - \ldots - \phi_p L^p$, with the backshift L satisfying $L^k V_t = V_{t-k}$ for all $k > 0$. Since the autoregressive order for the errors is unknown, one can compute the residuals by ordinary least squares and fit an AR model. Doing this, it is possible to use the initial estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ as initial values to be used in the minimization of (7). The algorithm bellow can help in this kind of fitting.

Algorithm for estimating the coefficient vector

(1) Estimate the coefficient vector $\boldsymbol{\beta}$ by ordinary least squares, and denote it by $\tilde{\boldsymbol{\beta}}$;

(2) Fit an autoregressive model to residuals of step (1), i.e., $\tilde{\epsilon}_t = Y_t - \mathbb{x}_t^\top \tilde{\boldsymbol{\beta}}$, say,

$$\tilde{\phi}_p(L)\tilde{\epsilon}_t = \tilde{\varepsilon}_t;$$

(3) Estimate $\boldsymbol{\eta}$ numerically, minimizing (7), using the estimates in steps (1) and (2) as initial values.

2.1 Selection of knots

In the estimation of functions by spline basis is important to pay attention in two questions: how to place and how much knots to be used. There are some techniques about the placement of the knots, and three of them are knots equally spaced, knots organized according to sample quantiles of the random variable and the approach named by free-knots (see, e.g., Hansen and Kooperberg, 2002, for more details about this methodology). For simplicity, and since it was provided good results in Huang and Shen (2004), we will employ here the equally spaced knots approach. About the number of knots to be used, as in Huang and Shen (2004), we will employ three criterion functions, which are AIC, AICc and BIC. Denoting the sample size by n, the number of parameters estimated by p and the residual mean square by RMS, they can be defined as

$$\text{AIC} = \log(\text{RMS}) + \frac{2p}{n}, \quad \text{AICc} = \text{AIC} + \frac{2(p+1)(p+2)}{n(n-p-2)} \quad \text{e} \quad \text{BIC} = \log(\text{RMS}) + \frac{p}{2}\log(n).$$

3 Numerical results

In this section, we will evaluate the performance of the coefficient function estimator presented in section 2. Here, we will study how these estimators behave to finite sample sizes, in a Monte Carlo simulation study and in a real data set application.

3.1 Simulation study

In order to evaluate the performance of the three criteria presented above, we use the square-root of average squared error (RASE), which is defined as

$$\text{RASE}^2 = \sum_{i=1}^{d} \text{RASE}_i^2, \quad \text{with} \quad \text{RASE}_i = \left\{ n_{\text{grid}}^{-1} \sum_{k=1}^{n_{\text{grid}}} \left[\hat{f}_i(u_k) - f_i(u_k) \right]^2 \right\}^{1/2},$$

where $\{u_k, k = 1, \ldots, n_{\text{grid}}\}$ corresponds to a grid of points equally spaced in a interval withing the range of the data set. The RASE has been used in works like Cai et al. (2000) and Huang and Shen (2004), but the grid points chose in these works are different. We decided to select the maximum of the 2.5 percentiles of the data sets to be the left boundary and the minimum of the 97.5 percentiles of the data sets to be the right boundary, as made by Huang and Shen (2004). Doing this, it is possible to eliminate the variation of the range in each data set.

The model simulated is the EXPAR (Haggan and Ozaki, 1981; Cai et al., 2000; Huang and Shen, 2004)

$$Y_t = f_1(Y_{t-1})Y_{t-1} + f_2(Y_{t-1})Y_{t-2} + \epsilon_t,$$

where $f_1(u) = 0.138 + (0.316 + 0.982u)e^{-3.89u^2}$, $f_2(u) = -0.437 + (0.659 + 1.260u)e^{-3.89u^2}$ and ϵ_t is an autoregressive model with coefficient $\theta = 0.6$. The white noise corresponds to i.i.d. $N(0; 0.16^2)$. The series, with length 400, is replicated 10,000 times. Quadratic and cubic spline functions are used in order to estimate the coefficient functions f_1 and f_2. For sake of simplicity, we have used the same spline basis in the estimation.

We used knots equally spaced with the boundary knots placed as the range of the threshold variable Y_{t-1} for each data set. We could choose boundary knots corresponding to sample percentiles of Y_{t-1} for a given data set and

use the same equally spaced knots in the simulation runs. However, since in practical situations the choice of knots depends on the data, we have decided to let the knots varying in each simulated data. The number of knots is selected according to the criteria AIC, AICc and BIC, presented in section 2, in a range from 2 to 10 (including the boundary knots). For sake of simplicity, the same number of knots is used in estimating f_1 and f_2. The sample mean (standard error) of RASE^2, for each criterion, is shown in Table 1, where we can see that in both cases (quadratic and cubic splines) AIC and AICc are providing the best results. This indicates that when the error of the FCR model has a correlation structure (autoregressive in this case) the AIC and AICc criteria still provide better results of fitting than BIC, as happened when independent errors were considered (Huang and Shen, 2004).

Table 1: *Sample mean (SE) of RASE^2, for quadratic and cubic splines, using the criteria AIC, AICc and BIC in the number of knots selection in the range from 2 to 10.*

Spline	AIC	AICc	BIC
Quadratic	0.006 (0.00005)	0.006 (0.00005)	0.009 (0.00006)
Cubic	0.006 (0.00004)	0.006 (0.00004)	0.009 (0.00006)

3.2 Application to Sunspot data

In this work, we extend the bootstrap forecasting technique presented by Huang and Shen (2004). Here, it is acceptable an autoregressive dependence for the errors. Once fitted the FCR model with autoregressive residuals, it is possible to compute and then bootstrap the predicted white noise, which are uncorrelated and homoscedastic. Thus, one can forecast the residuals and hence forecast the time series.

A very interesting and extensively studied data set corresponds to the Wolf's annual sunspot number (Ghaddar and Tong, 1981; Tong, 1983; Chen and Tsay, 1993; Cai et al., 2000), which is known to be a challenging (see Tong, 1993). Calling the series as X_t, the convention in literature suggests the transformation $Y_t = 2(\sqrt{X_t + 1} - 1)$.

The interest here is to fit a FAR model and then forecast some values to compare to other well-known models in literature, as the approach of Chen and Tsay (1993), that fit the model making use of arranged local regression, and of Cai et al. (2000), that use local linear regression method. In these two works was fitted a FAR model using the data on the years 1700-1979, corresponding the FAR of the form

$$Y_t = f_1(Y_{t-3})Y_{t-1} + f_2(Y_{t-3})Y_{t-2} + f_3(Y_{t-3})Y_{t-3} + f_6(Y_{t-3})Y_{t-6} + f_8(Y_{t-3})Y_{t-8} + \epsilon_t, \quad (8)$$

under the assumption of independent errors. After the fitting, it was computed the absolute prediction error (APE) based on the forecasts for the years 1980–1987. Thus, to compare our results, we also fitted the same model to the data in the years 1700–1979, and then forecast the values in the years 1980–1987 to compute the APE, as it was done by these authors. In our fitting, were tested quadratic and cubic splines, with knots in a range from 2 to 6. The AIC criterion has selected a cubic basis, with 4 knots, to estimate the coefficient functions. In a residual analysis (ACF, PACF and Ljung-Box tests), it was indicated that could be interesting to fit the model under assumption of autoregressive errors. The indicated order was of AR(8) and the resulting estimates suggested

$$\epsilon_t = -0.087\epsilon_{t-1} - 0.161\epsilon_{t-2} - 0.231\epsilon_{t-3} + 0.035\epsilon_{t-4} - 0.063\epsilon_{t-5} - 0.184\epsilon_{t-6} + 0.094\epsilon_{t-7} - 0.124\epsilon_{t-8} + \varepsilon_t, \quad (9)$$

with ε_t corresponding to a white noise.

With the fitted model it was possible to compute the forecasts and thus the APEs.The comparisons of forecasts were done by relative APE, as can be seen in Table 2. It is important to mention that the APEs from the models fitted by Chen and Tsay (1993) (denoted here by (CT93)) and Cai et al. (2000) (denoted here by (CFY00)) were picked from Cai et al. (2000). We replicated our forecasts 5,000 times and then computed the point forecasts by using the sample mean. We denote our APEs by (S.AR). It is possible to observe in the table that both the APEs (CFY00) and (S.AR) provide better forecasts than the benchmark (the superiority of the APEs (CFY00) with respect to the benchmark (CT93) has already been discussed in Cai et al. (2000)). In this case, the fit of Cai et al. (2000) has APE outperformed for the benchmark fit in the years 1981, 1982 and 1985, while the APE (S.AR) was not better than the APE of the benchmark model just in the year 1985. This indicates that our procedure provides better forecasts. Moreover, comparing the APEs (CFY00) and (S.AR), it is possible to observe that our approach was able to predict more values closer to the observed values (all of them, but for the year 1986) than the Cai et al. (2000) method.

Table 2: *APE of sunspot data series, for the years 1980–1987. Second column corresponds to the APEs (CT93), used here as benchmark. Tird and fourth columns correspond to the relative APEs (CFY00) and (S.AR).*

Year	APE (CT93)	APE (CFY00) / (CT93)	APE (S.AR) / (CT93)
1980	13.8	0.101	0.035
1981	3.8	2.737	0.084
1982	16.4	1.262	0.109
1983	0.8	0.875	0.030
1984	5.6	0.268	0.065
1985	1.7	2.000	1.479
1986	2.5	0.280	0.402
1987	23.6	0.555	0.098

4 Conclusions and comments

In this work we used polynomial splines, but (quadratic and cubic) B-Splines were used in the simulation study and in the application as well. It is worth to mention that similar results were found in the simulation study, compared to the presented in Table 1. However, to forecast values in the application, B-Splines were not able to perform as good as the quadratic splines. Thus, we believe that in situation of fitting FCR model with autoregressive errors it is more appropriate to use just polynomial splines.

Acknowledgments

The authors are thankful to FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) for the financial support (Proc. 2008/51097-6 and 2009/09588-5).

Bibliography

Cai, Z., J. Fan, and Q. Yao (2000). Functional-Coefficient Regression Models for Nonlinear Time Series. *Journal of the American Statistical Association 95*(451), 941–956.

Chen, R. and R. S. Tsay (1993). Functional-Coefficient autoregressive models. *Journal of the American Statistical Association 88*(421), 298–308.

Cochrane, D. and G. H. Orcutt (1949). Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American Statistical Association 44*(245), 32–61.

Ghaddar, D. K. and H. Tong (1981). Data transformation and Self-Exciting threshold autoregression. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 30*(3), 238–248.

Haggan, V. and T. Ozaki (1981). Modelling nonlinear random vibrations using an Amplitude-Dependent autoregressive time series model. *Biometrika 68*(1), 189–196.

Hansen, M. H. and C. Kooperberg (2002). Spline adaptation in extended linear models (with discussion). *Statistical Science 17*(1), 2–51.

Huang, J. Z. and H. Shen (2004). Functional Coefficient Regression Models for Non-linear Time Series: A Polynomial Spline Approach. *Scandinavian Journal of Statistics 31*(4), 515–534.

Tong, H. (1983). *Threshold Models in Non-linear Time Series Analysis*. Lecture Notes in Statistics. New York: Springer.

Tong, H. (1993). *Non-Linear Time Series: A Dynamical System Approach*, Volume 6 of *Oxford Statistical Science Series*. Oxford: Clarendon Press.