CARNEGIE MELLON UNIVERSITY

NONPARAMETRIC TIME SERIES ANALYSIS USING GAUSSIAN PROCESSES

A DISSERTATION SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

in

STATISTICS

by

SOTIRIOS DAMOURAS

Department of Statistics

Carnegie Mellon University

Pittsburgh, Pennsylvania 15213

May 2008

Abstract

We propose a new approach to nonlinear time series analysis. We build on the functional-coefficient autoregressive model of Chen and Tsay [22] by adopting a Bayesian nonparametric perspective. We treat the coefficient functions as stochastic parameters which are modeled with Gaussian processes. We suggest an empirical Bayes estimation method that is well-suited for the time series models in hand. In particular, our method allows increased modeling flexibility while offering parsimonious results. We compare our method to parametric and nonparametric alternatives and highlight some of its conceptual and practical advantages. We also develop an approximate framework for inference that is computationally efficient and allows us to work with large data sets. Based on this framework, we extend our method to multivariate and state-space models. Moreover, we give some theoretical insights into our approach by proving the consistency of our nonparametric estimates in a frequentist setting. We address further questions of inference by suggesting an integrated model selection procedure and various diagnostics. Finally, we illustrate our methodology with three applications to real data sets in different contexts. Specifically, we present examples on a univariate series, a large bivariate series, and a state-space model, the first coming from natural sciences and the last two from financial econometrics. For all three examples we also present results from competing models, and we demonstrate the improvements that our method can provide.

Contents

Chapter 1

Introduction

This thesis concerns nonlinear time series analysis using nonparametric estimation based on Gaussian Process (GP) regression. The subject of analysis is real-valued series, we do not explicitly aim at discrete observations such as count or categorical data. Nonparametric methods based on GPs have recently attracted a lot of attention from both the statistics and the machine learning community and have found diverse applications in regression and classification, see Rasmussen and Williams [96] for an overview. We concentrate on GP regression, which can be viewed as the Bayesian counterpart of nonparametric estimation techniques such as kernel regression and smoothing splines. We propose a novel methodology for performing time series analysis by extending the range of applications of GPs to this setting. Our treatment of time series with GP regression has significant departures from that of i.i.d. data which is typically found in the literature. Our main contribution lies in customizing the methodology, paying special attention to the nature and characteristics of time series data. In particular, we address practical issues of estimation and prediction, computational efficiency, model selection and diagnostics, as well as the theoretical properties of our model. The significance of all these becomes apparent in subsequent parts of the thesis.

In this chapter, we establish the relevant framework by reviewing existing approaches for time series analysis. First we briefly look at linear time series, since they constitute the foundation of the field; we point out important properties but also their limitations. We move on to nonlinear time series and we define the particular area that we focus on within the broad spectrum of nonlinear behavior. We make a distinction between parametric and nonparametric estimation and we discuss different methods within each class. The methods in the later class are of particular interest because they are directly comparable to ours. Finally, we give an outline of the remaining chapters of the thesis, highlighting the most important points.

1.1 Linear Time Series Models

Linear models play a dominant role in the field of time series analysis, the autoregressive moving average (ARMA) model

$$X_t = \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{j=1}^{p} \theta_j \epsilon_{t-j} + \epsilon_t \tag{1.1}$$

being the most widely used and extensively studied time series tool. Brockwell and Davis [14] give a comprehensive account of linear time series, providing the methodological and theoretical framework around the ARMA model and its equivalent frequency domain analysis. The significance of linear models stems mainly from the powerful Wold representation theorem, which states that every zero-mean, covariance-stationary purely stochastic process can be represented as an infinite order MA process $X_t = \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i}$, where $\{\epsilon_t\}$ is a white noise sequence. Moreover, this infinite MA process can be closely approximated by a finite ARMA process, which can be treated more efficiently for estimation and prediction.

Despite their ease of use and the generality of the previous result, linear models are not a universal solution to all time series problems. For one, the order of the ARMA model might be too high for practical purposes, in cases where nonlinear models can be more succinct. More importantly, the representation specifies the sequence $\{\epsilon_t\}$ only up to second moments, which does not require it to be independent of the past of $\{X_t\}$. Consequently, the best mean square predictor of the series can be quite different from the best linear predictor of the MA representation. The full strength of ARMA models is realized for Gaussian time series which are uniquely determined by their second order properties, i.e. the autocorrelation function, and the same applies to spectral methods. For more general processes, however, the second order properties are not always sufficient and there exists a whole range of behaviors which can not be adequately described by linear models. Examples include the presence of limit cycles, time irreversibility and bistability, among others; more details on nonlinear characteristics of time series are given in Tong [114] and Chen [20].

1.2 Nonlinear Time Series Models

The limitations of linear models have spurred interest in the field of nonlinear time series analysis and during the last couple of decades there has appeared a plethora of new models and estimation techniques. The majority of these are parametric in nature, but recently and with the advent of increased computational power attention is shifting to nonparametric methods. The range of nonlinear models is vast since it encompasses every departure from linearity, but we narrow the scope of our work to models for the conditional mean of the process. For example, we do not consider nonlinear models for the variance, such as the autoregressive conditional heteroskedastic (ARCH) model of Engle [34] and its numerous variants.

In what follows, we give a brief description of the relevant nonlinear models; more details

are given in the review article of Härdle, Lütkepohl and Chen [51] and in Fan and Yao [36]. All of the models we will be looking at fall under the nonlinear autoregressive (NLAR) category

$$X_t = f(X_{t-1}, \ldots, X_{t-p}) + \epsilon_t \tag{1.2}$$

where $\{\epsilon_t\}$ is a white noise sequence, independent of the past of $\{X_t\}$. The function f defines the conditional mean of X_t given the past values $(X_{t-1}, \ldots, X_{t-p})$ of the series; it does not include lagged error terms, in a MA fashion, because they are unobserved and complicate estimation considerably. In the parametric setting, we assume that the function f belongs to a specific class that is characterized by a fixed and finite number of parameters. Choosing a particular parametric form usually requires some knowledge of the characteristics of the data and can lead to modeling biases if the class of functions is too narrow. Another suggestion is to use neural networks that have universal approximation properties, but the resulting models tend to be over-parametrized. For this reason, nonparametric estimation techniques have been proposed which allow f to belong to some flexible class of functions. Three such common techniques are kernel, local polynomial and smoothing spline regression; a description of these can be found in Wasserman [120]. Despite its generality, the model in (1.2) has the significant disadvantage that it suffers from the curse of dimensionality. The term is used to describe the challenges that arise in nonparametric estimation in high dimensions and is a well known phenomenon, see Hastie, Tibshirani and Friedman [58]. Thus, it is helpful to impose a more parsimonious structure on model (1.2).

A popular approach for avoiding the curse of dimensionality in nonparametric regression is to assume an additive form for f. This leads to the generalized additive model (GAM) of Hastie and Tibshirani [55]; in the context of time series, it becomes the nonlinear additive

autoregressive (NLAAR) model

$$X_t = f_1(X_{t-1}) + \ldots + f_p(X_{t-p}) + \epsilon_t. \tag{1.3}$$

Estimation for the NLAAR models is discussed in Chen and Tsay [23] who propose two backfitting algorithms, the alternating conditional expectation (ACE) of Breiman and Friedman [11] and the BRUTO algorithm of Hastie and Tibshirani [55]. Chen and Tsay point out that for time series data which are serially correlated, the performance of backfitting algorithms can be poor. As an improvement on backfitting Masry and Tjøstheim [81] suggest an estimation procedure for the functions based on projections, a similar method for independent data using marginal integration appearing in Linton and Nielsen [75]. Nevertheless, the projections are based on the empirical distribution which can be inaccurate for higher orders and, on top of that, additive models are generally susceptible to non-identifiability.

A nonlinear model which is more favored in time series, and which is also the focus of our work, is the functional-coefficient autoregressive (FAR) model of Chen and Tsay [22]. It is the time series analogue of the varying-coefficient regression model of Hastie and Tibshirani [57] and is given by

$$X_t = f_1(U_t^{(1)})X_{t-1} + \ldots + f_p(U_t^{(p)})X_{t-p} + \epsilon_t. \tag{1.4}$$

Note that the arguments $U_t^{(i)}$ to the functional coefficients f_i are not necessarily equal to X_{t-i}, although they have to be $\mathcal{F}_{t-1}$-measurable w.r.t. the observations and will, in general, depend on lagged values of X_t. Thus, the FAR model (1.4) does not strictly nest the NLAAR model (1.3), in fact it is usually the case that all functional coefficients share the same argument. A considerable advantage of the FAR over the NLAAR model is that it is less prone to non-identifiability, since the functional coefficients f_i are multiplied with

the lagged regressor variables X_{t-i}. Moreover, it retains all the benefits of the NLAAR model, such as parsimony and ease of interpretation. Some of the most popular time series models belong to the FAR family, and we next review their characteristics and estimation.

1.2.1 Parametric Models

We begin our exposition of parametric models by introducing the concept of regimes, which allows us to decompose complex dynamical systems into simpler subsystems. More specifically, we consider piecewise linear models where each regime's dynamics are described by an autoregression. To fix ideas, we present a simple first order autoregressive model with two regimes

$$X_t = \begin{cases} \alpha_0^{(1)} + \alpha_1^{(1)} X_{t-1} + \epsilon_t^{(1)}, & \text{if } I_t = 1 \\ \alpha_0^{(2)} + \alpha_1^{(2)} X_{t-1} + \epsilon_t^{(2)}, & \text{if } I_t = 2 \end{cases} \tag{1.5}$$

where $\{\alpha_0^{(i)}, \alpha_1^{(i)}\}_{i=1,2}$ are the autoregressive coefficients within each regime and $I_t \in \{1, 2\}$ is an $\mathcal{F}_{t-1}$ measurable variable. In terms of the FAR paradigm, the coefficients can be viewed as piecewise linear functions of the common argument variable I_t, where we also permit the means and error variances to change between regimes. The variable I_t which controls the regime is instrumental for the properties and statistical analysis of these models and there are different approaches for specifying it. The most common is to make I_t depend on some lagged value X_{t-d} of the process itself, where X_{t-d} is called the threshold variable and d the time delay. The resulting model is known as the self-exciting threshold autoregressive (SETAR) model or simply as the threshold autoregressive (TAR) model, the latter term sometimes applied to more general settings. It was introduced by Howell Tong who gives a comprehensive survey in his book [114], together with some justification drawn from dynamical systems. The regimes of the TAR model are defined according to the region where the threshold variable X_{t-d} lies, so for the example in (1.5) we would have to select

a set $A \subset \mathbb{R}$ such that $I_t = 1$ if $X_{t-d} \in A$ and $I_t = 2$ otherwise. Notable alternatives to the TAR specification include the Markov switching model of Hamilton [50], where I_t as an independent hidden Markov process, and the smooth transition autoregressive (STAR) model of Teräsvirta [111], which instead of picking a single regime averages the conditional mean dynamics across all regimes and with weights based on X_{t-d}. Formally, the Markov switching model does not belong to the FAR family because I_t is not observable. We will concentrate on the TAR model, since it is the most relevant for our purposes.

It is conceptually easy to extend the TAR model by allowing higher order autoregression, more regimes and more variables for defining the regimes. Higher order autoregression can be dealt with easily, but this is not true for higher number of regimes or, especially, of variables that define them. The reason is that we decide the regime in terms of partitions of $\mathbb{R}^q$, where q is the number of threshold variables, and these can become very complicated as q or the size of the partition grow. For practical applications, TAR models are restricted to a few, say k, regimes and a single threshold variable X_{t-d}. The general form under which TAR models are usually encountered is

$$X_t = \sum_{i=1}^{k} \left\{ \alpha_0^{(i)} + \alpha_1^{(i)} X_{t-1} + \ldots + \alpha_p^{(i)} X_{t-p} + \epsilon_t^{(i)} \right\} I(X_{t-d} \in A_i) \tag{1.6}$$

where $I(\cdot)$ is the indicator function and the sets $\{A_i\}$ of the partition are intervals, $A_i = (r_{i-1}, r_i]$. Sometimes, the autoregressive order p is allowed to vary within each regime. The parameters of interest are the autoregressive coefficients $\{\alpha_j^{(i)}\}$, which are referred to as the autoregressive parameters, and the order p, the number of regimes k, the time delay d and the thresholds $\boldsymbol{r} = \{r_0 = -\infty, r_1, \ldots, r_{k-1}, r_k = \infty\}$, which are referred to as the structural parameters.

Estimation of the autoregressive parameters given knowledge of the structural parameters is

straightforward, using either conditional least squares or conditional maximum likelihood. The real challenge lies in the selecting k, d and estimating the thresholds. The last poses a serious difficulty since both the likelihood and the sum of squares functions are discontinuous w.r.t. $\boldsymbol{r}$, as demonstrated by Tong [114]. Chan [18] treats the case where $k = 1$ and p is known, and shows that least squares estimation for the remaining parameters is consistent, also providing their asymptotic distribution. However, maximization of either objective function w.r.t. $\boldsymbol{r}$ requires a discrete search. Tong and Lim [115] suggest using the quantiles of X_{t-d} as candidates for r_1 in the two regime case and use maximum likelihood, together with AIC for the remaining parameters. Notice that for $k > 2$, the discrete search must over done over ordered sequences. To simplify matters, Tsay [116] proposes a model fitting procedure based on graphical methods. He plots the t-ratio statistic of the coefficients of a two regime TAR model versus candidate thresholds and selects the (possibly more than one) final thresholds by visual inspection for abrupt changes. The delay d is selected by a nonlinearity test, similar to the one of Petrucelli and Davies [93], and other parameters are treated by AIC. Finally, Geweke and Terui [41] look at estimation from a Bayesian perspective using MCMC. In practice, either contextual knowledge or ad hoc choices are often used for defining some of the structural parameters. Despite any issues in fitting, the TAR model has proved successful for applications in various fields and, moreover, it usually affords an easy interpretation.

There are also parametric models that are not based on the regime principle. The most important one, belonging to the FAR class, is the exponential autoregressive (EXPAR) model of Haggan and Ozaki [49]

$$X_t = \sum_{i=1}^{p} \left(\alpha_i + (\beta_i + \gamma_i X_{t-d}) \exp\{-\theta_i X_{t-d}^2\}\right) X_{t-i} + \epsilon_t. \tag{1.7}$$

which is readily estimated by conditional least squares. The EXPAR model was developed

to describe amplitude dependent frequency phenomena typically found in random vibrations, but is not widely used outside this context. For completeness, we also mention two eminent parametric models which depart from the FAR dynamics. The first is the bilinear model of Granger and Anderson (see Subba Rao [95]), which is an ARMA model with additional cross-terms between the AR and MA parts. The second is the random coefficient autoregressive (RCAR) model of Nicholls and Quinn [87], for which the autoregressive coefficients at each time are independent random draws from a fixed distribution. Both have attracted attention in the literature, but we will not dwell on them because they are narrower in scope than the TAR model and not particularly pertinent to the remainder.

1.2.2 Nonparametric Models

We turn our attention to nonparametric estimation, which allows more flexible functional forms for the coefficients in (1.4). A first nonparametric estimation approach was proposed by Chen and Tsay [22], who use the arranged local regression (ALR) procedure. This procedure requires that all functional coefficients share the same argument variable $U_t^{(1)} = \ldots = U_t^{(p)} = U_t$, similarly to the unique threshold variable that defines the regimes in the TAR model. The main idea dates back to Tsay [116] and it involves arranging the data $\left\{X_t, [X_{t-1}, \ldots, X_{t-p}]\right\}$ (viewed as the response and regressor variables) w.r.t. the common argument U_t. In order to estimate the functional coefficients at a particular value U, we put a window around U and perform a linear regression over the arranged data points with U_t falling within that window

$$X_t = a_1 X_{t-1} + \ldots + a_p X_{t-p}; \;\; \forall t \text{ such that } U_t \in [U-h, U+h] \tag{1.8}$$

The estimates of the functional coefficients at U are given by the fitted regression parameters $\hat{f}_i(U) = \hat{a}_i$, for $i = 1, \ldots, p$. The estimation of the function is thus conducted by a series of local regressions, in which the data are arranged with regard to their position relative

to the function's argument variable U_t. This is essentially a variant of kernel regression with a simple boxcar kernel, also known as binning. In its original version, Chen and Tsay actually suggest using both a window and a minimum number of data K for controlling the smoothing. This avoids poor estimation in regions were the data are sparse, such as the boundaries of the range of U_t. For a given data set, the resulting function estimates from the ALR procedure will be step functions, the fitted parameters in the regression change only according to whether data points U_t enter or exit the window. The authors in [22] do not actually work with the nonparametric estimates of the functions, but they use them to infer a parametric functional form for the coefficients. They then use the data again to estimate the parameters of the hypothesized model by least squares. In particular, they do not address the selection of the smoothing parameters h and K, but they advocate repeating the procedure for different values and inspecting the results. They do provide, however, mean square consistency results for the ALR estimated functions.

A fully nonparametric approach, similar in spirit to ALR, was taken up by Cai, Fan and Yao [16], who use local linear regression (LLR). The locality is again induced by a common function argument variable U_t, but the weighting scheme for neighboring observations is different since a non-flat kernel is used. Moreover, the authors use a first order (linear) Taylor approximation of the coefficient function around U_0 in the following fashion

$$f_i(U) \approx a_i + b_i(U - U_0). \tag{1.9}$$

The local linear estimates of the functional coefficients at U are given by $\hat{f}_i(U) = \hat{a}_i$, where $\{\hat{a}_i, \hat{b}_i\}$ are such that they maximize the weighted sum of squares

$$\sum_t \left(X_t - \sum_{i=1}^{p} \Big(a_i + b_i(U_t - U) \Big) X_{t-i} \right)^2 K_h(U_t - U), \tag{1.10}$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$ is a kernel function with bandwidth $h > 0$. The resulting functional coefficient estimates are smooth functions, with the smoothness controlled by h. The authors use the Epanechnikov kernel and suggest a multifold cross-validation procedure for selecting h, which they also extend for the selection of the lag used for U and the autoregressive order. They prove the asymptotic normality of the estimates at a given point and also provide convergence rates.

We point out two limitations of the previous approaches which are both due to the use of weighting techniques for smoothing. First, all of the coefficient functions must share the same argument U_t, and second, all of the estimated functions must share the same smoothing parameter, either the window or the the kernel bandwidth. The reason for this is that estimation is performed locally, based on some weighting scheme, and locality in the data must be defined w.r.t. a common variable. TAR models have a common argument/threshold variable for all coefficients for the same reason; they can be viewed as ALR estimates with fixed windows covering each regime. In practice, these limitations can lead to substantial modeling and estimation disadvantages. It is possible to use backfitting algorithms as in the NLAAR model to relax the common argument restriction, but their performance in time series does not guarantee an improvement. In general, integrated estimation is preferable to profile methods such as backfitting, and its theoretical development is also easier. The problem of common smoothness was acknowledged by Fan and Zhang [64] for i.i.d. data and they proposed a two-step local polynomial regression procedure to rectify it, an initial local linear and a subsequent local cubic. In their approach, both local regressions suffer from the aforementioned problem, but they show that the resulting functional coefficient estimates achieve optimal rates of convergence. Their results do not apply directly to time series, however, so this method was not suggested for the FAR model.

For the varying coefficient model in the regression setting, other estimation approaches have been proposed in order overcome these limitations. Hastie and Tibshirani [57] use smoothing splines, which allow the coefficient functions to have different arguments. They estimate the model by minimizing the least squares plus a smoothness penalty on the functions, where the later can vary between functions. This approach leads to a cubic spline specification for each function f_i, with knots at the locations where each argument is observed. They show that estimating the spline coefficients is computationally expensive in this setting, and fall back to backfitting algorithms for efficiency. Huang and Shen [60] propose a regression spline method for time series which bypasses the need for backfitting. The authors do not use regularization because they control the smoothness of the functions by the number of knots. The resulting representation of the functional coefficients is

$$f_i(U) = \sum_{j=1}^{k_i} \alpha_{i,j} B_{i,j}(U), \quad i = 1, \ldots, p, \tag{1.11}$$

where $B_{i,j}$ are the spline basis functions, $\alpha_{i,j}$ are their parameters and k_i is the number of spline bases used (k_i is directly related to the number of knots). The parameters $\alpha_{i,j}$ are estimated by minimizing the conditional sum of squares

$$\sum_t \left(X_t - \sum_{i=1}^{p} \left(\sum_{j=1}^{k_i} \alpha_{i,j} B_{i,j}(U_t^{(i)}) \right) X_{t-i} \right)^2 \tag{1.12}$$

The authors suggest using equally spaced knots and AIC to decide their number, motivated by the performance of their method in simulation experiments. They also suggest AIC for selecting the argument variables and for identifying the model specification with a greedy stepwise search. They argue that for most applications, the number of required knots will be small, usually less than five for each function, and the resulting estimation procedure will be fast. For their examples, the authors use the same arguments and they sometimes even

use the same number of knots for each function. This practice avoids the need to calculate multiple spline bases and speeds up the procedure even more. On the theoretical side, they show the consistency of their estimates as the number of data and knots increases.

1.3 Outline

The rest of the thesis is organized as follows.

- Chapter 2 presents our approach to Bayesian nonparametric analysis within the FAR framework. We describe estimation and prediction, and provide a systematic way of setting up the prior specification and selecting the hyperparameters. We demonstrate our approach with a real example and comment on qualitative and practical issues.

- Chapter 3 concerns approximate analysis, which is necessary for the computational efficiency of our method. We review the relevant techniques in the literature and adapt them to the requirements of our model. The resulting procedure is outlined and illustrated, and it is also extended to a multivariate and a state-space setting.

- Chapter 4 sheds light on the theoretical aspects of our model. We present its relation to reproducing kernel Hilbert space methods and use it to establish the consistency of the functions' estimators. To this end, we give important conditions for identifiability and ergodicity. We also present theoretical results for the case of approximate estimation.

- Chapter 5 further develops our methodology by addressing model identification and diagnostics. For the former, we use information criteria for creating a greedy model selection procedure, and for the latter, we describe useful residual-based and graphical procedures.

- Chapter 6 contains three applications of our methodology to real data sets in different contexts. We look at the famous sunspot series, a bivariate cointegrated system and

a state-space model for stochastic volatility. We assess the performance and discuss the characteristics of our approach in relation to competing ones.

- Chapter 7 concludes the thesis by synopsizing our work and its main contributions and by identifying directions for future research.

Chapter 2

Proposed Model

This chapter introduces our proposed modeling and estimation approach to nonlinear time series analysis. We essentially build upon the FAR model from a Bayesian nonparametric perspective, using GPs to describe the uncertainty in the functional coefficients. We describe the model and present the resulting estimation and prediction procedures. On the practical side, we discuss the prior specification of our model and focus on empirical Bayes estimation. We give details on the actual implementation through an example using the Canadian lynx data. In the end, we comment on our method and compare it to the relevant alternatives.

2.1 Model Description

In this section we describe the formulation of our model and some of its implications. We begin our exposition by considering the following Markovian FAR model of order p

$$X_t = f_1(U_t^{(1)})X_{t-1} + \ldots + f_p(U_t^{(p)})X_{t-p} + \epsilon_t \tag{2.1}$$

where $\{\epsilon_t\}$ is a white noise sequence and the variables $\{U_t^{(i)}\}_{i=1}^p$ depend on a finite number d of lagged values of X_t, i.e. they are $\sigma(X_{t-1}, \ldots, X_{t-d})$ measurable. According to the

Bayesian paradigm, we put a prior on the functions $\{f_i\}_{i=1}^p$ to describe the uncertainty about them. To this end we use Gaussian Processes, which are stochastic processes whose finite-dimensional distributions are multivariate normal. This is the most convenient and popular approach owing to the conjugacy properties of normals, but also because of the superior theoretical understanding and practical handle that these processes offer. We say that f follows a GP with mean function $\mu(\cdot)$ and covariance function $C(\cdot,\cdot)$, denoted by $f \sim \mathcal{GP}(\mu, C)$, if for every d-dimensional set of indices $\{x_1, \ldots, x_d\}$, the vector of process evaluations $[f(x_1), \ldots, f(x_d)]^\top$ follows a d-dimensional normal distribution with mean vector $\boldsymbol{\mu} = [\mu(x_1), \ldots, \mu(x_d)]^\top$ and covariance matrix $\boldsymbol{C} = \left[\{C(x_i, x_j)\}_{i,j=1}^d\right]$. Thus, the formulation of the FAR model in this setting becomes

$$X_t = f_1(U_t^{(1)})X_{t-1} + \ldots + f_p(U_t^{(p)})X_{t-p} + \epsilon_t \tag{2.2}$$

$$f_i \sim \mathcal{GP}(\mu_i, C_i); \quad i = 1, \ldots, p \tag{2.3}$$

where μ_i is the mean function and C_i is the covariance function of the GP for f_i, and where the functional coefficients $\{f_i\}$ are independent of the error sequence $\{\epsilon_t\}$ and among themselves.

We interpret model (2.2-2.3) as follows: first, the functional coefficients f_i are drawn independently from (2.3) and then the series $\{X_t\}$ is evolved so that it satisfies the usual FAR dynamics in (2.2), given the function draws. This interpretation of the model as a data generating mechanism has the disadvantage that it is hard to establish the stability of the resulting series. There are functional coefficient draws for which the series will be stationary and others for which it will be explosive. From a modeling perspective, it is not appealing having to describe the general behavior of a given series in such a probabilistic manner. Conceptually, we could restrict the sample space of the coefficient functions so as to ensure stationarity, but we do not pursue this for two reasons. First, there is no

strict characterization of the functions that lead to either behavior, there are at best only sufficient conditions on these functions which are pretty restrictive. Chen [20] gives such conditions for stationarity and we go over them in detail in Chapter 4, where we look at theoretical aspects of our model. Second, even if we thus restricted the space of possible functions to ensure stability, it would be very difficult to carry out the calculations required for estimation because the conjugacy would no longer hold. Therefore, we use model (2.2-2.3) mainly as a vehicle for statistical estimation and prediction, for all other considerations we assume the data come from a true, fixed FAR model.

The primary function of our interpretation of model (2.2-2.3) is to analyze the paths of the process conditionally on the past. The advantage in this case is that the process can be evolved sequentially in time. To give an example, suppose the first $q = p \vee d$ data points $X_1, \ldots, X_q$ are given and we want to generate T subsequent observations $X_{q+1}, \ldots, X_{q+T}$ from our model. Using the previous definition, we would first generate the functions $\{f_i\}_{i=1}^p$ from (2.3) and then create the series $X_{q+1}, \ldots, X_{q+T}$ iteratively, from their conditional dynamics in (2.2). In practice, we can not really draw an entire function, but fortunately the generated data depend only on a finite number of evaluations from these random functions. Moreover, this finite vector of function evaluations follows a multivariate normal distribution which can be factorized in a sequence of conditional normal distributions. Specifically, letting $f_{i,t} = f_i(u_t^{(i)})$ for fixed arguments $u_t^{(i)}$, we have

$$\pi(f_{i,q+1}, \ldots, f_{i,q+T}) = \pi(f_{i,q+1})\pi(f_{i,q+2}|f_{i,q+1}) \ldots \pi(f_{i,q+T}|f_{i,q+T-1}, \ldots, f_{i,q+1}) \qquad (2.4)$$

where all the distributions are normal with moments derived from the mean and covariance functions μ_i and C_i. Therefore, we can readily simulate a path with the following scheme: at every time $t = q+1, \ldots, q+T$, we know the values of $X_1, \ldots, X_{t-1}$, $\{U_{q+1}^{(i)}, \ldots, U_t^{(i)}\}_{i=1}^p$ and $\{f_{i,q+1}, \ldots, f_{i,t-1}\}_{i=1}^p$. First, we generate $f_{i,t}$ given $f_{i,q+1}, \ldots, f_{i,t-1}$

from a conditional normal for each i and an independent error ϵ_t, and then we calculate $X_t = f_{1,t}X_{t-1} + \ldots + f_{p,t}X_{t-p} + \epsilon_t$. The random vector $X_{q+T}, \ldots, X_{q+1}$ thus created has the same conditional distribution, given $X_1, \ldots, X_p$, as one coming from model (2.2-2.3). This sequential approach is possible because we condition on the initial observations, whose distribution is generally intractable. In the same way, we can calculate the conditional likelihood of our model, construct predictions and update estimations in an on-line fashion as new data arrive.

We also consider extensions of our model by allowing general regressors and functional coefficient arguments, $X^{(i)}$ and $U^{(i)}$ respectively, which can now be exogenous to the response Y. The model becomes

$$Y_t = f_i(U_t^{(1)})X_t^{(1)} + \ldots + f_i(U_t^{(p)})X_t^{(p)} + \epsilon_t \tag{2.5}$$

$$f_i \sim \mathcal{GP}(\mu_i, C_i); \;\; i = 1, \ldots, p \tag{2.6}$$

which is viewed as a time series regression with varying coefficients. The only requirement on the variables $\{X_t^{(i)}, U_t^{(i)}\}$ is that they be known by time $t-1$, i.e. they are $\mathcal{F}_{t-1}$-measurable for some observable filtration $\mathcal{F}$, which need not be generated only by Y_t. The time subscript is somewhat counter-intuitive, but it helps in keeping the notation simple and representing the model as a regression. If there is at least one exogenous variable in (2.5), the model is not sufficient as a data generating mechanism or for making predictions unless we know the dynamics of the exogenous variables. Therefore, model (2.5) is used for describing the conditional dependence structure of Y_t on the rest of the variables. Other variations include the NLAAR specification by setting all regressors $X^{(i)}$ equal to one. If only $X^{(1)} = 1$, then our model can accommodate a varying mean level. We can also allow a linear AR specification by forcing the coefficients to be random constant functions, which removes the dependence on the argument variables $U^{(i)}$. This can be achieved by imposing a constant

covariance function $C(\cdot,\cdot) = \nu$. Moreover, we can create hybrid specifications by combining varying or constant coefficient functions in multiplicative or additive terms for the dynamics, both for endogenous or exogenous variables. This flexibility will prove useful later on, when we look at model selection procedures. Finally, we also mention an important characteristic that our model inherits from its FAR dynamics. The model is scale invariant but not, in general, location invariant, unless it includes a varying mean function with argument variable the Cartesian product of all other coefficients' arguments. This should be kept in mind when transforming the data prior to the analysis, since, for example, demeaning can lead to inconsistent results.

2.2 Estimation

We now turn attention to estimating the coefficients using our model. We assume throughout that we observe data $\left\{y_t, \{x_t^{(i)}, u_t^{(i)}\}_{i=1}^{p}\right\}_{t=1}^{T}$ from the general model (2.5-2.6), which also covers the Markov FAR model (2.2-2.3) as a special case. If there are endogenous variables in the model dynamics, we treat the first q observations as fixed, where q is the maximum lag of Y_t used in defining $X_t^{(i)}$ and $U_t^{(i)}$. For simplicity we assume the sample runs from 1 to T. We work with the conditional likelihood of the data, where the conditioning is with respect to the regressor and functional coefficient argument variables. Using the conditional likelihood is a common approach in nonlinear time series analysis, since the exact likelihood is almost always intractable. This approach is justified by the ergodic theorem for densities which states that, under certain stability conditions on the series, the conditional likelihood converges to the exact likelihood as the amount of data increases, see Barron [5]. We also assume that the mean and covariance functions of our prior specification in (2.6) are fixed, we discuss how to choose these in later sections. Finally, we assume the error terms are independent and identically distributed normal random variables, $\epsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, in order to take advantage of the conjugacy properties of normals.

First, we introduce some notation: let $f_{i,t} = f_i(u_t^{(i)})$ and define the following

$$\begin{aligned}
\boldsymbol{f}_i^\top &= [f_{i,1}, f_{i,2}, \ldots, f_{i,T}] \\
\boldsymbol{f}^\top &= \left[\boldsymbol{f}_1^\top, \boldsymbol{f}_2^\top, \ldots, \boldsymbol{f}_p^\top\right] \\
\boldsymbol{\mu}_i^\top &= \left[\boldsymbol{\mu}_i(u_1^{(i)}), \boldsymbol{\mu}_i(u_2^{(i)}), \ldots, \boldsymbol{\mu}_i(u_T^{(i)})\right] \\
\boldsymbol{\mu}^\top &= \left[\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top, \ldots, \boldsymbol{\mu}_p^\top\right] \\
\boldsymbol{C}_i &= \left[\left\{C_i\left(u_s^{(i)}, u_t^{(i)}\right)\right\}_{s,t=1}^{T}\right] \\
\boldsymbol{C} &= \begin{bmatrix}
\boldsymbol{C}_1 & \mathbf{0}_{T\times T} & \ldots & \mathbf{0}_{T\times T} \\
\mathbf{0}_{T\times T} & \boldsymbol{C}_2 & \ldots & \mathbf{0}_{T\times T} \\
\vdots & \vdots & \ddots & \vdots \\
\mathbf{0}_{T\times T} & \mathbf{0}_{T\times T} & \ldots & \boldsymbol{C}_p
\end{bmatrix}
\end{aligned}$$

The vector $\boldsymbol{f}$ is the random vector of the functional coefficient evaluations of the model, ordered first by function and then by time. The vector $\boldsymbol{\mu}$ is the prior mean and the matrix $\boldsymbol{C}$ is the prior covariance matrix of $\boldsymbol{f}$, given by the functions $\{\mu_i\}$ and $\{C_i\}$. $\boldsymbol{C}$ has a block diagonal structure as a result of the prior independence of the functions f_i. The prior distribution of the functional coefficient evaluations becomes

$$\pi(\boldsymbol{f}|\boldsymbol{u}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{\mu})^\top \boldsymbol{C}^{-1}(\boldsymbol{f}-\boldsymbol{\mu})\right\} \tag{2.7}$$

where $\boldsymbol{u}$ stands for all the functional coefficient arguments' values. Even if the $\{U^{(i)}\}$ variables depend on the response variable Y, we can still use the sequential conditioning approach we mentioned previously in order to express the prior of $\boldsymbol{f}$ as a multivariate normal. For the likelihood, we also define

$$\boldsymbol{y}^\top = [y_1, y_2, \ldots, y_T]$$

$$\begin{aligned}
\boldsymbol{x}_i^\top &= \left[x_1^{(i)}, x_2^{(i)}, \ldots, x_T^{(i)}\right] \\
\boldsymbol{X}_i &= \text{diag}(\boldsymbol{x}_i) \\
\boldsymbol{X}^\top &= [\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p] \\
\boldsymbol{\Sigma} &= \sigma^2 \boldsymbol{I}_T
\end{aligned}$$

The vector $\boldsymbol{y}$ contains the responses and $\boldsymbol{X}$ can be thought of as an expanded design matrix. The conditional likelihood of the data is

$$\begin{aligned}
\mathcal{L}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{f}) &\propto \prod_{t=1}^{T} \pi(y_t | \{x_t^{(i)}, f_{i,t}\}_{i=1}^{p}) \\
&\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_t - f_{1,t}x_t^{(1)} \ldots - f_{p,t}x_t^{(p)}\right)^2\right\} \qquad (2.8)
\end{aligned}$$

where $\boldsymbol{x}$ stands for all the regressor variables' values. We rearrange the likelihood, expressing it with in terms of $\boldsymbol{f}$, in order to make it compatible with the prior

$$\mathcal{L}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{f}) \propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{f}^\top(\boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top)\boldsymbol{f} - 2\boldsymbol{f}^\top(\boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{y})\right]\right\} \qquad (2.9)$$

Combining the prior and the likelihood, the resulting posterior of the observed function evaluations $\boldsymbol{f}$ is multivariate normal with moments

$$\begin{aligned}
\text{E}[\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}] &= (\boldsymbol{C}^{-1} + \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top)^{-1}(\boldsymbol{C}^{-1}\boldsymbol{\mu} + \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}) \\
&= \boldsymbol{\mu} + \boldsymbol{C}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X} + \boldsymbol{\Sigma})^{-1}(\boldsymbol{y} - \boldsymbol{X}^\top\boldsymbol{\mu}) \qquad (2.10) \\
\text{Var}[\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}] &= (\boldsymbol{C}^{-1} + \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top)^{-1} \\
&= \boldsymbol{C} - \boldsymbol{C}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X} + \boldsymbol{\Sigma})^{-1}\boldsymbol{X}^\top\boldsymbol{C} \qquad (2.11)
\end{aligned}$$

The first expression for the mean comes in the usual Bayesian fashion of a weighted average between the prior and the data. The second expression is the one that we actually use in

practice because it involves the inverse of the smaller $T \times T$ matrix $\boldsymbol{X}^\top \boldsymbol{C} \boldsymbol{X} + \boldsymbol{\Sigma}$, instead of the $(pT) \times (pT)$ matrix $\boldsymbol{C}^{-1} + \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top$. Moreover, the smaller matrix inversion is numerically stable due to the added diagonal matrix $\boldsymbol{\Sigma}$. The posterior mean in (2.10) serves as our Bayes estimate of $\boldsymbol{f}$.

More generally, we show that each functional coefficient follows a GP a posteriori. To verify this, we study the posterior distribution of an arbitrary finite vector of function evaluations $\boldsymbol{f}_N$. This vector can contain evaluations of any function f_i at any set of points $\{u_j^{(i)}\}_{j=1}^{n_i}$

$$\boldsymbol{f}_N^\top = \left[f_1(u_1^{(1)}), \ldots, f_1(u_{n_1}^{(1)}), \ldots, f_p(u_1^{(p)}), \ldots, f_p(u_{n_p}^{(p)}) \right] \tag{2.12}$$

The vector $\boldsymbol{f}_N$ depends on the data only through the correlation of its elements with the observed function evaluations. That is, each new evaluation $f_i(u_j^{(i)})$ of the i^{th} function is only correlated with the evaluations contained in the vector $\boldsymbol{f}_i$ that appear in the likelihood. The prior mean of $\boldsymbol{f}_N$ is

$$\boldsymbol{\mu}_N^\top = \left[\mu_1(u_1^{(1)}), \ldots, \mu_1(u_{n_1}^{(1)}), \ldots, \mu_p(u_1^{(p)}), \ldots, \mu_p(u_{n_p}^{(p)}) \right] \tag{2.13}$$

and the prior covariance of $[\boldsymbol{f}^\top, \boldsymbol{f}_N^\top]^\top$ is

$$\left[\begin{array}{c|c} \boldsymbol{C} & \boldsymbol{C}_N \\ \hline \boldsymbol{C}_N^\top & \boldsymbol{C}_{NN} \end{array}\right] \tag{2.14}$$

which we view as a partitioned matrix, with partitioning according to the vectors $\boldsymbol{f}$, $\boldsymbol{f}_N$. Note that each element of $\boldsymbol{f}_N$ will only be correlated with elements from the same function, so both $\boldsymbol{C}_N$ and $\boldsymbol{C}_{NN}$ will be sparse. In particular, if $\boldsymbol{f}_N$ contains only one evaluation from each function then $\boldsymbol{C}_{NN}$ will be diagonal and $\boldsymbol{C}_N^\top$ will have nonzero elements only where

its i^{th} row is underneath $\boldsymbol{C}$'s i^{th} block. The joint prior of $[\boldsymbol{f}^\top, \boldsymbol{f}_N^\top]^\top$ is

$$\left[\begin{array}{c} \boldsymbol{f} \\ \boldsymbol{f}_N \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} \boldsymbol{\mu} \\ \boldsymbol{\mu}_N \end{array}\right], \left[\begin{array}{c|c} \boldsymbol{C} & \boldsymbol{C}_N \\ \hline \boldsymbol{C}_N^\top & \boldsymbol{C}_{NN} \end{array}\right]\right) \tag{2.15}$$

After multiplying with the likelihood and integrating out $\boldsymbol{f}$ we get that the posterior of $\boldsymbol{f}_N$ is normal, with moments

$$\mathrm{E}[\boldsymbol{f}_N|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}] = \boldsymbol{\mu}_N + \boldsymbol{C}_N^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X} + \boldsymbol{\Sigma})^{-1}(\boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{\mu}) \tag{2.16}$$

$$\mathrm{Var}[\boldsymbol{f}_N|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}] = \boldsymbol{C}_{NN} - \boldsymbol{C}_N^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X} + \boldsymbol{\Sigma})^{-1}\boldsymbol{X}^\top \boldsymbol{C}_N \tag{2.17}$$

By letting $\boldsymbol{f}_N$ contain evaluations from a single function only, it is obvious that the posterior distribution of each function is again a GP, with mean and variance functions depending on the data. The above formulas are the result of straightforward normal calculations, analogous to those for the Bayesian linear model, see e.g. Lindley and Smith [74].

Besides the functional coefficients themselves, it is practically and conceptually advantageous to look at the conditional mean of the observations. We define the conditional mean Z_t as the quantity

$$Z_t = \mathrm{E}[Y_t|\{X_t^{(i)}, U_t^{(i)}\}_{i=1}^p] = \sum_{i=1}^p X_t^{(i)} f_i(U_t^{(i)}) \tag{2.18}$$

where the expectation is taken with respect to the error term's distribution. The conditioning refers to the variables $X^{(i)}$ and $U^{(i)}$, and not the functional coefficients. This means that we treat $X^{(i)}$, $U^{(i)}$ as known, similarly to exogenous variables in a regression, and that all the uncertainty about Z_t comes from the random functional coefficients. Since the functional coefficients follow GPs and Z_t is a linear combination of them, it will follow a normal distribution. More specifically, letting $\boldsymbol{z}^\top = [Z_1, Z_2, \ldots, Z_T] = \boldsymbol{X}^\top \boldsymbol{f}$ we look at the

prior and posterior distribution of $\boldsymbol{z}$, which are both multivariate normal. The prior mean of z is

$$\mathrm{E}[\boldsymbol{z}|\boldsymbol{x},\boldsymbol{u}] = \boldsymbol{X}^\top \mathrm{E}[\boldsymbol{f}|\boldsymbol{u}] = \boldsymbol{X}^\top \boldsymbol{\mu} \tag{2.19}$$

where the expectation is taken with respect to the prior measure of the functional coefficients. The prior covariance is

$$\begin{aligned} \mathrm{Var}[\boldsymbol{z}|\boldsymbol{x},\boldsymbol{u}] &= \boldsymbol{X}^\top \mathrm{Var}[\boldsymbol{f}|\boldsymbol{u}]\boldsymbol{X} = \boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X} \\ &= \sum_{i=1}^{p} \boldsymbol{X}_i^\top \boldsymbol{C}_i \boldsymbol{X}_i = \sum_{i=1}^{p} (\boldsymbol{x}_i \boldsymbol{x}_i^\top) \circ \boldsymbol{C}_i \end{aligned} \tag{2.20}$$

where $\circ$ is the Hadamard product. The posterior mean of $\boldsymbol{z}$ is given by

$$\begin{aligned} \mathrm{E}[\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{u}] &= \boldsymbol{X}^\top \mathrm{E}[\boldsymbol{f}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{u}] = \boldsymbol{X}^\top (\boldsymbol{C}^{-1} + \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top)^{-1}(\boldsymbol{C}^{-1}\boldsymbol{\mu} + \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}) \\ &= \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\mu} + \boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X}(\boldsymbol{\Sigma} + \boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X})^{-1}\boldsymbol{y} \\ &= \boldsymbol{X}^\top \boldsymbol{\mu} + \boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X}(\boldsymbol{\Sigma} + \boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X})^{-1}(\boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{\mu}) \end{aligned} \tag{2.21}$$

Like before, the expectation is taken with respect to the posterior measure of the functional coefficients. Actually, the posterior mean of $\boldsymbol{z}$ is what we would consider as the model's fitted values. The second expression in (2.21) represents the posterior mean of $\boldsymbol{z}$ as a weighted average of its prior mean and the data. The third expression has an interpretation in terms of a frequentist nonparametric smoothing problem. The fitted values are given as a bias term $\boldsymbol{X}^\top \boldsymbol{\mu}$ plus the smoothed deviations $(\boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{\mu})$ of the data from the bias, where the smoothing or hat matrix is given by $\boldsymbol{H} = \boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X}(\boldsymbol{\Sigma} + \boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X})^{-1}$. Moreover, the posterior covariance of $\boldsymbol{z}$ is

$$\mathrm{Var}[\boldsymbol{z}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{u}] = \boldsymbol{X}^\top \mathrm{Var}[\boldsymbol{f}|\boldsymbol{y},\boldsymbol{x},\boldsymbol{u}]\boldsymbol{X}$$

$$
\begin{aligned}
&= \boldsymbol{X}^\top(\boldsymbol{C}^{-1} + \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top)^{-1}\boldsymbol{X} \\
&= \boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X} - \boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}(\boldsymbol{\Sigma} + \boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X} \\
&= \boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}(\boldsymbol{\Sigma} + \boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X})^{-1}\boldsymbol{\Sigma} \qquad (2.22)
\end{aligned}
$$

Finally, we demonstrate how sequential estimation can be efficiently achieved in our model. We can view the problem of sequential estimation as that of updating the inverse matrix $\boldsymbol{A}^{-1} = (\boldsymbol{\Sigma} + \boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X})^{-1}$, since this is where the main computational burden lies. Each new observation $(y_{T+1}, \{x_{T+1}^{(i)}, u_{T+1}^{(i)}\}_{i=1}^{p})$ adds another dimension to this matrix, so we need to invert the partitioned matrix

$$
\boldsymbol{A}' = \left[\begin{array}{c|c} \boldsymbol{A} & \boldsymbol{b} \\ \hline \boldsymbol{b}^\top & c \end{array}\right]
$$

where $c = \sigma^2 + \sum_{i=1}^{p} C(u_{T+1}^{(i)}, u_{T+1}^{(i)})(x_{T+1}^{(i)})^2$ and $\boldsymbol{b} = \sum_{i=1}^{p} \boldsymbol{X}_i\left[\{C(u_t^{(i)}, u_{T+1}^{(i)})\}_{t=1}^{T}\right]x_{T+1}^{(i)}$. Given knowledge of $\boldsymbol{A}^{-1}$, we can find $\boldsymbol{A}'^{-1}$ using the formula for partitioned matrix inverses. The matrix $\boldsymbol{A}'^{-1}$ is given in partitioned form as

$$
\boldsymbol{A}'^{-1} = \left[\begin{array}{c|c} \tilde{\boldsymbol{A}} & \tilde{\boldsymbol{b}} \\ \hline \tilde{\boldsymbol{b}}^\top & \tilde{c} \end{array}\right] \qquad (2.23)
$$

where $\tilde{c} = 1/(c - \boldsymbol{b}^\top\boldsymbol{A}^{-1}\boldsymbol{b})$, $\tilde{\boldsymbol{b}} = -\boldsymbol{A}^{-1}\boldsymbol{b}^\top\tilde{c}$ and $\tilde{\boldsymbol{A}} = \boldsymbol{A}^{-1} + \tilde{c}\boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{b}^\top\boldsymbol{A}^{-1}$. This is a typical normal Bayesian updating calculation (see e.g. Harrison and West [121]) and the required computations, given $\boldsymbol{A}^{-1}$, can be performed in time $\mathcal{O}(T^2)$. As a result, sequential estimation scales at the same $\mathcal{O}(T^3)$ rate as batch estimation in terms of computations, practical considerations aside.

2.3 Prediction

In this section we describe the procedure for making predictions. We consider the Markovian case which provides a self-contained model for the data, since the presence of exogenous variables does not permit predictions unless we know how they evolve. One-step-ahead predictions follow naturally from the posterior distribution of the functional coefficients. Let $\boldsymbol{x}_N^\top = [x_{T+1}^{(1)}, \ldots, x_{T+1}^{(p)}]$ and $\boldsymbol{f}_N^\top = [f_1(u_{T+1}^{(1)}), \ldots, f_p(u_{T+1}^{(p)})]$, where $\{x_{T+1}^{(i)}, u_{T+1}^{(i)}\}_{i=1}^p$ are known by time T. We use formulas (2.16-2.17) to get the posterior mean and covariance matrix of $\boldsymbol{f}_N$. The next value $Y_{T+1} = \sum_{i=1}^p f_i(u_{T+1}^{(i)}) x_{T+1}^{(i)} + \epsilon_{T+1}$ is a linear combination of $\boldsymbol{f}_N$ and the error term ϵ_{T+1}, so its predictive distribution is normal with moments

$$\begin{aligned}
\mathrm{E}[Y_{T+1}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}] &= \boldsymbol{x}_N^\top \mathrm{E}[\boldsymbol{f}_N|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}] \\
&= \boldsymbol{x}_N^\top \boldsymbol{\mu}_N + \boldsymbol{x}_N^\top \boldsymbol{C}_N^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X} + \boldsymbol{\Sigma})^{-1}(\boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{\mu}) \qquad (2.24) \\
\mathrm{Var}[Y_{T+1}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}] &= \boldsymbol{x}_N^\top (\mathrm{Var}[\boldsymbol{f}_N|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}]) \boldsymbol{x}_N + \sigma^2 \\
&= \boldsymbol{x}_N^\top \big(\boldsymbol{C}_{NN} - \boldsymbol{C}_N^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X} + \boldsymbol{\Sigma})^{-1} \boldsymbol{X}^\top \boldsymbol{C}_N\big) \boldsymbol{x}_N + \sigma^2 \qquad (2.25)
\end{aligned}$$

where $\boldsymbol{\mu}_N, \boldsymbol{C}_N$ correspond to the current definition of $\boldsymbol{f}_N$. Note that we account for both the estimation uncertainty in the coefficient functions and the inherent model uncertainty coming from the error term ϵ_{T+1}.

For multi-step-ahead prediction we cannot, in general, find the distribution of the process beyond time $T+1$ explicitly. The predicted values serve either as regressors or as arguments to the nonlinear functional coefficients and we thus lose the conditional normality structure. We can, however, use Monte Carlo simulation from the model posterior to approximate the predictive distributions, or any other quantity depending on them. For example, if we are interested in S-step-ahead predictions, we can use a sample of paths $\{y_{T+1}^{(b)}, \ldots, y_{T+S}^{(b)}\}_{b=1}^B$ coming from the model. One detail that requires attention is that each path $y_{T+1}^{(b)}, \ldots, y_{T+S}^{(b)}$

must come from an independent draw of coefficient functions from the posterior. To achieve this, we first draw $y_{T+1}^{(b)}$ from the normal distribution described above and then we calculate the regressor and argument variables for time $T+2$. The vector of functional coefficient evaluations required for $y_{T+2}^{(b)}$, must now be drawn from the posterior, also given the values we generated at time $T+1$. Repeating this procedure recursively within each path b, we have to update the posterior of the functional coefficients at each time, treating our draws as observations. This Bayesian updating can be performed more efficiently using the sequential estimation scheme we presented before.

Nevertheless, this approach can be slow if the number of observations T or the number of steps S is high, in which case we can use simpler alternatives. Given enough observations, we can disregard estimation uncertainty by treating the functional coefficients as known and equal to their posterior mean, the variability in the resulting generated paths coming solely from the error term. Moreover, if we are interested only in point estimates of the future values, we can just evolve the process iteratively by generating a single path where every function evaluation is equal to its posterior mean and the errors are zero. A more rigorous approach for finding the predictive distributions of the process can be adapted form the work of Girard et al. [45], who look at a GP regression with random inputs and propose an approximation scheme for the predictive distribution of the regression function at a normally distributed argument. Their approximation relies on a second order Taylor expansion of the posterior mean and the covariance functions of the regression surface and it preserves normality. Extending their approach in our setting, though, is more involved because we would have to deal with products of normals, since the regressors are multiplied with the coefficients. This could degrade the quality of the approximation significantly and we did not pursue the approach for this reason.

2.4 Prior Specification

In our discussion of estimation and prediction we assumed the GP prior is fully specified. In this section we propose a systematic way of setting up this prior. The choices we have to make concern the prior mean and covariance functions, $\mu_i(\cdot)$ and $C_i(\cdot,\cdot)$ respectively, for each f_i. For both of these we choose simple forms to describe them. It is obvious from formula (2.16) that the mean function defines the prior bias of each coefficient function. Since our procedure is nonparametric and it allows reasonable flexibility for the posterior estimates, we use a constant mean function $\mu_i(\cdot) = \mu_i$; the prior bias we thus introduce favors linearity. For the vast majority of GP regression applications the prior mean is set to zero, but we will allow it to assume arbitrary values. We do this because later on we look at models with constant coefficient functions, i.e. random but not varying autoregressive coefficients, and we do not necessarily want to shrink them towards zero.

The choice of the prior covariance function is more important because it affects the shape and properties of the coefficient functions, as well as smoothing. In effect, the covariance function quantifies how close should the function evaluations be depending on their arguments and the only requirement on it is that it be a positive definite kernel. There are quite a few covariance kernels suggested in the literature, but by far the most popular is the squared exponential

$$C(\boldsymbol{x}, \boldsymbol{x}') = \nu^2 \exp\left\{-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{h^2}\right\} \tag{2.26}$$

This is the one we will use throughout, mainly because it has few parameters with intuitive roles and because it is easy to compute. This covariance function is described by ν which controls the prior strength and by h, called the characteristic lengthscale, which controls the amount of smoothing and is the direct analogue of the bandwidth parameter in kernel

regression. Another important feature of this covariance function is that it is stationary, meaning that the smoothing depends only on the distance $d = \|\boldsymbol{x} - \boldsymbol{x}'\|$ between the inputs and not on their position in the input space. Other important alternatives are the exponential or Ornstein-Uhlenbeck function ($C_{OU}(d) = \nu^2 \exp\{-d/h\}$, where $h > 0$), the Matérn function ($C_{\text{Matérn}}(d) = \nu^2(2^{1-\kappa}/\Gamma(\kappa))(\sqrt{2\kappa}d/h)^\kappa J_\kappa(\sqrt{2\kappa}d/h)$, where $\kappa, h > 0$ and $J_\kappa(\cdot)$ is a modified Bessel function) which is a generalization of both the exponential and squared exponential covariance functions, and different methods for constructing non-stationary covariance functions. For a more detailed discussion on these alternatives and others see Stein [109], Rasmussen and Williams [96] and Paciorek and Schervish [91].

2.5 Hyperparameter Selection

Using a constant mean function μ_i and a squared exponential covariance function C_i for each functional coefficient f_i, we still have a number of hyperparameters we have to specify in order to implement our estimation procedure. We collect these, together with the variance σ^2 of the normal error terms, in a single vector $\boldsymbol{\theta}^\top = [\sigma, \{\mu_i, \nu_i, h_i\}_{i=1}^p]$. It seems impossible to assume exact prior knowledge of $\boldsymbol{\theta}$, so we need a method for selecting the hyperparameters. For this we rely on the (conditional) marginal likelihood of the data $\mathcal{L}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{u}; \boldsymbol{\theta}) = \int_{\boldsymbol{f}} \mathcal{L}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{f}; \boldsymbol{\theta})\pi(\boldsymbol{f}|\boldsymbol{u}; \boldsymbol{\theta})d\boldsymbol{f}$, where we marginalize with respect to the vector of function evaluations $\boldsymbol{f}$. Substituting the prior and the likelihood from (2.7) and (2.9) and carrying out the integration, we get

$$\begin{aligned} \mathcal{L}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{u}; \boldsymbol{\theta}) \quad \propto \quad & |\boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X} + \boldsymbol{\Sigma}|^{-1/2} \times \\ & \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}^\top\boldsymbol{\mu})^\top(\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X} + \boldsymbol{\Sigma})^{-1}(\boldsymbol{y} - \boldsymbol{X}^\top\boldsymbol{\mu})\right\} \end{aligned} \tag{2.27}$$

We can use this quantity in different ways; one possibility is fully Bayesian and relies on hierarchical modeling, treating the hyperparameters as random and putting a prior distri-

bution on them. However, there is no established framework for defining the hyperpriors, and the posterior distribution of $\boldsymbol{\theta}$ is analytically intractable and must be approximated by MCMC methods. Moreover, this procedure is practically unattractive because of the computational burden of MCMC methods, and because of the additional layer of uncertainty in the model which we have to take into account when making probabilistic statements or predictions.

For these reasons, we focus on the empirical Bayes approach which chooses the hyperparameters by maximizing the marginal log-likelihood of the data. The main advantage of this approach is that it is very convenient for selecting multiple parameters because the gradient of the marginal log-likelihood is also available. Let $\boldsymbol{S} = \boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X} + \boldsymbol{\Sigma}$ and $\boldsymbol{w} = \boldsymbol{S}^{-1}(\boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{\mu})$. Suppose θ_S is some parameter in $\boldsymbol{S}$ (one of σ, ν_i or h_i in our case) and let $\frac{\partial \boldsymbol{S}}{\partial \theta_S}$ be the matrix whose elements are the partial derivatives of the elements of $\boldsymbol{S}$ with respect to θ_S. The partial derivative of the marginal log-likelihood with respect to θ_S is given by:

$$\frac{\partial \ell}{\partial \theta_S} = \frac{1}{2} \operatorname{tr}\left((\boldsymbol{w}\boldsymbol{w}^\top - \boldsymbol{S}^{-1}) \frac{\partial \boldsymbol{S}}{\partial \theta_S} \right) \tag{2.28}$$

Also, suppose θ_μ is some parameter in $\boldsymbol{\mu}$ (a μ_i in our case) and let $\frac{\partial \boldsymbol{\mu}}{\partial \theta_\mu}$ be a vector whose elements are the partial derivative of the elements of $\boldsymbol{\mu}$ w.r.t. θ_μ. The partial derivative of the marginal log-likelihood w.r.t. θ_μ is given by:

$$\frac{\partial \ell}{\partial \theta_\mu} = (\boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{\mu})^\top \boldsymbol{S}^{-1} \boldsymbol{X}^\top \frac{\partial \boldsymbol{\mu}}{\partial \theta_\mu} \tag{2.29}$$

We use the gradient for maximizing the marginal log-likelihood of the data. Getting the Hessian is more involved and we do not pursue this, but even with the gradient there is a variety of available schemes to perform this unconstrained optimization task (for positive

hyperparameters we use a logarithmic transformation to avoid constraints) from simple gradient descent to quasi-Newton methods. Normally, the objective function will have multiple local maxima and we deal with this issue by a judicious choice of starting values which we describe at the end of this section.

There are two problems in selecting the hyperparameters of the model, in particular the prior uncertainties ν_i, in an empirical fashion. First, we describe these problems and then we present our approach on how to overcome them, together with the intuition behind it. The first problem has to do with the distribution of the prior uncertainty among the functional coefficients. Notice that the prior covariance of the conditional means $\boldsymbol{z}$ in (2.20) is the sum of the covariances of the individual coefficient functions $\boldsymbol{C}_i$ weighted by the outer product $\boldsymbol{x}_i\boldsymbol{x}_i^\top$ of the variables by which the coefficients are multiplied. This can result in a kind of variance non-identifiability, meaning that, in some cases, we can decrease one ν_i and still get almost the same covariance for $\boldsymbol{z}$ by increasing some other ν_j or σ. This behavior is not absolute since it depends on the form of $\boldsymbol{x}_i\boldsymbol{x}_i^\top$, but it is nevertheless common. What typically happens when we use empirical Bayes to select each ν_i independently is that the functions which are very smooth tend to have almost zero prior uncertainty and they end up being treated as known a priori; we give an example of this behavior in the next section. In order to overcome this problem we suggest distributing the prior uncertainty evenly among functions. We present this approach by assuming we have a desired prior uncertainty level $p\tau^2$ for the conditional mean of the observations, i.e. we want the diagonal elements of $\mathrm{Var}[\boldsymbol{z}|\boldsymbol{x},\boldsymbol{u}]$ to be around $p\tau^2$. The prior uncertainty of $Z_t|\boldsymbol{x},\boldsymbol{u} = \sum_{i=1}^p x_t^{(i)} f_i(x_t^{(i)})$ depends also on the regressor values, so we want to balance the contribution of each term $x_t^{(i)} f_i(u_t^{(i)})$ in such a way that $(x_t^{(i)})^2 \,\mathrm{Var}[f_i(u_t^{(i)})] = (x_t^{(i)})^2\nu_i^2 \approx \tau^2$. To this end we set $\nu_i^2 = \tau^2/v_i^2$ where $v_i^2 = \sum_{t=1}^T (x_t^{(i)})^2/T$, which serves our purpose in the sense that the prior variance contribution of each term is on average equal to τ^2. Essentially, we are rescaling the re-

gressor variables by their empirical second order moment, so that in the rescaled model $X'^{(i)} = X^{(i)}/v_i$ and the prior uncertainty is equally distributed as $\nu_i'^2 = \tau^2$. The only difference is that the rescaling is done through the parameters ν_i of the functions' priors. Moreover, the original and rescaled models are equivalent because of the scale invariance of the FAR model. Besides making the variance contributions of each term in the model comparable, this practice is extremely helpful for the numerical stability of the relevant computations.

Our treatment of prior uncertainty is related to that of empirical Bayes shrinkage estimation; see Efron and Morris [33] for an overview. To demonstrate this argument we consider empirical Bayes estimation in the multiple linear regression model, which is equivalent to assuming a priori that the coefficient functions in model (2.5) are constant, i.e. $f_i(\cdot) = \beta_i$. Let $\boldsymbol{\beta}^\top = [\beta_1, \ldots, \beta_p]$ be the vector of regression coefficients and $\boldsymbol{X}_d = [\boldsymbol{x}_i \cdots \boldsymbol{x}_p]$ be the design matrix. Oman [89] looks at priors of the form $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \tau^2 \boldsymbol{V})$, for which the Bayes estimator (posterior mean) is $(\boldsymbol{X}_d^\top \boldsymbol{X}_d + \sigma^2/\tau^2 \boldsymbol{V}^{-1})(\boldsymbol{X}_d^\top \boldsymbol{X}_d)\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the least squares estimate. He specifically considers two cases, one where $\boldsymbol{V} = (\boldsymbol{X}_d^\top \boldsymbol{X}_d)^{-1}$ gives rise to Stein-type shrinkage and a second where $\boldsymbol{V} = \boldsymbol{I}$ results in ridge-type shrinkage; the term σ^2/τ^2 controlling the shrinking strength in both. Our prior covariance specification uses $\boldsymbol{V} = \mathrm{diag}([v_1, \ldots, v_p]) = (\mathrm{diag}(\boldsymbol{X}_d^\top \boldsymbol{X}_d/T))^{-1}$ which, being diagonal, raises similarities to ridge-type estimation. However, we are not actually shrinking the estimates because the prior mean is data dependent (it maximizes the marginal likelihood) and so it will be close to $\hat{\boldsymbol{\beta}}$.

The second problem concerns the prior uncertainty of the conditional mean of the observations $p\tau^2$ itself. In general, selecting the hyperparameter τ by maximizing the marginal likelihood can also lead to problematic behavior. The problem arises when all coefficient

functions are close to constant, in which case τ will tend to zero, and is again due to non-identifiability between τ^2 and σ^2. We address this by defining ν^2 in terms of σ^2 in a sensible way, we propose setting $\tau^2 = \sigma^2$. We look at the consequences of this choice by drawing an analogy to the previously discussed regression setting. Our prior on the coefficient functions becomes $\boldsymbol{\beta} \sim \mathcal{N}\big(\boldsymbol{\mu}_\beta, \sigma^2 \operatorname{diag}(\boldsymbol{X}_d^\top \boldsymbol{X}_d/T)^{-1}\big)$, where $\boldsymbol{\mu}_\beta$ denotes the prior mean of $\boldsymbol{\beta}$. The posterior variance of $\boldsymbol{\beta}$ is $\sigma^2 \left(\boldsymbol{X}_d^\top \boldsymbol{X}_d + \frac{1}{T} \operatorname{diag}(\boldsymbol{X}_d^\top \boldsymbol{X}_d)\right)^{-1}$, which is close to the Fisher information matrix $\sigma^2(\boldsymbol{X}_d^\top \boldsymbol{X}_d)^{-1}$, it falls short of it by approximately a factor of $T/(T+1)$. The prior covariance choice gives, in relative terms, as much information for the parameter as that obtained in a single observation. Thus, we manage to relate the prior uncertainty of the coefficients to that of the observations and make the posterior behave in a frequentist (square root of T) manner. When the coefficient functions are varying the situation is more complicated since we perform local estimation. Besides the prior uncertainty ν_i, the smoothing parameter h_i is also important because it controls information sharing between observations. As a result, the posterior variance of a varying function evaluated at a given point will be higher than that of a constant coefficient.

We give a simple demonstration of our treatment of prior uncertainty by looking at the varying mean model $Y_t = f(U_t) + \epsilon_t$. We consider the two extreme cases $h = 0$ and $h = \infty$ for the smoothing parameter of f, the former corresponding to estimating a separate mean for each observation and the latter corresponding to estimating a common constant mean. For $h = \infty$ we have $\boldsymbol{C} = \sigma^2 \mathbf{1}_{T\times T}$, the vector of functional coefficients being perfectly correlated since we effectively estimate only one parameter. The posterior variance of $\boldsymbol{f}$ is

$$\operatorname{Var}[\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{u}] = \sigma^2\big(\boldsymbol{I}_T - (\boldsymbol{I}_T + \mathbf{1}_{T\times T})^{-1}\big) = \frac{\sigma^2}{T+1}\mathbf{1}_{T\times T}$$

which is approximately the inverse Fisher information matrix we got before. On the other hand, if $h = 0$ we have $\boldsymbol{C} = \sigma^2 \boldsymbol{I}_T$, so that each evaluation of f is completely independent

of all others. The posterior variance of $\boldsymbol{f}$ is

$$\mathrm{Var}[\boldsymbol{f}|\boldsymbol{y},\boldsymbol{u}] = \sigma^2\big(\boldsymbol{I}_T - (\boldsymbol{I}_T + \boldsymbol{I}_T)^{-1}\big) = \frac{\sigma^2}{2}\boldsymbol{I}_T$$

The elements of $\boldsymbol{f}$ are still independent and for each one we essentially use the information in only one observation, thus resulting in a posterior variance of $\sigma^2/2$. For positive but finite values of h, its magnitude will define the extent of information sharing between observations and the posterior variance of any observed function evaluation will be between $\sigma^2/2$ and $\sigma^2/(T+1)$. This behavior holds in general for more complex models; the posterior variance of any functional coefficient evaluation will be a compromise between the two extremes depending on the amount of smoothing we do, the locations of the arguments and the design matrix $\boldsymbol{X}$.

Our discussion implicitly assumed that σ is known, but in fact it also has to be selected from the data. The greatest danger in setting $\tau^2 = \sigma^2$ is that we tie the error variance with the parameter of the prior covariance function, which also plays a role in smoothing. However, we have seen in practice that σ is selected based primarily on the error variance and does not change a lot whether we define ν_i in terms of σ or let it be a free parameter. There are two additional important advantages that result from this practice. For one, we significantly reduce the number of hyperparameters and this simplifies the nonlinear optimization procedure by making it faster and less prone to local maxima. Moreover, we have only one hyperparameter, h_i, to control the amount of smoothing and this helps later, when we make comparisons between models with varying and constant coefficients.

The resulting vector of hyperparameters, after fixing the prior uncertainties $\{\nu_i\}$ in relation to σ, becomes $\boldsymbol{\theta}^\top = [\sigma, \{\mu_i, h_i\}_{i=1}^p]$ and we select it by a gradient descent scheme. We decide the starting values of the algorithm in such a way so that it converges fast to some

reasonable local maximum. From intuition and some experimentation we have found that a good choice of starting values is as follows. First, we fit a linear regression to model (2.5), assuming the coefficients are fixed constants. If there are terms with the same regressor variables but different arguments, we only keep one of each regressor variable in the design matrix in order to have full rank. We set the initial value of σ equal to the regression standard error estimate and the initial values of the prior means μ_i of the functions equal to the estimated regression coefficients. If there is more than one function multiplied with the same regressor, we set the initial value for the mean of each function to be equal to the estimated coefficient of that regressor divided by the number of functions that share it, thus splitting the effect of each functional coefficient equally. For the starting values of the characteristic lengthscale h_i we use the sample standard deviation of the argument variable $U^{(i)}$, which implies moderate smoothness. The starting values selected in this way tend to give relatively stable and reasonable results.

2.6 Example

We apply our model to a real data set, we look at the famous Canadian lynx data which is the annual record of the number of Canadian lynx trapped in the MacKenzie river district from 1821 to 1934. This time series has traditionally served as an example of the need for nonlinear and nonparametric time series models. A first analysis of the data with an AR model was attempted by Moran [85], with later overviews and analyses by Tong [113] and Campbell and Walker [17]. However, the appropriateness of the AR model and its variations were disputed early on, so this data set quickly became a testing ground for new time series models and estimation techniques and has achieved benchmark status. Tong [114] provides a detailed exploratory and qualitative analysis and applies different TAR, RCAR and bilinear models. More recently, Cai, Fan and Yao [16] used LLR for this data set, whereas Lin and Pourahmadi [72] reviewed various nonparametric estimation techniques

from regression applied to it. We work with the base 10 logarithm of the data in order to stabilize the variance; the plot of the resulting series is shown in Fig. 2.1. As suggested in the aforementioned literature, we use a second order FAR model with functional coefficients depending on the lag-two series, i.e. $X_t^{(1)} = X_{t-1}$, $X_t^{(2)} = X_{t-2}$ and $U_t^{(1)} = U_t^{(2)} = X_{t-2}$. The dynamics of the series are described by

$$X_t = f_1(X_{t-2})X_{t-1} + f_2(X_{t-2})X_{t-2} + \epsilon_t \tag{2.30}$$

We apply our estimation procedure, assuming the two functional coefficients f_1, f_2 follow independent GP priors with constant mean functions and squared exponential covariance kernels. We specify the prior as described in section 2.6 and we end up with hyperparameters σ and $\{\mu_i, h_i\}_{i=1,2}$ which are chosen by maximizing the marginal likelihood. We use a simple gradient descent algorithm with initial values derived from the AR model. The selected hyperparameters are given in Table 2.1, and the resulting posterior estimates of the functional coefficients are presented in Fig. 2.2, together with pointwise 95% posterior confidence intervals. As we can see, the fitted model suggests that f_1 is almost constant and f_2 becomes more negative as X_{t-2} increases.

Table 2.1: The hyperparameters that maximize the marginal likelihood of the Canadian lynx series.

		f_1		f_2	
σ	0.2091714	μ_1	1.3747400	μ_2	-0.3486145
		h_1	2.535278	h_2	0.736689
Induced		ν_1	0.07078407	ν_2	0.07099573
$\ell = 9.258848$					

For comparison, we fit the model with the two alternative prior specifications we discussed in the previous section, i.e. we treat (ν_1, ν_2) as a free parameters or treat τ^2 as a free parameter. The results for the first case are shown in Fig. 2.3 and for the second case in

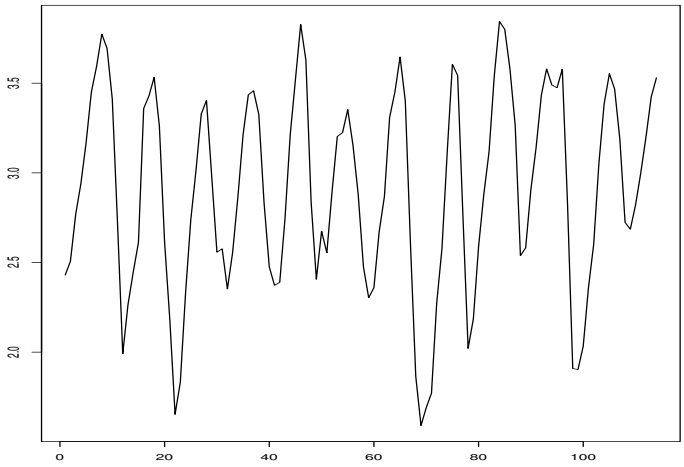
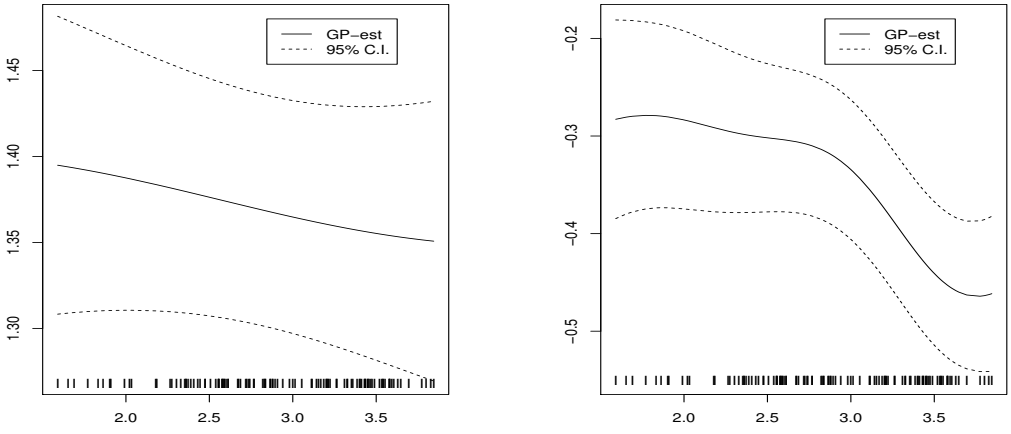


Figure 2.1: Plot of the logarithm of the Canadian lynx time series.

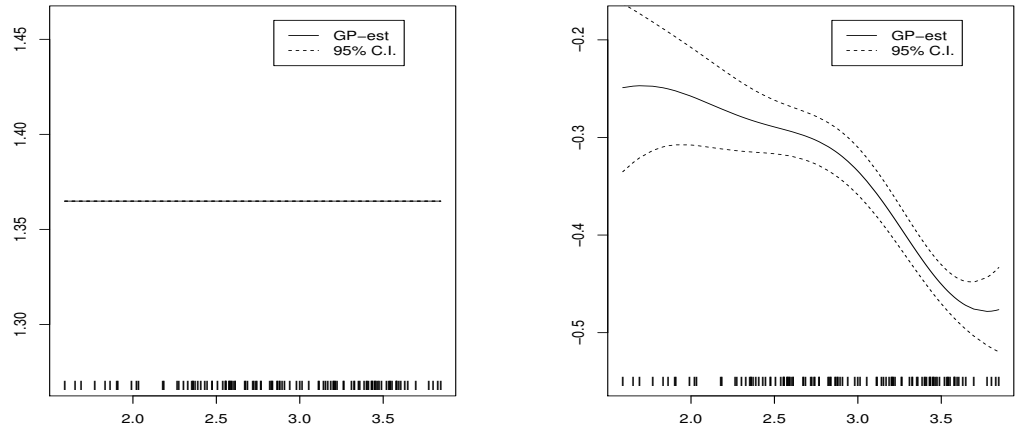


(a) (b)

Figure 2.2: Plots of GP estimated functional coefficients for (a) f_1 and (b) f_2 for the Canadian lynx series (the marks at the bottom indicate the locations of the observed arguments).

Fig. 2.4. We also provide the values of the hyperparameters that maximize the marginal log-likelihood in Table 2.2 and Table 2.3 respectively. We note that both specifications resulted in a constant first functional coefficient. In the case where ν_1, ν_2 are free, we see

that the uncertainty for the first coefficient is almost zero. The characteristic lengthscale h_1 is high but having ν_1 so small makes it irrelevant; the posterior is practically equal to the prior (which is constant) independently of the value of h_1. The posterior for the second functional coefficient has similar shape and slightly smaller variance than the other two methods. In the case where τ is a free parameter we get almost the same results as our suggested method, the most striking difference being the huge smoothing parameter h_1. Our proposed specification's estimate of f_1 is also very smooth, and in Chapter 5 we look at procedures for deciding whether a function should be treated as constant or not. Notice also that the optimal value of τ is close to that of σ so the induced prior uncertainties are similar, and those for f_2 are also comparable to the freely selected ν_2. In general, the posterior means of the functional coefficients from all three specifications are quite close, so there are not dramatic changes in fit. The specifications are close in terms of marginal likelihood as well, even though the ones with more parameters always achieve higher likelihoods.

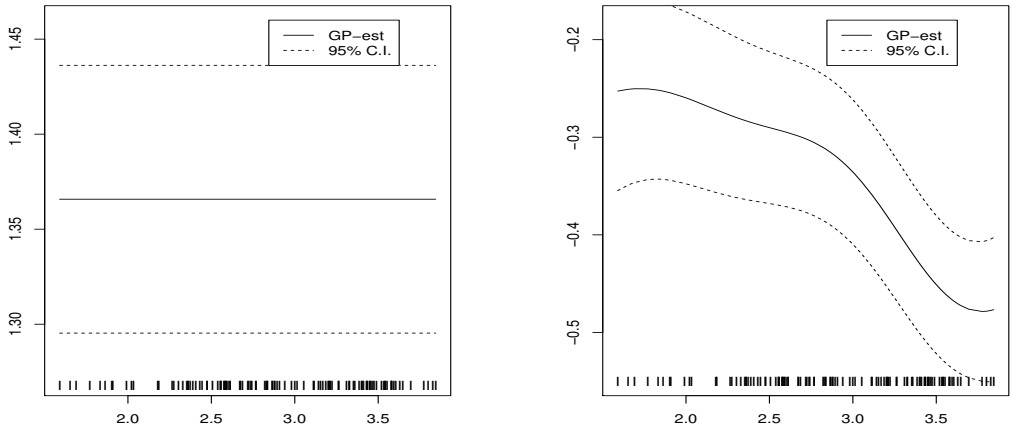


(a) (b)

Figure 2.3: Plots of GP estimated functional coefficients for (a) f_1 and (b) f_2 for the Canadian lynx series, with ν_1, ν_2 treated as free parameters.

Table 2.2: The hyperparameters that maximize the marginal likelihood of the Canadian lynx series, with ν_1, ν_2 treated as free parameters.

			f_1		f_2
σ	0.2080146	μ_1	1.3648734	μ_2	-0.3393747
		h_1	174.8650339	h_2	0.7949713
		ν_1	6.575×10^{-6}	ν_2	0.08858933
$\ell = 10.12103$					



(a) (b)

Figure 2.4: Plots of GP estimated functional coefficients for (a) f_1 and (b) f_2 for the Canadian lynx series, with τ treated as a free parameter.

Table 2.3: The hyperparameters that maximize the marginal likelihood of the Canadian lynx series, with τ treated as a free parameter.

			f_1		f_2
σ	0.2084579	μ_1	1.365773	μ_2	-0.341352
τ	0.2460876	h_1	2.121209×10^{5}	h_2	0.7860304
Induced		ν_1	0.08327662	ν_2	0.08352563
$\ell = 9.268973$					

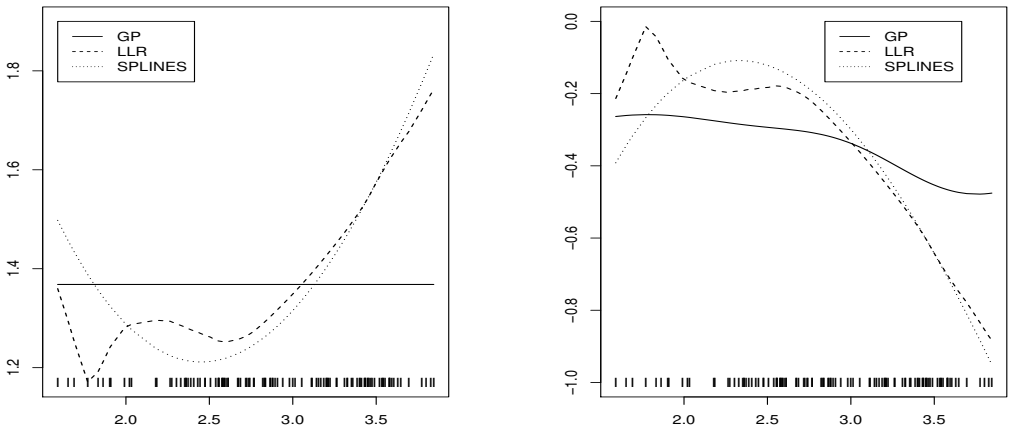
We also apply nonparametric estimation for the FAR model in (2.30) using LLR and splines, as well as a parametric TAR model. For the LLR procedure we follow Cai, Fan and Yao [16] which include an analysis of the lynx data in their paper and for the splines method, we

follow the procedure proposed in Huang and Shen [60], using the same number of knots for each function. The estimated functional coefficients for the two nonparametric methods are presented in Fig. 2.5, superimposed with our posterior mean functions for comparison. The least squares estimated second order TAR model is copied from Tong [114] (p. 377)

$$X_t = \begin{cases} 0.59 + 1.25X_{t-1} - 0.42X_{t-2} + \epsilon_t, & \text{if } X_{t-2} \leq 3.25 \\ 2.23 + 1.52X_{t-1} - 1.24X_{t-2} + \epsilon_t, & \text{if } X_{t-2} > 3.25 \end{cases} \tag{2.31}$$

The comparative plot of the estimated functional coefficients shows that our method's estimates are a lot smoother. It is important to mention here the flexibility of our method compared to the TAR and the LLR estimation methods. We can allow different degrees of smoothing for different functions, and this is why we estimated f_1 to be almost constant. For the TAR model and the LLR method, in order to estimate a constant autoregressive coefficient we would have to use profile least squares. For the spline method this can be avoided, but we would still have to treat the number of knots for each function separately and introduce a constant basis function. Notice also that for both LLR and spline methods the estimates of the coefficient functions f_1 and f_2 have a lot more curvature and that they look like flipped versions of each other, which implies that the estimates are highly correlated.

We next look at the fitted values from the four models, which are plotted in Fig 2.6 and they are all practically indistinguishable. We also look at the predictive performance of these models. We refit the models to the first 102 values from the series and try to predict the remaining 12. We employ two prediction schemes, in the first we do one-step-ahead prediction for the next observation, where we use all previous data as they come along. In the second, we do multi-step-ahead predictions for all 12 future values by iteratively applying one-step-ahead predictions and treating the predicted values as the real data. For our



(a) (b)

Figure 2.5: Plots of nonparametric estimates of functional coefficients (a) f_1 and (b) f_2 for the Canadian lynx data using GP, LLR and splines.

method this means we disregard parameter uncertainty, and we did this in order to make the results comparable. For all models, the parameters are chosen based on the first 102 data. The one-step-ahead predictions are shown in Fig. 2.7 and the multi-step-ahead predictions are shown in Fig. 2.8. The one-step-ahead predictions are very close for all models but this is not surprising, since they all have almost identical fitted values. However, our modeling procedure seems to give improved multi-step-ahead predictions, which follow the true process more closely, with the TAR model coming second.

Based on this last observation, we also perform a graphical test of the dynamics of each fit. We look at the deterministic component of eqn. (2.30) under the different estimates for f_1, f_2, also referred to as the deterministic skeleton of the series. The dynamics of this component play an important role in the dynamics of the resulting series, see e.g. Meyn and Tweedie [84]. We can view the deterministic component for the lynx data as a second order nonlinear difference equation, whose behavior we can study in a 2-dimensional phase space.

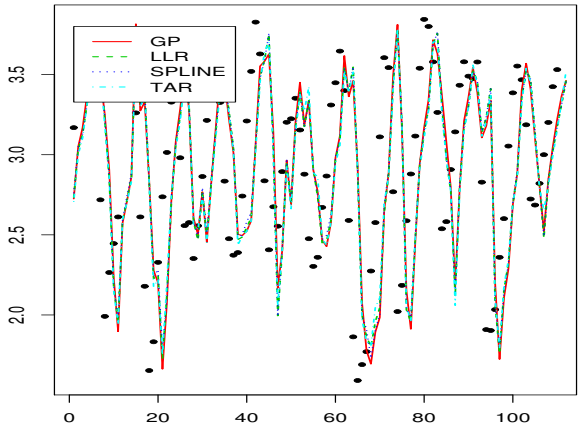


Figure 2.6: Fitted values from all four models applied to the Canadian lynx data, dots represent true values.

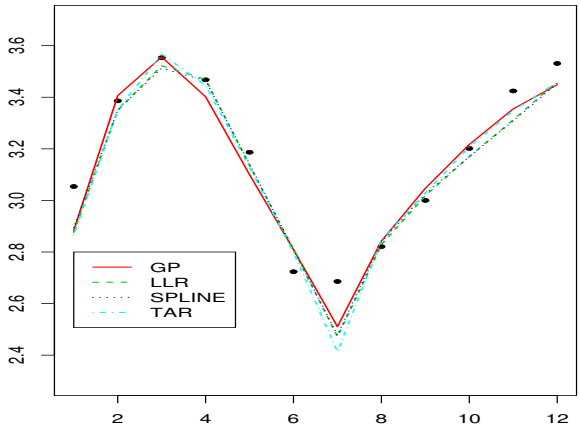


Figure 2.7: One-step-ahead predictions from all three models applied to the Canadian lynx data, dots represent true values.

Fig. 2.9 shows the phase space plot of the data, i.e. the scatter plot of X_t versus X_{t-1}. It is obvious from the plot, and a well documented fact, that the Canadian lynx data have an approximate limit cycle, with a period of around 9.5 years (see Tong [114], p. 375). We

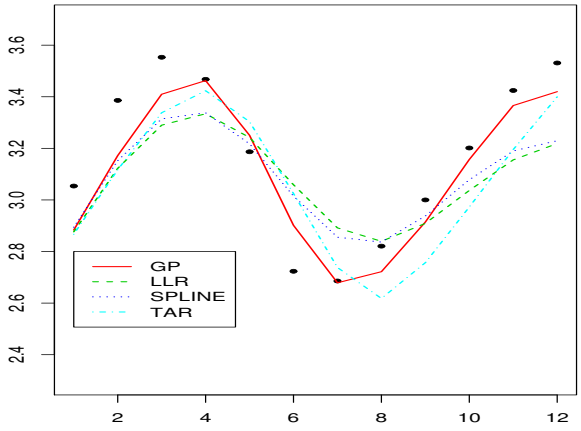


Figure 2.8: Multi-step-ahead predictions from all three models applied to the Canadian lynx data, dots represent true values.

present similar phase space plots for the dynamics of each fit in Fig. 2.10. The initial value of the system, indicated by a cross, correspond to the point (2.033,1.903), the 100^{th} and 99^{th} observation in the series. Although in general the evolution of the system depends on the initial values, we have tried a range of different values and got the same long term behavior. As we can see, only our method and TAR give sustained cyclic behavior, whereas LLR and splines converge to a stable point of around (3.1,3.1). Even though the function estimates of LLR and splines have a lot more curvature, the fitted model dynamics did not capture the qualitative behavior of the series. We believe this happens because the coefficient functions have reflected shapes and opposite signs and the arguments are correlated, so the effects of the nonlinearity of each term cancel out. The important message from this example is that we do not need to allow too much flexibility in the coefficient functions in order to capture the characteristics of the series. In dynamical settings, even small nonlinearities can lead to a very wide spectrum of behaviors and the popularity and practical success of TAR models can attest to that. Our method also tends to produce simple models because it penalizes each function's variability separately. In Chapter 5 we discuss model selection procedures

which are well suited for parsimony.

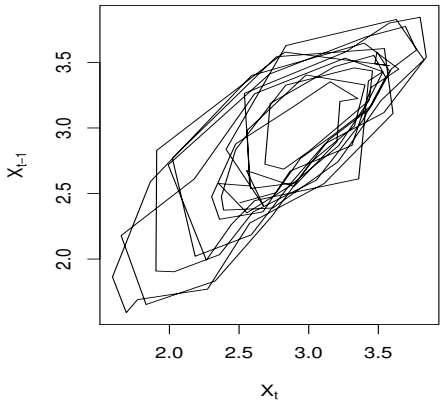


Figure 2.9: Phase space plot of the Canadian lynx data.

2.7 Comments

We now make some general comments on our proposed approach; we begin by putting our Bayesian model in context with respect to the literature. The use of GPs for nonparametric regression dates back to the seminal paper of O'Hagan [88], although it is closely related to earlier methods from spatial statistics, e.g. see Cressie [25]. Incidentally, O'Hagan looks at the varying coefficient model, which is exactly our FAR model for independent data. He provides the relevant posterior formulas, but apart from similarities in the general setting our treatment is very different from his. First of all, O'Hagan uses the same argument for all functional coefficients, so that their prior covariance matrix has a Kronecker product form and they can be correlated a priori. We purposely allow different arguments for each function in order to control their smoothness individually. Moreover, O'Hagan assumes full knowledge of the prior and does not discuss its specification. Finally, we also differ in the practical and theoretical issues that we address in subsequent chapters, and which are

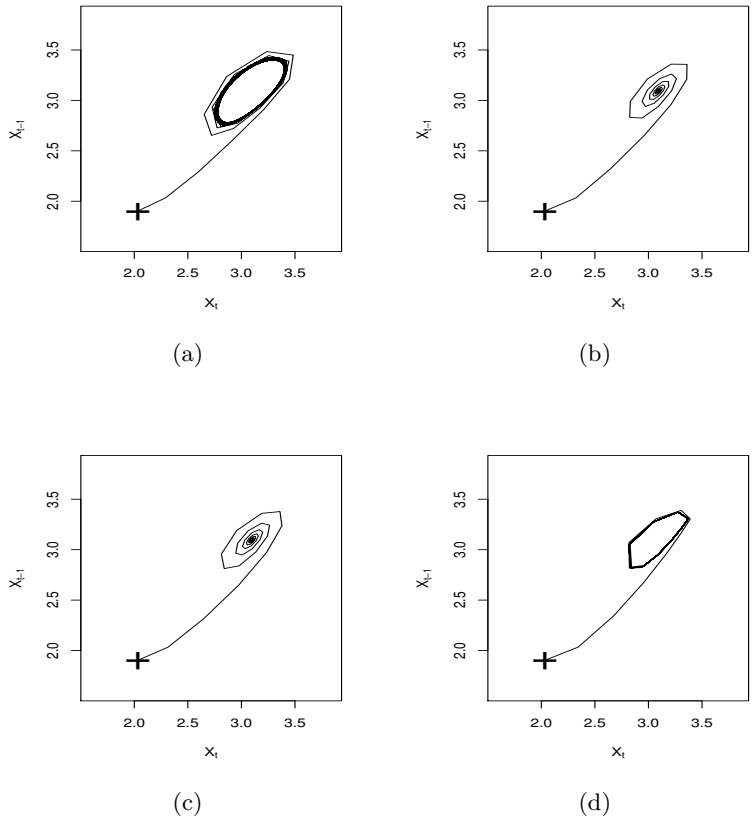


Figure 2.10: Phase space plots for the dynamics of the FAR model (2.30), fitted to the Canadian lynx data using (a) GP, (b) LLR, (c) splines and (d) TAR.

pertinent to time series. Since the early work of O'Hagan, nonparametric GP methods have witnessed tremendous development and there have lately appeared applications explicitly for time series, see Girard et al. [45] and Wang et al. [119]. However, these are based on the NLAR model; to our knowledge, we are the first to approach the nonparametric analysis of the FAR model using GPs.

We also look at the relationship of our method to other models. Our treatment of the coefficients as random variables resembles the RCAR model, but is in fact markedly different. We assume the whole series is generated from a single random draw of coefficient functions,

whereas in the RCAR model each observation is generated using a new i.i.d. coefficient draw. Due to this important distinction we cannot apply available results for the RCAR model, such as those for stability given by Bougerol and Picard [10]. A stronger connection exists between our model and hidden Markov models (HMMs), since we can view the coefficient evaluations as hidden variables, and the observations are conditionally independent given the coefficients and the past. The main discrepancy is that each coefficient evaluation in our model is correlated to all the previous ones, while in HMMs the hidden layer has a finite dependence structure. As a result, our sequential inference requires conditioning on all the past observations, whereas for finite memory HMMs we can apply faster methods, e.g. Kalman filtering. Nevertheless, we exploit this connection after we develop the approximate inference method for our model, which fits into the HMM framework. On a more practical level, we look at the posterior means of the coefficient functions in (2.16) as point estimates. It is not difficult to see that the posterior mean of each f_i is given as a linear combination of squared exponential kernels $\{C_i(\cdot, x_t^{(i)})\}_{t=1}^T$ centered at the observed arguments, plus a prior bias term given by the prior mean function μ_i. This relates to the EXPAR model (1.7), which also uses squared exponentials to represent the coefficient functions, but does so in a parametric way. Each coefficient involves only one exponential, and the locations and shapes for each one are governed by independent parameters. The basic consequence of this similarity between the two methods is that they always give bounded estimates, which is important for the stability of the fitted model.

We move on to give some justification for our choices on the prior, which were motivated by the general characteristics of FAR models and time series. First, we look at the use of arbitrary constant prior means, which is a departure from the conventional use of zero means. This choice is practically inconsequential for the region (in functional coefficient input space) where the main body of the data lies, but it is significant for extrapolating

the functions. By selecting the constant prior means by maximum marginal likelihood, we introduce a bias toward a linear AR model which becomes stronger outside the observed range of the functions. This is helpful in two ways, first it makes the coefficient functions closer to constant, so that our model is more parsimonious. Second and most important, it improves the behavior of predictions or simulations from our model. Basically, when we iterate the FAR model dynamics to generate future paths we cannot control the points at which we have to evaluate the coefficients, which can often lie close to or outside the boundaries of their observed range. This renders the so called boundary behavior of the estimators crucial for the accuracy and stability of the paths. By using our data dependent prior means, we impose a linear AR structure over regions where we do not have enough data, which is more informative than the zero mean alternative and still leads to relatively stable paths.

Changing perspective, we now compare our method to the other nonparametric estimation schemes that have appeared in the literature. As we pointed out before, local/kernel methods, in order to avoid backfitting, are restricted to the case where all functional coefficients share the same argument and, consequently, the same smoothing parameter. Another complication arises when we need to predict the coefficients at the boundary, especially when using kernels with finite support. Often in such cases, the local linear system we need to fit is ill-conditioned or even undefined when it does not contain enough data points, and some correction is required. Turning to our method, it is well known that there is a strong correspondence between GP regression and smoothing splines. In particular, Kimeldorf and Wahba [67] show that a smoothing spline is the posterior mean of a GP regression problem with a special prior and, conversely, that the posterior mean of a GP regression is the solution to a regularization (penalized least-squares) problem in an appropriate function space. The former viewpoint has been applied for Bayesian spline estimation, see Ansley et

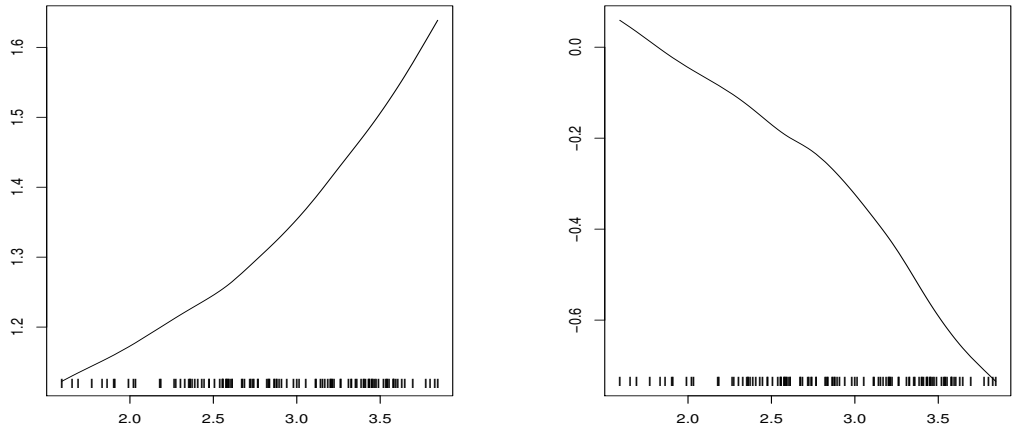
al. [3], and for creating confidence bands around spline curve estimates, see Wahba [118]. Despite this correspondence, there are some substantial practical differences between the two methods. In terms of estimation and prediction, computations for both methods scale the same. In practice, smoothing splines are almost always employed with fewer number of knots than observations, which makes them considerably faster; even so, our approximation scheme, presented in Chapter 3, makes the methods comparable. However, for choosing the smoothing parameters spline methods usually rely on CV and involve grid searches, while for our method we can perform a gradient-based search. The approach of Huang and Shen [60] for the FAR model, also known as regression splines, uses only the number of knots to control the smoothness. This approach is a lot faster than regularized regression since it only requires a grid search over the number of knots, which is usually low, but the search still scales exponentially in the number of functions. In addition, the absence of regularization can lead to unstable estimates at the boundary.

The most important difference between our GP method and splines, though, are with respect to the shape of the estimated functions; and we demonstrate this using our previous example. We fit a smoothing spline to each functional coefficient in model (2.30) for the Canadian lynx data, where we allow different roughness penalties λ_1, λ_2 for each function. We minimize the following objective function

$$\min_{f_1,f_2} \left\{ \sum_t \Big(x_t - x_{t-1} f_1(x_{t-2}) - x_{t-2} f_2(x_{t-2}) \Big)^2 + \sum_{i=1}^{2} \lambda_i \|f_i''\|_2^2 \right\}$$

where the smoothing parameters $\lambda_1 = 0.261$ and $\lambda_2 = 0.099$ are selected by the ordered multi-fold CV procedure suggested in Cai et al. [16]; the plots of the resulting estimates are shown in Fig. 2.11. The two estimates are close to linear and have opposite slopes, because cubic splines only penalize the second derivative of the function. Therefore, a linear function can always be fit at no cost, even for high roughness penalties. Moreover, any type

of splines, regression or smoothing, do linear extrapolation outside the observed range. This means that virtually every function estimate will be unbounded, which affects the stability of the fitted model. In contrast, our GP method, using the squared exponential kernel, penalizes deviations form the mean level, so the coefficient estimates become constant with more smoothing. From a modeling perspective, we claim that the behavior of GPs is more attractive than that of splines because of the type of functions we want to estimate. Imagine a term of the form $X^{(i)} f_i(U^{(i)})$ that we want to fit with regularized regression and with an infinite roughness penalty under both splines and GP. In the former case, we would end up with a term $\alpha X^{(i)} U^{(i)}$ whereas in the latter we would have a term $\alpha X^{(i)}$, for some $\alpha \in \mathbb{R}$. It seems more natural to assume that, in their smoothest form, the coefficient functions are flat rather than linear in the argument, and this also makes interpretation a lot easier because of the absence of interactions between the argument and the regressor, and the similarity to a linear AR model. We believe splines lend themselves better for estimating additive functions instead of coefficient functions, especially in our dynamic setting.



(a) (b)

Figure 2.11: Plots of smoothing spline estimated functional coefficients for (a) f_1 and (b) f_2 for the Canadian lynx series.

Chapter 3

Approximation Methods

Gaussian process regression is computationally expensive, requiring $\mathcal{O}(n^3)$ operations where n is the number of observations. This significantly limits its applicability to relatively small data sets and this has led to the development of more efficient approximate inference methods. Different approximation schemes have been proposed in the literature in order to cope with this problem for the nonparametric regression setting. Gibbs and McKay [44] use iterative solutions to the linear system of equations involved in estimation which scale as $\mathcal{O}(mn^2)$, where m is the number of iterations. Although this offers an improvement, the quadratic factor can still be prohibitive. A faster alternative is based on reduced rank approximations of the prior covariance matrix of the functions. The idea is to approximate the covariance matrix by $\boldsymbol{C} = \boldsymbol{W}\boldsymbol{V}\boldsymbol{W}^\top$, where $\boldsymbol{W}$ is an $n \times m$ matrix and $\boldsymbol{V}$ is a nonsingular $m \times m$ matrix, with $m \ll n$. This affords the use of the Woodbury-Sherman-Morrison formula, also known as the matrix inversion lemma, for the inversion of the matrices used in estimation. The formula states that

$$(\boldsymbol{\Sigma} + \boldsymbol{C})^{-1} = (\boldsymbol{\Sigma} + \boldsymbol{W}\boldsymbol{V}\boldsymbol{W}^\top)^{-1} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{W}(\boldsymbol{V}^{-1} + \boldsymbol{W}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{\Sigma}^{-1} \quad (3.1)$$

The cost of carrying out the computations in (3.1) scales as $\mathcal{O}(m^2 n)$, where m is also the rank of $\boldsymbol{C}$. This is a significant improvement that allows us to work with large numbers of observations. In the following section we give a review of the different reduced rank approximations as they have appeared for nonparametric regression. Next, we extend the method to our FAR setting and describe its practical implementation. We give an example of approximate inference and compare it to exact inference. Finally, we describe how the extension can be used for treating multivariate FAR and state-space models.

3.1 Review of Reduced Rank Approximations

There are three common ways in the literature to approximate the covariance matrix in a reduced rank form. An overview can be found in Rasmussen and Williams [96] and in Quiñonero-Candela and Rasmussen [94]. The first is the Nyström method, which relies on an approximate eigenvalue decomposition of the covariance kernel. The second is the subset of regressors (SR) method which, in effect, represents the functions as a linear combination of kernels centered at a subset of the observations. This is an analogue of smoothing splines with a smaller number of knots than number of data points. The third is the projected process (PP) method, which performs estimation by conditioning (projecting) the Gaussian process on a subset of the data. All three methods lead to reduced rank approximations and they are closely related.

3.1.1 Nyström Method

The Nyström method, which was introduced by Williams and Seeger [122], is based on the numerical eigendecomposition of a kernel $k(\cdot, \cdot)$. The eigenvalues λ_i and eigenfunctions $\phi_i(\cdot)$ of k with respect to the probability measure P satisfy $\lambda_i \phi_i(x') = \int_x k(x', x)\phi_i(x) dP(x)$ (see also Def. A.5 in the Appendix). Assuming we have a sample $\{x_1, \ldots, x_n\}$ from P, we can

perform the integration numerically, to get

$$\lambda_i \phi_i(x') = \int_x k(x', x)\phi_i(x)dP(x) \approx \frac{1}{n}\sum_{k=1}^{n} k(x', x_k)\phi_i(x_k) \tag{3.2}$$

If we substitute the sampled x's in the place of x' we get the system of equations

$$\lambda_i \phi_i(x_j) = \frac{1}{n}\sum_{k=1}^{n} k(x_j, x_k)\phi_i(x_k) \ , \ \ \forall i, j = 1, \ldots, n$$

which translates to the matrix eigenproblem $\lambda_i \boldsymbol{\phi}_i = \boldsymbol{K}\boldsymbol{\phi}_i$ for $i = 1, \ldots, n$, where $\boldsymbol{K} = [\{k(x_i, x_j)\}_{i,j=1}^{n}]$ and $\boldsymbol{\phi}_i^\top = [\phi_i(x_1), \ldots, \phi_i(x_n)]$. Based on this, we look at the eigendecomposition $\lambda_i^\star \boldsymbol{u}_i = \boldsymbol{K}\boldsymbol{u}_i \ , \ \ i = 1, \ldots, n$ of the matrix $\boldsymbol{K}$, where $\lambda_i^\star$ is the i^{th} eigenvalue and $\boldsymbol{u}_i$ is it's corresponding normalized eigenvector. A straightforward estimator of the eigenvalues is $\tilde{\lambda}_i = \lambda_i^\star / n$, and the elements of $\boldsymbol{u}_i$ are estimates of the eigenfunction evaluations at the observed x's, that is $\tilde{\phi}_i(x_j) = \sqrt{n}(\boldsymbol{u}_i)_j$. The $\sqrt{n}$ factor is a result of the different normalization between the eigenvectors and the eigenfunctions.

The Nyström method uses equation (3.2) to extend the eigenfunction approximation to the entire range of x, giving $\tilde{\phi}_i(x) = \frac{\sqrt{n}}{\lambda_i^\star}\boldsymbol{k}(x)^\top \boldsymbol{u}_i$, where $\boldsymbol{k}(\cdot) = [k(\cdot, x_1), \ldots, k(\cdot, x_n)]^\top$. The index i runs only up to n, thus we can only approximate the first n eigenvalues and eigenvectors of the kernel. The kernel k is approximated by

$$\begin{aligned} \tilde{k}(x, x') &= \sum_{i=1}^{n} \tilde{\lambda}_i \tilde{\phi}_i(x)\tilde{\phi}_i(x') = \sum_{i=1}^{n} \frac{1}{\lambda_i^\star}\boldsymbol{k}(x)^\top \boldsymbol{u}_i \boldsymbol{k}(x')^\top \boldsymbol{u}_i \\ &= \boldsymbol{k}(x)^\top \left(\sum_{i=1}^{n} \frac{1}{\lambda_i^\star}\boldsymbol{u}_i \boldsymbol{u}_i^\top \right) \boldsymbol{k}(x') = \boldsymbol{k}(x)^\top \boldsymbol{K}^{-1}\boldsymbol{k}(x') \end{aligned} \tag{3.3}$$

This does not yet provide any computational improvement since the inversion of $\boldsymbol{K}$ requires $\mathcal{O}(n^3)$ calculations. The main idea behind the Nyström method is to base the approximation on a small subset of $m \ll n$ data points. Without loss of generality, let this subset be

$\{x_1, \ldots, x_m\}$ and also suppose $k(\cdot, \cdot)$ is a covariance kernel. We want to approximate the covariance matrix $\boldsymbol{K}$, which can be written in partitioned form as

$$\boldsymbol{K} = \left[\begin{array}{c|c} \boldsymbol{K}_{m,m} & \boldsymbol{K}_{m,n-m} \\ \hline \boldsymbol{K}_{n-m,m} & \boldsymbol{K}_{n-m,n-m} \end{array}\right]$$

where $\boldsymbol{K}_{m,m} = [\{k(x_i, x_j)\}_{i,j=1}^{m}]$, $\boldsymbol{K}_{n-m,n-m} = [\{k(x_i, x_j)\}_{i,j=m+1}^{n}]$, and $\boldsymbol{K}_{m,n-m} = \boldsymbol{K}_{n-m,m}^{\top} = [\{k(x_i, x_j)\}_{i=1,j=m+1}^{m,n}]$. The Nyström approximation of $\boldsymbol{K}$, based on the eigendecomposition of $\boldsymbol{K}_{m,m}$, becomes

$$\tilde{\boldsymbol{K}} = \boldsymbol{K}_{n,m} \boldsymbol{K}_{m,m}^{-1} \boldsymbol{K}_{m,n} \tag{3.4}$$

where $\boldsymbol{K}_{m,n} = \boldsymbol{K}_{n,m}^{\top} = [\boldsymbol{K}_{m,m}, \boldsymbol{K}_{m,n-m}]$.

As an example of how the approximation is applied, consider a simple nonparametric GP regression, where we observe data points$\{(y_i, x_i)\}_{i=1}^{n}$ coming from the model

$$\begin{aligned} y_i | f, x_i &\sim \mathcal{N}(f(x_i), \sigma^2) \\ f &\sim \mathcal{GP}(0, k) \end{aligned}$$

The covariance function of f is given by the kernel $k(\cdot, \cdot)$ which we approximate using the Nyström method. The resulting formulas for the posterior mean and variance of f, evaluated at an arbitrary point x', are

$$\mathrm{E}_{Nys}[f(x') | \boldsymbol{x}, \boldsymbol{y}] = \boldsymbol{k}(x')^{\top} (\tilde{\boldsymbol{K}} + \boldsymbol{\Sigma})^{-1} \boldsymbol{y} \tag{3.5}$$

$$\mathrm{Var}_{Nys}[f(x') | \boldsymbol{x}, \boldsymbol{y}] = k(x', x') - \boldsymbol{k}(x')^{\top} (\tilde{\boldsymbol{K}} + \boldsymbol{\Sigma})^{-1} \boldsymbol{k}(x) \tag{3.6}$$

with the notation carrying over from our previous exposition and that of Chapter 2 in an

obvious way. Moreover, The marginal likelihood is approximated by

$$\ell_{Nys}(\boldsymbol{y}|\boldsymbol{x}) = -\frac{1}{2}\left(\log|\tilde{\boldsymbol{K}}+\boldsymbol{\Sigma}| + \boldsymbol{y}^\top(\tilde{\boldsymbol{K}}+\boldsymbol{\Sigma})^{-1}\boldsymbol{y} + n\log(2\pi)\right) \tag{3.7}$$

The matrix $\tilde{\boldsymbol{K}}$ is in reduced rank form and the computations for estimation, i.e. inversion and determinant of $(\tilde{\boldsymbol{K}}+\boldsymbol{\Sigma})^{-1}$, scale as $\mathcal{O}(m^2 n)$. For the latter operation, we use the analogue of the matrix inversion lemma for determinants, which states that $|\boldsymbol{\Sigma}+\boldsymbol{W}\boldsymbol{V}\boldsymbol{W}^\top| = |\boldsymbol{\Sigma}|\ |\boldsymbol{V}|\ |\boldsymbol{V}^{-1}+\boldsymbol{W}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{W}|$. For comparison, we also give the exact formulas for GP regression.

$$\begin{aligned}
\mathrm{E}[f(x')|\boldsymbol{x},\boldsymbol{y}] &= \boldsymbol{k}(x')^\top(\boldsymbol{K}+\boldsymbol{\Sigma})^{-1}\boldsymbol{y} && (3.8)\\
\mathrm{Var}[f(x')|\boldsymbol{x},\boldsymbol{y}] &= k(x',x') - \boldsymbol{k}(x')^\top(\boldsymbol{K}+\boldsymbol{\Sigma})^{-1}\boldsymbol{k}(x') && (3.9)\\
\ell(\boldsymbol{y}|\boldsymbol{x}) &= -\frac{1}{2}\left(\log|\boldsymbol{K}+\boldsymbol{\Sigma}| + \boldsymbol{y}^\top(\boldsymbol{K}+\boldsymbol{\Sigma})^{-1}\boldsymbol{y} + n\log(2\pi)\right) && (3.10)
\end{aligned}$$

Before we present the subset of regressors method, we take a closer look at the previous non-parametric regression model. The posterior mean of f, as given in (3.8), can be represented as a linear combination of n kernels centered at the observed x's

$$\mathrm{E}[f(x')|\boldsymbol{x},\boldsymbol{y}] = \sum_{i=1}^{n}\alpha_i k(x',x_i) = \boldsymbol{\alpha}^\top\boldsymbol{k}(\cdot) \tag{3.11}$$

where $\boldsymbol{\alpha} = (\boldsymbol{K}+\boldsymbol{\Sigma})^{-1}\boldsymbol{y}$. Based on this observation in a similar regularized regression setting, Silverman [107] proposed a finite dimensional Bayesian model that gives the same posterior mean. In particular, we can view the function f as a linear combination $f(\cdot) = \boldsymbol{\alpha}^\top\boldsymbol{k}(\cdot)$ of fixed kernels centered at the observed data, with random coefficients $\boldsymbol{\alpha}$. Putting the specific Gaussian prior $\boldsymbol{\alpha} \sim \mathcal{N}(0, \boldsymbol{K}^{-1})$ on the coefficients, their posterior distribution becomes

$$\boldsymbol{\alpha}|\boldsymbol{x},\boldsymbol{y} \sim \mathcal{N}\big((\boldsymbol{K}+\boldsymbol{\Sigma})^{-1}\boldsymbol{y}, (\boldsymbol{K}+\boldsymbol{K}\boldsymbol{\Sigma}^{-1}\boldsymbol{K})^{-1}\big)$$

As a result, the posterior mean of f for this finite dimensional model is exactly the one in (3.11). It is important to note that this finite dimensional Bayes model is not equivalent to GP regression, the difference lies in the posterior covariances. Specifically, the posterior variance of f at a point x' is $\boldsymbol{k}(x')^\top(\boldsymbol{K}+\boldsymbol{K}\boldsymbol{\Sigma}^{-1}\boldsymbol{K})^{-1}\boldsymbol{k}(x')$ for the finite dimensional Bayes model. Thus, if the point x' is far from the observed x_i's (so that the vector $\boldsymbol{k}(x')$ is close to zero) the posterior variance will tend to zero, whereas for GP it will always be positive, going to $k(x',x')$.

3.1.2 Subset of Regressors

As the name suggests, the SR method uses only a subset of the data on which to base the representation. Again, we assume a subset $\{x_1,\ldots,x_m\}$ of size m, so that f can be represented as

$$f(\cdot)=\sum_{i=1}^{m}\alpha_i k(\cdot,x_i)=\boldsymbol{\alpha}_m^\top\boldsymbol{k}_m(\cdot) \tag{3.12}$$

and we put the prior $\boldsymbol{\alpha}_m\sim\mathcal{N}(0,\boldsymbol{K}_{m,m}^{-1})$ on the coefficients. The formulas for the posterior mean and variance using the SR method are

$$\mathrm{E}_{SR}[f(x')|\boldsymbol{x},\boldsymbol{y}] = \boldsymbol{k}_m(x')^\top(\boldsymbol{K}_{m,m}+\boldsymbol{K}_{m,n}\boldsymbol{\Sigma}^{-1}\boldsymbol{K}_{n,m})^{-1}\boldsymbol{K}_{m,n}\boldsymbol{\Sigma}^{-1}\boldsymbol{y} \tag{3.13}$$

$$\mathrm{Var}_{SR}[f(x')|\boldsymbol{x},\boldsymbol{y}] = \boldsymbol{k}_m(x')^\top(\boldsymbol{K}_{m,m}+\boldsymbol{K}_{m,n}\boldsymbol{\Sigma}^{-1}\boldsymbol{K}_{n,m})^{-1}\boldsymbol{k}_m(x') \tag{3.14}$$

and the marginal likelihood is equal to that of the Nyström method. To verify that the SR method is a reduced rank approximation, we look at the prior covariance matrix of the vector $\boldsymbol{f}=[f(x_1),\ldots,f(x_n)]^\top$ of function evaluations at the observed data points, which is simply $\boldsymbol{K}_{n,m}\boldsymbol{K}_{mm}^{-1}\boldsymbol{K}_{m,n}$. This is exactly the matrix $\tilde{\boldsymbol{K}}$ from equation (3.4), and in fact there is a close relationship between the Nyström and the SR method, they both reach the

same reduced rank approximation through different avenues. The main difference is that the SR method is formulated as a probabilistic model, so it gives consistent results, whereas the Nyström method only substitutes an approximation $\tilde{\boldsymbol{K}}$ for the covariance matrix $\boldsymbol{K}$ wherever it appears in the exact GP regression formulas. The latter approach is inconsistent because some posterior variances can be negative. However, Rasmussen and Williams [96] show that if we systematically substitute the Nyström approximation (3.3) for the kernel k in the exact formulas (3.8-3.9), we get the SR method. For all other practical considerations, the two methods behave similarly.

3.1.3 Projected Process

The SR method has the disadvantage that, being a finite dimensional model, the posterior variance of predictions can go to zero; the projected process approximation of Seeger et al. [103] can amend this. Again, the method is based on a subset of the data $\{x_1, \ldots, x_m\}$. Consider a vector of evaluations at this set $\boldsymbol{f}_m = [f(x_1), \ldots, f(x_m)]^\top$, and let $\boldsymbol{f}_{n-m}$ be the rest of the evaluations. From the GP prior, we know that $\mathrm{E}[\boldsymbol{f}_{n-m}|\boldsymbol{f}] = \boldsymbol{K}_{n-m,m}\boldsymbol{K}_{m,m}^{-1}\boldsymbol{f}_m$. The idea behind the method is to replace the data likelihood $\mathcal{L}(\boldsymbol{y}|\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}, \boldsymbol{\Sigma})$, which depends on the entire $\boldsymbol{f}$, with the following approximation

$$\tilde{\mathcal{L}}(\boldsymbol{y}|\boldsymbol{f}) \quad = \quad \mathcal{N}(\mathrm{E}[\boldsymbol{f}|\boldsymbol{f}_m], \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{K}_{n,m}\boldsymbol{K}_{m,m}^{-1}\boldsymbol{f}_m, \sigma^2\boldsymbol{I}) \tag{3.15}$$

which depends only on $\boldsymbol{f}_m$. The name of the method comes from the fact that we project the entire process on $\boldsymbol{f}_m$. Seeger [101] gives a theoretical justification behind this approximation in terms of the Kullback-Leibler divergence of the posterior distributions $p(\boldsymbol{f}|\boldsymbol{y})$ and $\tilde{p}(\boldsymbol{f}|\boldsymbol{y})$ of $\boldsymbol{f}$ under the exact and the approximate method, respectively. In particular, he shows that $\tilde{p}(\boldsymbol{f}|\boldsymbol{y})$ minimizes $D_{KL}\big(q(\boldsymbol{f}|\boldsymbol{y})\|p(\boldsymbol{f}|\boldsymbol{y})\big)$ over all distributions q of the form $q(\boldsymbol{f}|\boldsymbol{y}) \propto q_1(\boldsymbol{y}|\boldsymbol{f}_m)q_2(\boldsymbol{f})$. In our case, $q_2(\boldsymbol{f})$ is the GP prior and $q_1(\boldsymbol{y}|\boldsymbol{f}_m)$ is the likelihood approximation in (3.15). The resulting posterior mean and marginal likelihood is exactly

the same as that of SR, but the variance now becomes

$$\begin{aligned} \mathrm{Var}_{PP}[f(x')|\boldsymbol{x},\boldsymbol{y}] &= k(x',x') - \boldsymbol{k}_m(x')^\top \boldsymbol{K}_{m,m}^{-1}\boldsymbol{k}_m(x') \\ &+ \boldsymbol{k}_m(x')^\top(\boldsymbol{K}_{m,m} + \boldsymbol{K}_{m,n}\boldsymbol{\Sigma}^{-1}\boldsymbol{K}_{n,m})^{-1}\boldsymbol{k}_m(x') \end{aligned} \tag{3.16}$$

This means that even if x' is far form the range of the basis points, the posterior variance of $f(x')$ will still be positive.

3.2 Reduced Rank Approximations for FAR Model

In this section we present our adaptation of the PP reduced rank approximation to the general FAR model $Y_t = X_t^{(1)} f_1(U_t^{(1)}) + \ldots + X_t^{(p)} f_p(U_t^{(p)}) + \epsilon_t$. Suppose that we have T observations from this time series model (instead of n for regression) , where each function has a GP prior with mean function $\mu_i(\cdot)$ and covariance kernel $C_i(\cdot,\cdot)$, and let $\{B_j^{(i)}\}_{j=1}^{m_i}$ be a collection of points in the space of $U^{(i)}$, which need not be observed values of $U^{(i)}$. For example, if $U^{(i)}$ is one dimensional they can be an equally spaced or quantile sequence in the range of $U^{(i)}$. These are the points on which we base the m_i^{th} order reduced rank approximation for each f_i, and they can be thought of as the knots in smoothing splines. We define the following

$$\begin{aligned} \boldsymbol{f}_{B,i}^\top &= \left[f_i(B_1^{(i)}), f_i(B_2^{(i)}), \ldots, f_i(B_{m_i}^{(i)})\right], \qquad \boldsymbol{f}_B^\top = [\boldsymbol{f}_{B,1}^\top, \boldsymbol{f}_{B,2}^\top, \ldots, \boldsymbol{f}_{B,p}^\top] \\ \boldsymbol{\mu}_{B,i}^\top &= \left[\mu_i(B_1^{(i)}), \mu_i(B_2^{(i)}), \ldots, \mu_i(B_{m_i}^{(i)})\right], \qquad \boldsymbol{\mu}_B^\top = \left[\boldsymbol{\mu}_{B,1}^\top, \boldsymbol{\mu}_{B,2}^\top, \ldots, \boldsymbol{\mu}_{B,p}^\top\right] \\ \boldsymbol{C}_{B,i} &= \left[\left\{C_i\left(B_j^{(i)}, B_k^{(i)}\right)\right\}_{j,k=1}^{m_i}\right], \quad \boldsymbol{C}_{U,i} = \left[\left\{C_i\left(U_t^{(i)}, B_j^{(i)}\right)\right\}_{t=1,j=1}^{T,m_i}\right] \end{aligned}$$

$$\boldsymbol{C}_B = \begin{bmatrix} \boldsymbol{C}_{B,1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C}_{B,2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{C}_{B,p} \end{bmatrix}, \quad \boldsymbol{C}_U = \begin{bmatrix} \boldsymbol{C}_{U,1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C}_{U,2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{C}_{U,p} \end{bmatrix}$$

The vector $\boldsymbol{f}_B$ represents the functional coefficient evaluations at the basis set and $\boldsymbol{\mu}_B, \boldsymbol{C}_B$ are its corresponding prior mean and covariance matrix. The matrix $\boldsymbol{C}_U$ provides the covariance between the function evaluations at the basis points and the observed arguments. After standard calculations, the posterior mean and covariance of $\boldsymbol{f}_B$ are

$$\begin{aligned} \mathrm{E}[\boldsymbol{f}_B|\boldsymbol{y}] &= \boldsymbol{\mu}_B + \boldsymbol{C}_B(\boldsymbol{C}_B + \boldsymbol{C}_U^\top \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top \boldsymbol{C}_U)^{-1}\boldsymbol{C}_U^\top \boldsymbol{X}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}^\top\boldsymbol{\mu}) \quad (3.17) \\ \mathrm{Var}[\boldsymbol{f}_B|\boldsymbol{y}] &= \boldsymbol{C}_B(\boldsymbol{C}_B + \boldsymbol{C}_U^\top \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top \boldsymbol{C}_U)^{-1}\boldsymbol{C}_B \quad (3.18) \end{aligned}$$

and the logarithm of the marginal likelihood of the data is

$$\begin{aligned} \ell(y) &= -\frac{1}{2}\Big[T\log(2\pi) + \log(|\boldsymbol{\Sigma} + \boldsymbol{X}^\top \boldsymbol{C}_U \boldsymbol{C}_B^{-1}\boldsymbol{C}_U^\top \boldsymbol{X}|) + \\ &\quad + (\boldsymbol{y} - \boldsymbol{X}^\top\boldsymbol{\mu})^\top \left(\boldsymbol{\Sigma} + \boldsymbol{X}^\top \boldsymbol{C}_U \boldsymbol{C}_B^{-1}\boldsymbol{C}_U^\top \boldsymbol{X}\right)^{-1}(\boldsymbol{y} - \boldsymbol{X}^\top\boldsymbol{\mu})\Big] \\ &= -\frac{1}{2}\Big[T\log(2\pi|\boldsymbol{\Sigma}|) + \log(|\boldsymbol{I} + \boldsymbol{C}_U^\top \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top \boldsymbol{C}_U \boldsymbol{C}_B^{-1}|) + \\ &\quad + \boldsymbol{r}^\top \left(\boldsymbol{\Sigma} - \boldsymbol{X}^\top \boldsymbol{C}_U(\boldsymbol{C}_B + \boldsymbol{C}_U^\top \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top \boldsymbol{C}_U)^{-1}\boldsymbol{C}_U^\top \boldsymbol{X}\right)\boldsymbol{r}\Big] \quad (3.19) \end{aligned}$$

where $\boldsymbol{r} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}^\top\boldsymbol{\mu})$. For predictions, we want to find the posterior distribution of a vector f_N of function evaluations at arbitrary points $\{N_k^{(i)}\}_{k=1}^{n_i}$ for each f_i. Similarly, we define

$$\begin{aligned} \boldsymbol{f}_{N,i}^\top &= \left[f_i(N_1^{(i)}), f_i(N_2^{(i)}), \dots, f_i(N_{n_i}^{(i)})\right], \qquad \boldsymbol{f}_N^\top = [\boldsymbol{f}_{N,1}^\top, \boldsymbol{f}_{N,2}^\top, \dots, \boldsymbol{f}_{N,p}^\top] \\ \boldsymbol{\mu}_{N,i}^\top &= \left[\mu_i(N_1^{(i)}), \mu_i(N_2^{(i)}), \dots, \mu_i(N_{n_i}^{(i)})\right], \qquad \boldsymbol{\mu}_N^\top = \left[\boldsymbol{\mu}_{N,1}^\top, \boldsymbol{\mu}_{N,2}^\top, \dots, \boldsymbol{\mu}_{Np}^\top\right] \end{aligned}$$

$$\boldsymbol{C}_{NN,i} = \left[\left\{C_i\left(N_j^{(i)}, N_k^{(i)}\right)\right\}_{j,k=1}^{n_i}\right], \quad \boldsymbol{C}_{N,i} = \left[\left\{C_i\left(N_k^{(i)}, B_j^{(i)}\right)\right\}_{k=1,j=1}^{n_i,m_i}\right]$$

$$\boldsymbol{C}_{NN} = \begin{bmatrix} \boldsymbol{C}_{NN,1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C}_{NN,2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{C}_{NN,p} \end{bmatrix}, \quad \boldsymbol{C}_{N} = \begin{bmatrix} \boldsymbol{C}_{N,1} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C}_{N,2} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{C}_{N,p} \end{bmatrix}$$

The posterior distribution of $\boldsymbol{f}_N$ is induced from (3.17-3.18) and is normal with mean and variance given by

$$\mathrm{E}[\boldsymbol{f}_N|\boldsymbol{y}] = \boldsymbol{\mu}_N + \boldsymbol{C}_N(\boldsymbol{C}_B + \boldsymbol{C}_U^\top \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top \boldsymbol{C}_U)^{-1}\boldsymbol{C}_U^\top \boldsymbol{X}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}^\top\boldsymbol{\mu}) \quad (3.20)$$

$$\mathrm{Var}[\boldsymbol{f}_N|\boldsymbol{y}] = C_{NN} - \boldsymbol{C}_N\boldsymbol{C}_B^{-1}\boldsymbol{C}_N^\top + \boldsymbol{C}_N(\boldsymbol{C}_B + \boldsymbol{C}_U^\top \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top \boldsymbol{C}_U)^{-1}\boldsymbol{C}_N^\top \quad (3.21)$$

In terms of computational efficiency, our PP approximation scheme scales as $\mathcal{O}\left(T(\sum_{i=1}^p m_i)^2\right)$ for the posterior means and variances, and the marginal likelihood (we only look at variances, because for covariances we would have to compute the entire $T \times T$ covariance matrix). In order to carry out these operations, we only need to invert matrices of the dimension of $\boldsymbol{C}_B$, which scale as $\mathcal{O}\left((\sum_{i=1}^p m_i)^3\right)$, and all the relevant matrix multiplications can be performed in $\mathcal{O}\left(T(\sum_{i=1}^p m_i)^2\right)$ computations by using the special structure of $\boldsymbol{\Sigma}$ and $\boldsymbol{X}$, and following a convenient ordering for multiplication. Thus, the computations in this case depend also on the order/number of regressors p of the FAR model, unlike the exact inference where they only depend on the number of observations. For calculating the score functions in addition, the procedure scales as $\mathcal{O}\left(pT(\sum_{i=1}^p m_i)^2\right)$ because we have $2p+1$ hyperparameters; the formulas for the score functions follow readily from eqn. (3.19) upon differentiation. To summarize, in a realistic setting where we estimate the hyperparame-

ters and use the same number of bases m for each function, the whole procedure scales as $\mathcal{O}\left(Tp^3m^2\right)$. This can still be high in relation to the number of terms p, but later we look at methods for substituting varying coefficients with constants (in effect replacing m basis points with only one), which offers further improvements. Finally, we discuss online estimation for the approximation method, which essentially reduces to updating the inverse of $\boldsymbol{A} = (\boldsymbol{C}_B + \boldsymbol{C}_U^\top \boldsymbol{X}\boldsymbol{\Sigma}^{-1}\boldsymbol{X}^\top \boldsymbol{C}_U)$. To isolate the contribution of each observation, we rewrite the last matrix as $\boldsymbol{A} = \boldsymbol{C}_B + \sum_{t=1}^{T} \boldsymbol{w}_t \boldsymbol{w}_t^\top$, where $\boldsymbol{w}_t$ is the t^{th} row of the matrix $\boldsymbol{W} = \boldsymbol{X}^\top \boldsymbol{C}_U \boldsymbol{\Sigma}^{-1/2}$. Each row of $\boldsymbol{W}$ involves only variables associated with the corresponding observation and the basis points. Now, assuming we get a new observation with a new $\boldsymbol{w}_{T+1}$, we need the inverse of $\boldsymbol{A}'^{-1} = (\boldsymbol{C}_B + \sum_{t=1}^{T+1} \boldsymbol{w}_t \boldsymbol{w}_t^\top)^{-1} = (\boldsymbol{A} + \boldsymbol{w}_{T+1}\boldsymbol{w}_{T+1}^\top)^{-1}$. Given knowledge of $\boldsymbol{A}^{-1}$, we can use the matrix inversion lemma to get

$$\boldsymbol{A}'^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{w}_{T+1}(\boldsymbol{I}_T + \boldsymbol{w}_{T+1}^\top \boldsymbol{A}^{-1}\boldsymbol{w}_{T+1})^{-1}\boldsymbol{w}_{T+1}^\top \boldsymbol{A}^{-1} \tag{3.22}$$

which can be computed in quadratic time in the dimension of $\boldsymbol{C}_B$. In effect, each new observation adds a rank one perturbation to the matrix $\boldsymbol{A}$, and sequential estimation scales the same as batch estimation.

3.3 Implementation

Before we discuss practical choices in the implementation of our reduced rank approximation, we point out some important qualitative differences between the regression setting and our FAR model. First, in GP regression the input space is typically high dimensional, contrary to the FAR model where each argument $U_t^{(i)}$ of f_i will usually be one dimensional. This is a consequences of our intension to avoid the curse of dimensionality. Moreover, the behavior of the functions we try to estimate is fundamentally different. In GP regression, there is no limitation on the shape of the regression function f. In the FAR model on

the other hand, the functions f_i we try to estimate are coefficients to be multiplied with lagged values of the process, or other exogenous time series, in a dynamic setting, so even relatively simple functions f_i can represent a large spectrum of behaviors. Moreover, in the Markov case we expect the functions to be small in absolute value, usually less than one, for the process not to be explosive. Therefore, and for reasons of parsimony, we believe that the functional coefficients should not be very variable, and we should be able to even allow them to be constant.

We now translate these observations in the context of the practical implementation of reduced rank approximations to the FAR model. For GP regression, the covariance matrix $\boldsymbol{K}$ usually has high rank, so the order of the approximation and the composition of the basis set play an important role. For this reason greedy algorithms are used to decide which and how many basis points should be used, e.g. see Lawrence et al. [71] and Seeger et al. [102]. For the FAR model, however, the rank of the covariance matrix of each function is typically much smaller than the number of observations, because the functions are smooth and low dimensional. This allows us to make ad hoc choices which do not significantly compromise the accuracy of the approximation. Practically, we only need a small rank for each function, in the order of tens, with the basis set for each function f_i being an equally or quantile spaced sequence in the observed range of its argument $U^{(i)}$. We generally use equal spacing, except when the observed arguments take extreme values, in which case it could be preferable to use quantile-based spacing. We have found that for applications where the functional coefficients are one dimensional this approach behaves quite well.

Another important aspect of the implementation of our reduced rank approximation is the invertibility of the basis prior covariance matrix $\boldsymbol{C}_B$. This is a block diagonal matrix, so its invertibility is determined by that of its diagonal blocks $\boldsymbol{C}_{B,i}$, i.e. the basis prior covariance

matrices for each functional coefficient f_i. For practical applications, the matrices

$$\boldsymbol{C}_{B,i} = \left[\{C_i(B_j, B_k)\}_{j,k=1}^{m_i}\right] = \left[\nu_i^2 \left\{\exp\left(-\frac{\|B_j - B_k\|^2}{h_i^2}\right)\right\}_{j,k=1}^{m_i}\right]$$

end up being computationally singular even for moderate values of the characteristic length-scale h_i and small numbers of basis points m_i. This singularity does not pose a problem for the exact method, since we need the inverse of $(\boldsymbol{X}^\top \boldsymbol{C} \boldsymbol{X} + \boldsymbol{\Sigma})^{-1}$ which is stable because of the added diagonal matrix $\boldsymbol{\Sigma}$. But the formulas for our reduced rank approximation necessitate the inversion of $(\boldsymbol{C}_B + \boldsymbol{C}_U^\top \boldsymbol{X} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}^\top \boldsymbol{C}_U)$, which relies on $\boldsymbol{C}_B^{-1}$ for applying the matrix inversion lemma. It might seem that we could use a QR decomposition on $\boldsymbol{C}_B = \boldsymbol{Q}\boldsymbol{R}$ and apply the lemma in the opposite direction

$$(\boldsymbol{C}_B + \boldsymbol{C}_U^\top \boldsymbol{X} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}^\top \boldsymbol{C}_U)^{-1} = (\boldsymbol{Q}\boldsymbol{R} + \boldsymbol{F})^{-1} = \boldsymbol{F}^{-1} - \boldsymbol{F}^{-1}\boldsymbol{Q}(\boldsymbol{I} + \boldsymbol{R}\boldsymbol{F}^{-1}\boldsymbol{Q})^{-1}\boldsymbol{R}\boldsymbol{F}^{-1}$$

where $\boldsymbol{F} = \boldsymbol{C}_U^\top \boldsymbol{X} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}^\top \boldsymbol{C}_U$. Unfortunately, for practical applications the matrix $\boldsymbol{F}$ is almost always singular and this approach does not work in general. One could argue in favor of reducing the number of basis points for f_i in order to make $\boldsymbol{C}_i$ invertible. This approach would be valid if we knew the values of the hyperparameters, especially h_i, beforehand, so that we could adjust the number of bases according to the smoothness of the function. But for any realistic application the hyperparameters need to be selected by maximizing the marginal likelihood and this requires evaluations at arbitrary points in the hyperparameter space. Neither can we adjust the number of bases on the fly, as we move through the hyperparameter space, because we would be essentially comparing different model specifications, sacrificing the consistency of our method and introducing discontinuities in the likelihood (also prohibiting a gradient-based search). Therefore, we cannot bypass the need to ensure that $\boldsymbol{C}_B$ is invertible for any number of basis points and hyperparameter values.

The way we choose to address this issue is by slightly altering the covariance function in order to always give us an invertible $\boldsymbol{C}_B$. We propose two different schemes, the first one uses the covariance function

$$C'(x,y) = \nu^2 \left[\exp\left(-\frac{\|x-y\|^2}{h^2} \right) + \epsilon\delta(x-y) \right]$$

where ϵ is some small positive number and $\delta(\cdot)$ is the Kronecker delta function. What we actually achieve by this is to add a small diagonal term to $\boldsymbol{C}_B$ in order to stabilize the matrix; the new basis covariance matrix being $\boldsymbol{C}'_B = \boldsymbol{C}_B + \epsilon\boldsymbol{I}$. This method can be quite rough, because the perturbation only appears on $\boldsymbol{C}'_B$ and not on $\boldsymbol{C}'_U$ (unless there are exact ties between the basis set and the observed arguments). A smoother perturbation results from the covariance function

$$C'(x,y) = \nu^2 \left[\exp\left(-\frac{\|x-y\|^2}{h^2} \right) + \epsilon \exp\left(-\frac{\|x-y\|^2}{h^{\star 2}} \right) \right]$$

where ϵ is as before, and $h^\star$ is a fixed bandwidth, specified from the basis set in order to always give a non singular $\boldsymbol{C}_B$. For practical applications, we propose setting $h^\star$ equal to the minimum distance between the basis points. This perturbation is smoother, in the sense that it also affects the rectangular matrix $\boldsymbol{C}'_U$.

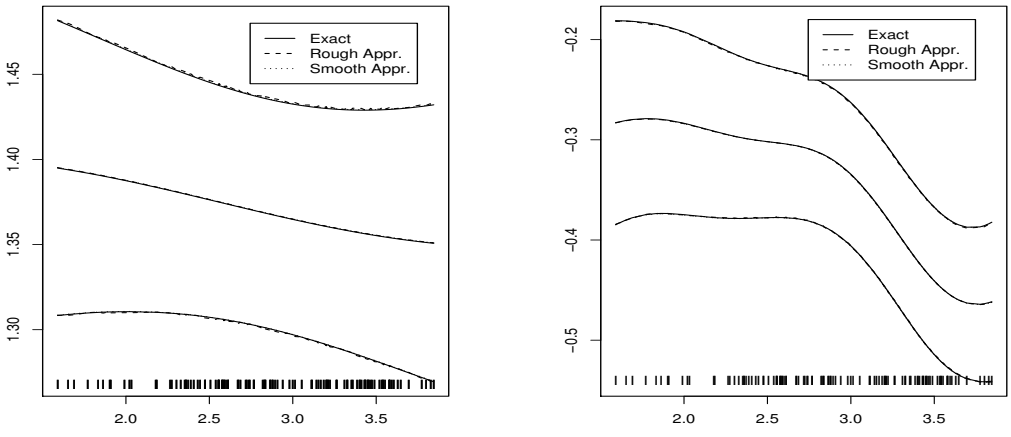
We now give a heuristic discussion about the magnitude of ϵ, the parameter in the covariance function perturbation. Assume that we want to approximate the covariance matrix of a function, and we have decided on the number m of basis points. We want to make sure that the covariance matrix of the basis is numerically stable for inversion, and for this reason we can just look at a worst case scenario which happens when the bandwidth is infinite (i.e. the function is constant). The parameter ν^2 is irrelevant because it divides the entire matrix, so without loss of generality we can assume it is equal to one and that

we need to invert $\boldsymbol{C}_B = \mathbf{1}$, an $m \times m$ matrix of ones. Using the first (rough) type of perturbation, we have $\boldsymbol{C}'_B = \mathbf{1} + \epsilon \boldsymbol{I}$, which is theoretically invertible for any $\epsilon > 0$. In practice, the feasibility and numerical stability of matrix inversion is related to its condition number, which is defined as the ratio of its greatest over its smallest eigenvalue. The matrix $\mathbf{1} + \epsilon \boldsymbol{I}$ has maximum eigenvalue $\lambda_{max} = m + \epsilon$ and minimum eigenvalue $\lambda_{min} = \epsilon$, with multiplicity $m - 1$. Thus, its condition number is approximately equal to m/ϵ. Most numerical computation packages can handle reasonably sized matrices with condition numbers up to 10^{15} so this can give us an approximate lower bound on ϵ, since we need $\epsilon \geq m10^{-15}$. As we mentioned before m is usually in the order of tens, so we have found that a good practical choice of ϵ is around 10^{-5}. This choice of ϵ also works for the smooth perturbation because the eigenvalues of the smooth and rough perturbation are relatively close, at least for h^* equal to the minimum distance between bases points. So far, we have only looked at the invertibility of each function's basis covariance matrix, but in fact we need the inverse $(\boldsymbol{C}_B + \boldsymbol{C}_U^\top \boldsymbol{X} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}^\top \boldsymbol{C}_U)^{-1}$ which also depends on the data. We can still justify our previous choice of ϵ by noticing that we can rewrite the above matrix as $\boldsymbol{C}_B^{-1} - \boldsymbol{C}_B^{-1} \boldsymbol{F} (\boldsymbol{I} + \boldsymbol{C}_B^{-1} \boldsymbol{F})^{-1} \boldsymbol{C}_B^{-1}$, with $\boldsymbol{F}$ defined as before. This expression involves only $\boldsymbol{C}_B^{-1}$ and the inversion of the matrix $(\boldsymbol{I} + \boldsymbol{C}_B^{-1} \boldsymbol{F})$ is generally stable.

3.4 Example

In this section we use the PP approximation and present results from applying it on the Canadian lynx data. We consider both the smooth and rough perurbations, and select the hyperparameters by maximizing the corresponding approximate marginal likelihood. The results for the exact method and the rough and smooth PP approximations using 10 equally spaced bases for each coefficient and $\epsilon = 10^{-5}$ are shown in Fig. 3.1, the corresponding optimal hyperparameters are given in Table 3.1. As we can see, both the hyperparameters and the posteriors are indistinguishable; the functional coefficients in this example are quite

smooth so, as we would expect, the approximation works very well. In order to find discrepancies between the fits we have to use fewer bases, so we try our method with 5 and 3 bases for each function and the results are shown in Fig. 3.2 and Fig. 3.3 respectively. Even with 5 bases, the exact and approximate estimation give very similar results and it requires reducing the number of bases down to three to see significant differences in the shape of the posterior functions. In the latter case we can clearly see how the first coefficient is a mixture of three Gaussian curves. In general, for relatively smooth one-dimensional coefficient functions we will not require more than 10 bases in our approximation to adequately capture their shape and this is why we believe our approximation will work well in practice. In terms of the comparison between the rough and smooth approximation, they both gave identical results in this example but we have found that in certain cases the smooth approximation behaves better, so it is the one we will adopt throughout.

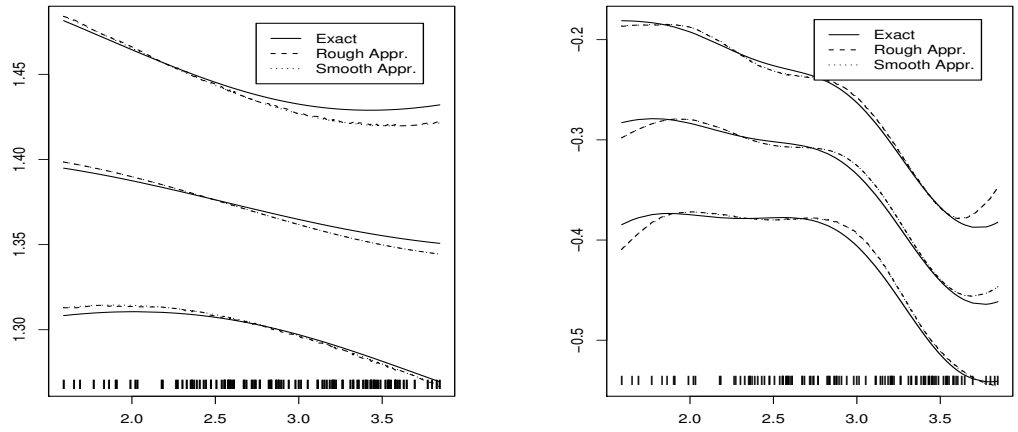


(a) (b)

Figure 3.1: Plot of posterior distributions of functional coefficients (a) f_1 and (b) f_2 for the Canadian lynx data using the exact method and the rough and smooth PP approximations with 10 bases.

Table 3.1: Selected hyperparameters by maximizing the marginal likelihood of the Canadian lynx data for the exact method and the rough and smooth PP approximations with 10 bases.

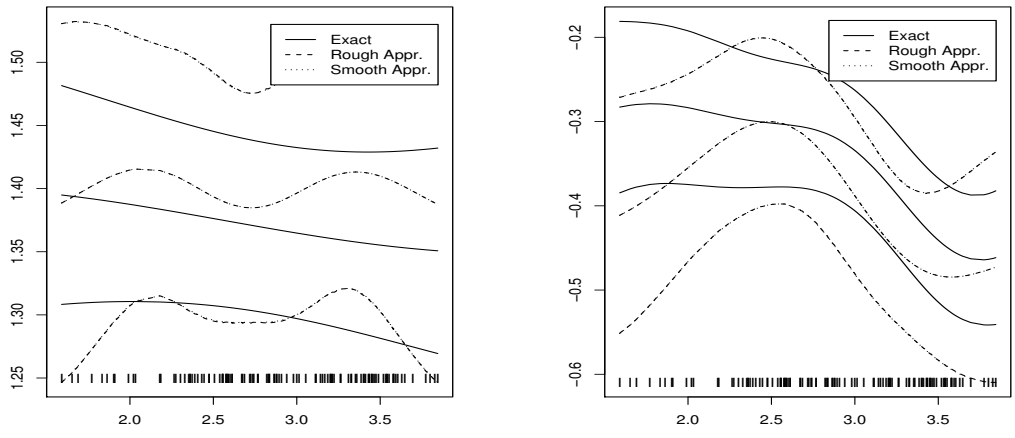
	Exact	Approximate	
		Rough	Smooth
σ	0.20917	0.20920	0.20920
μ_1	1.37474	1.37548	1.37549
μ_2	-0.34861	-0.34871	-0.34873
h_1	2.53528	2.53522	2.53490
h_2	0.73669	0.74166	0.74177
max ℓ	9.25885	9.26058	9.2606



(a) (b)

Figure 3.2: Plot of posterior distributions of functional coefficients (a) f_1 and (b) f_2 for the Canadian lynx data using the exact method and the rough and smooth PP approximations with 5 bases.

We also demonstrate the main difference between the PP and the SR (and by extension the Nystr om) approximation methods. We use both of them to fit the Canadian lynx data with the same 10 equally spaced basis points and the smooth covariance perturbation. As we have pointed out, the marginal likelihoods and posterior means are identical for the two methods, their only difference being their posterior covariance. In Fig. 3.4 we present a plot of the posterior distribution of the second functional coefficient under both methods. We



(a) (b)

Figure 3.3: Plot of posterior distributions of functional coefficients (a) f_1 and (b) f_2 for the Canadian lynx data using the exact method and the rough and smooth PP approximations with 3 bases.

have purposely extended the range of the argument beyond the observed range in order to capture their discrepancy. The variance of the SR method decays quickly to zero because we essentially represent the function as a (random) combination of Gaussian kernels, centered within the observed range. Any evaluation of these kernels outside the range will approach zero at an exponential rate, and this leads to the unfortunate interpretation that we are more sure about the value of the function outside the range where we observe data. On the other hand, the PP method preserves the local character of the estimation; the posterior variance of an evaluation outside the observed range approaches the prior variance. This is the reason why we chose the PP method, especially because when we want to simulate path from the process, taking estimation uncertainty into account, we might have to make evaluations outside the observed range of the functional coefficients.

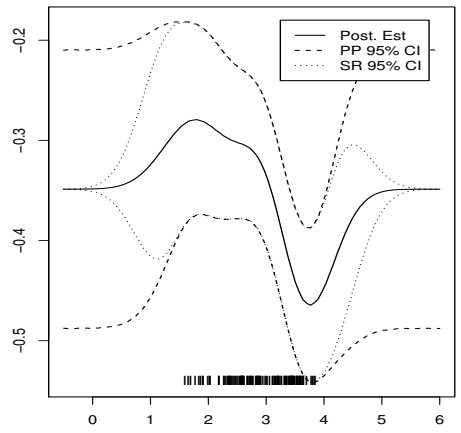


Figure 3.4: Plot of posterior distribution of functional coefficient f_2 for the Canadian lynx data under the PP and SR approximations with 10 bases.

3.5 Extensions

We present two extensions of our methodology which are based on reduced rank approximations. First, we look at multivariate FAR models, a natural nonlinear generalization of the vector autoregressive (VAR) model. Our treatment of multivariate nonlinear time series does not strictly necessitate approximate inference, but it is much more efficient if we do so. Second, we look at state space (SS) models and how we can incorporate nonlinear terms in the dynamics of the latent variables. This class of models is a generalization of the FAR model, but also encompasses other important sub-cases, such as factor models. Our treatment of SS models relies on reduced rank approximations.

3.5.1 Multivariate Models

Many practical applications of time series analysis extend beyond the univariate setting, with the aim of describing the interrelations across multiple series. In these cases, the VAR model has been the most successful, flexible and easy to use model; a thorough exposition

is given in Lütkepohl [78]. For a $d-$dimensional series $\{\boldsymbol{Y}_t\}$, the p^{th} order VAR dynamics are

$$\boldsymbol{Y}_t = \boldsymbol{\mu} + \boldsymbol{A}_1\boldsymbol{Y}_{t-1} + \ldots + \boldsymbol{A}_p\boldsymbol{Y}_{t-p} + \boldsymbol{\epsilon}_t$$

where $\boldsymbol{\mu}$ is a d-dimensional mean vector, $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_p$ are fixed $d \times d$ coefficient matrices, and $\{\boldsymbol{\epsilon}_t\}$ is a d-dimensional white noise sequence. It is straightforward to generalize the VAR model to a multivariate FAR model, by allowing the coefficients to be functions of some argument variable. The resulting specification is

$$\boldsymbol{Y}_t = \boldsymbol{X}_t^\top F(\boldsymbol{U}_t) + \boldsymbol{\epsilon}_t \tag{3.23}$$

where

$$\boldsymbol{X}_t^\top = \begin{bmatrix} X_t^{(1,1)} & \cdots & X_t^{(1,p_1)} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & X_t^{(2,1)} & \cdots & X_t^{(2,p_2)} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & X_t^{(d,1)} & \cdots & X_t^{(d,p_d)} \end{bmatrix}$$

$$F(\boldsymbol{U}_t)^\top = \begin{bmatrix} f_{11t} & \cdots & f_{1p_1t} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & f_{21t} & \cdots & f_{2p_2t} & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & \cdots & f_{d1t} & \cdots & f_{dp_dt} \end{bmatrix}$$

and $f_{ijt} = f_{i,j}(U_t^{(i,j)})$, with $\{X_t^{(i,j)}\}$, $\{U_t^{(i,j)}\}$ being $\mathcal{F}_{t-1}$ measurable variables, for any $i = 1, \ldots, d$ and $j = 1, \ldots, p_i$. The specification in (3.23) allows each component Y_{it} of $\boldsymbol{Y}_t = [Y_{1t}, \ldots, Y_{dt}]^\top$ to have p_i of its own regressor variables $\{X^{(i,1)}, \ldots, X^{(i,p_i)}\}$ multiplied

with p_i of its own functional coefficients, with arguments $\{U^{(i,1)}, \ldots, U^{(i,p_i)}\}$. As usual, the regressor and argument variables can depend on lagged values of any component of $\boldsymbol{Y}_t$, or they can be exogenous.

In terms of estimation, we use an independent GP prior for every function and work with the resulting conditional likelihood. We do not provide formulas, since they are readily extended from the univariate setting. However, there is a subtlety that arises from conditioning; if the error covariance matrix Σ_ϵ is diagonal we can equivalently fit a separate FAR model for each coordinate. In this case, it does not matter for estimation if some components of $\boldsymbol{Y}$ depend on the same regressor or argument variables. Of course, this argument does not hold for multi-step-ahead predictions or simulations from the model. If, on the other hand, Σ_ϵ is full, we must treat the system in a unified way and our likelihood-based method can account for the interactions between components. Note that the last is not true for nonparametric estimation based on conditional least squares where each coordinate is treated independently, unless the conditional errors are weighted by a fixed covariance matrix. In terms of computation, estimation in d dimensions increases the burden by roughly a factor of d^3. This is exactly true for the exact method; for the approximate method computation actually scales as $\mathcal{O}\Big(dT(\sum_{i=1}^{d}\sum_{j=1}^{p_i} m_i, j)^2\Big)$, where $m_{i,j}$ is the number of bases used for representing $f_{i,j}$. For reasons of efficiency, we suggest using approximate inference for multivariate models, and we provide such an example in section 6.2 where we look at a bivariate financial time series.

3.5.2 State Space Models

A SS model describes the evolution of an observable random sequence $\{\boldsymbol{Y}_t\}$ based on that of a latent random sequence $\{\boldsymbol{Z}_t\}$. The dynamics of a general linear state space (LSS)

model are

$$\begin{aligned} \boldsymbol{Y}_t &= \boldsymbol{G}_t + \boldsymbol{H}_t \boldsymbol{Z}_t + \boldsymbol{W}_t \\ \boldsymbol{Z}_{t+1} &= \boldsymbol{F}_t + \boldsymbol{E}_t \boldsymbol{Z}_t + \boldsymbol{V}_t \end{aligned}$$

where the $\boldsymbol{Y}_t$ and $\boldsymbol{Z}_t$ are d_Y- and d_Z-dimensional vectors, $\{\boldsymbol{W}_t\}$ and $\{\boldsymbol{V}_t\}$ are independent vector error sequences of the same dimensions, and $\boldsymbol{G}_t$, $\boldsymbol{H}_t$, $\boldsymbol{E}_t$, and $\boldsymbol{F}_t$ are nonstochastic, possibly time-varying matrices of dimensions $d_Y \times 1$, $d_Y \times d_Z$, $d_Z \times 1$ and $d_Z \times d_Z$, respectively. Gaussian LSS models, for which the error terms are normal, are by far the most popular SS models, the main reason being their generality and ease of use. In particular, they include ARMA and ARIMA models, factor models and dynamic regression. The statistical analysis of these models relies on the celebrated Kalman filter which, in turn, takes advantage of the conjugacy of the normal distribution to perform the analysis in a sequential manner. The LSS model appears under various names in the literature such as Structural Time Series or Dynamic Linear model. Good overviews of the model and the associated Kalman filtering procedures are given in Harvey [54], West and Harrison [121] and chapter 12 of Brockwell and Davis [14].

We also look at SS models for which the error terms are normal, but we relax the requirement of linearity in the dynamics. Our goal is to allow more flexibility without sacrificing the convenience of the conjugate calculations, and in order to achieve this we have to impose some restrictions on the model dynamics. Specifically, we look at nonlinear SS (NLSS) models of the form

$$\boldsymbol{Y}_t = \boldsymbol{G}(\boldsymbol{U}_{t-1})\boldsymbol{X}_{t-1} + \boldsymbol{H}_t \boldsymbol{Z}_t + \boldsymbol{W}_t \tag{3.24}$$

$$\boldsymbol{Z}_{t+1} = \boldsymbol{F}(\boldsymbol{U}_t)\boldsymbol{X}_t + \boldsymbol{E}_t \boldsymbol{Z}_t + \boldsymbol{V}_t \tag{3.25}$$

where $\boldsymbol{G}(\cdot)$, $\boldsymbol{F}(\cdot)$ are general matrix functions of arguments $\boldsymbol{U}_t$, multiplied by regressor vectors $\boldsymbol{X}_t$. The sequences of arguments and regressors $\{\boldsymbol{U}_t\}$ and $\{\boldsymbol{X}_t\}$ can be stochastic, but they have to be observed by time t. We estimate the functions in $\boldsymbol{G}$, $\boldsymbol{F}$ using reduced rank approximations of GPs, as in the FAR model. The basic restriction we impose on the dynamics of the NLSS model (3.24-3.25) is that the latent process $\{\boldsymbol{Z}_t\}$ can serve neither as an argument nor as a regressor in the nonlinear terms, in order to preserve the conditional normality of the process. At any time t, we want to describe the conditional distribution of the state variable $\boldsymbol{Z}_t$ by a normal distribution. Since function evaluations in our GP framework are also normal, we cannot multiply them with the state variable or use the state variable as an argument.

We describe how we address estimation from model (3.24-3.25). The main idea is to treat every function evaluation as another latent variable and do sequential analysis, which is essentially what we do when we perform online estimation for the FAR model. Compared to a SS model, the difference in the FAR model is that the state space dimension expands with the number of observations, and this prevents us from using Kalman filtering techniques which rely on a fixed dimension state variable with a Markovian structure. We can overcome this problem by using a reduced rank approximation and treating the random function evaluations as linear combinations of a finite basis with random coefficients, as in subsection 3.1.2. The random coefficients are treated as latent variables and are attached to the vector $\boldsymbol{Z}_t$. Since every function evaluation in our model comes from the same function draw, the coefficients are the same for every evaluation and therefore satisfy the Markovian assumption. Thus, we can recast the model in SS form using the reduced rank approximation. Before we treat model (3.24-3.25) in full generality, and in order to fix ideas, we look at two simple cases.

First, consider the simple NLSS model

$$\begin{aligned} y_t &= f(y_{t-1}) + z_t + \epsilon_t \\ z_{t+1} &= \alpha z_t + r_t \end{aligned}$$

where we want to estimate function f nonparametrically. We put a $\mathcal{GP}(\mu, C)$ prior on f and use an order m reduced rank approximation $f(\cdot) = \mu + \sum_{i=1}^{m} \beta_i C(\cdot, b_i) = \mu + C(\cdot, \boldsymbol{b})^\top \boldsymbol{\beta}$ with basis points $\boldsymbol{b} = [b_1, \ldots, b_m]^\top$ and coefficient vector $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_m]^\top$. A priori, $\boldsymbol{\beta}$ follows a normal distribution with zero mean and covariance matrix $\boldsymbol{C}_b^{-1}$, where $\boldsymbol{C}_b = \left[\{C(b_i, b_j)\}_{i,j=1}^m\right]$. We expand the state variable z_t into the $(m+1)-$dimensional state vector $\boldsymbol{z}'_t = [z_t, \beta_1, \ldots, \beta_m]^\top$. The dynamics of the approximate SS model become

$$\begin{aligned} y_t &= \mu + [1, C(y_{t-1}, b_1), \ldots, C(y_{t-1}, b_m)]\boldsymbol{z}'_t + \epsilon_t \\ \boldsymbol{z}'_{t+1} &= \begin{bmatrix} \alpha & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \boldsymbol{z}'_t + \begin{bmatrix} r_t \\ 0 \\ \vdots \\ 0 \end{bmatrix} \end{aligned}$$

Now, suppose function f appears in the the latent process $\{z_t\}$

$$\begin{aligned} y_t &= z_t + \epsilon_t \\ z_{t+1} &= f(y_t) + \alpha z_t + r_t \end{aligned}$$

Although it seems that we just shifted the observed mean level from one equation to another, the two models are very different. In the first case the effect of the function is only instantaneously expressed, whereas in the second case it carries over to subsequent observations, because the state process has autoregressive dynamics. We can approximate this

model, using the same idea, by

$$y_t = [1, 0, \ldots, 0]\boldsymbol{z}'_t + \epsilon_t \tag{3.26}$$

$$\boldsymbol{z}'_{t+1} = \begin{bmatrix} \mu \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} \alpha & C(y_t, b_1) & \ldots & C(y_t, b_m) \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \boldsymbol{z}'_t + \begin{bmatrix} r_t \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{3.27}$$

We now extend this approach to the general NLSS model (3.24-3.25). For every function in the matrices $\boldsymbol{G}$ and $\boldsymbol{F}$ we use a reduced rank approximation to express it as a linear combination of a collection of $\boldsymbol{U}_t$-dependent basis functions times some $\boldsymbol{X}_t$-dependent arguments. The function uncertainty in this representation is described by the distribution of the random basis coefficients, which we treat as latent variables. We expand the latent process $\boldsymbol{Z}_t$ by these coefficients in order to form a new state vector $\boldsymbol{Z}'_t$. We also expand the matrices $\boldsymbol{H}_t$ and $\boldsymbol{E}_t$ to accommodate the basis terms, getting the new matrices $\boldsymbol{H}'(\boldsymbol{U}_t, \boldsymbol{X}_t)$ and $\boldsymbol{E}'(\boldsymbol{U}_t, \boldsymbol{X}_t)$. The resulting approximate representation of the NLSS model (3.24-3.25) is

$$\boldsymbol{Y}_t = \boldsymbol{M}_G \boldsymbol{X}_{t-1} + \boldsymbol{H}'(\boldsymbol{U}_{t-1}, \boldsymbol{X}_{t-1})\boldsymbol{Z}'_t + \boldsymbol{W}_t \tag{3.28}$$

$$\boldsymbol{Z}_{t+1} = \boldsymbol{M}_F \boldsymbol{X}_t + \boldsymbol{E}'(\boldsymbol{U}_t, \boldsymbol{X}_t)\boldsymbol{Z}'_t + \boldsymbol{V}'_t \tag{3.29}$$

where $\boldsymbol{M}_G \boldsymbol{X}_{t-1}$ and $\boldsymbol{M}_F \boldsymbol{X}_t$ are terms accounting for the prior mean level of the functions $\boldsymbol{G}$ and $\boldsymbol{F}$ respectively, and $\boldsymbol{V}'_t$ is the appropriately expanded latent error vector. We impose the prior on the functions through the initial distribution of the state variable.

Our approximate NLSS model (3.28-3.29) falls into the category of conditionally Gaussian

random processes and lends itself to filtering techniques. Lipster and Shiryaev [76] give an extensive treatment of conditional Gaussian filtering. On the theoretical side, the authors show that there exists a solution to (3.28-3.29) whenever the elements of $\boldsymbol{M}_G\boldsymbol{X}_{t-1}$ and $\boldsymbol{M}_F\boldsymbol{X}_t$ have finite second moments and the elements of $\boldsymbol{H}'$ and $\boldsymbol{E}'$ are almost surely bounded, plus some additional conditions on the initial state distribution (see p.76 of Lipster and Shiryaev [76]). In our setting, these conditions will always be satisfied if the nonlinear functions are additive (the regressors $\boldsymbol{X}_t$ are unity) and we use a squared exponential kernel, because the basis functions will be bounded. For more general models the conditions do not hold by default, but we can still apply the same statistical procedure to obtain the Kalman recursions, which only rely on the conditional normality assumption.

Next, we put our approach in context with respect to competing methods. The idea of linearizing the functions in a NLSS model is not new and dates back at least to Ghahramani and Roweis [42], who use Gaussian kernel bases for representing the nonlinear functions. The authors use the extended Kalman filter for approximating the filtering equations and an EM algorithm for choosing the coefficients in the representation, which are free parameters and are not regularized. Wang, Fleet and Hertzmann [119] explicitly put Gaussian priors for the nonlinear functions, but they treat the latent variables $\{Z_t\}$ as parameters over which the likelihood is maximized. This is different from filtering methods, for which the likelihood is marginalized over the latent variables. Both of these references are focused toward general models of the form

$$\begin{aligned} \boldsymbol{Y}_t &= \boldsymbol{G}(\boldsymbol{Z}_t) + \boldsymbol{W}_t \\ \boldsymbol{Z}_{t+1} &= \boldsymbol{F}(\boldsymbol{Z}_t) + \boldsymbol{V}_t \end{aligned}$$

where the state variable is nonlinear in both the state and the observation equation. In contrast, we require the state variable to appear only linearly in the model, and permit

nonlinear terms only for observed variables. Our contribution lies in identifying the appropriate NLSS model form and using reduced rank approximations in such a way that nonparametric estimation can be tackled with essentially linear methods. Even though the dynamics of the model are restricted, we can perform inference very simply and efficiently. In particular, this method allows us to treat FAR models with a linear moving average term.

We describe the implementation of our NLSS estimation method through a simulated example. We generate 1000 observations from the following model

$$y_t = z_t + \epsilon_t \tag{3.30}$$

$$z_{t+1} = f(y_t) + \alpha z_t + r_t \tag{3.31}$$

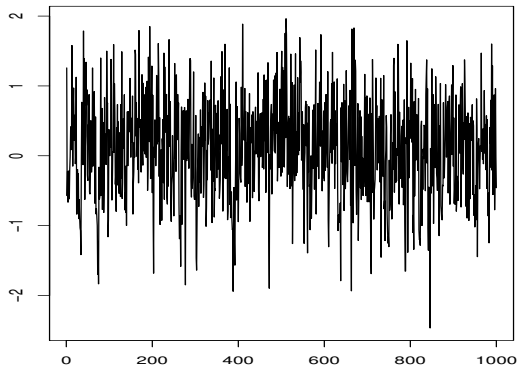
where $\epsilon_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$, $r_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \omega^2)$, $\sigma = .5$, $\omega = .2$, $\alpha = .5$, and f is a sinusoidal function given by $f(x) = \cos(3x)/2$. The initial value of the state variable z_1 is drawn from $\mathcal{N}(0, \omega^2)$, and the plot of the generated series is given in Fig. 3.5. Our goal is to estimate the nonlinear function nonparametrically, so we put a $\mathcal{GP}(\mu, C)$ prior on f, where $C(x, x') = \nu^2 \exp\{(x - x')^2/h^2\}$ is a squared exponential kernel. We represent the function in reduced form using 10 equally spaced basis points $\{b_i\}_{i=1}^{10}$ in the observed range of y_t, as $f(\cdot) = \mu + \sum_{i=1}^{m} \beta_i C(\cdot, b_i)$. The state vector becomes $\boldsymbol{z}_t = [z_t, \beta_1, \ldots, \beta_{10}]^\top$ and the approximate NLSS model is the same as in (3.26-3.27). We do not assume knowledge of the model parameters (σ,ω,α) or the kernel hyperparameters (μ,ν,h), but we choose them by maximizing the marginal likelihood of the model, calculated using the Kalman filter. For this, the initial distribution of the state variable must be be normal. Assuming the initial distribution of the original state z_1 is $\mathcal{N}(m_{z_1}, s_{z_1}^2)$, the initial distribution of the extended

state z_1' becomes

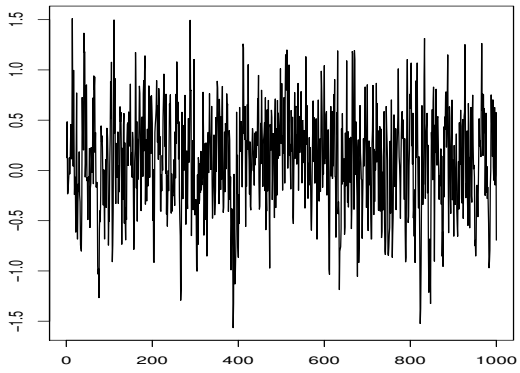
$$\mathcal{N}\left(\begin{bmatrix} m_{z_1} \\ \mathbf{0}_{10\times 1} \end{bmatrix}, \begin{bmatrix} s_{z_1}^2 & \mathbf{0}_{1\times 10} \\ \mathbf{0}_{10\times 1} & \boldsymbol{C}_b^{-1} \end{bmatrix}\right)$$

where $\boldsymbol{C}_b$ is the covariance matrix of the basis points (to avoid singularities in $\boldsymbol{C}_b^{-1}$, we use a smooth perturbation of the covariance kernel). From this point, it is straightforward to apply Kalman recursions for filtering, smoothing and likelihood calculation in the expanded model, e.g. see Harvey [54].

We give more details on the practical choices we make for fitting our model to the simulated data. First, we treat the prior uncertainty ν as a free parameter, and do not fix its value relative to σ as in the FAR model. We do this because we have not found a satisfactory method for describing ν in relation to the other parameters. Our heuristic argument regarding Fisher information in FAR models does not carry over to NLSS models, since, in the later case, the information in the observations is divided between estimation of the latent states $\{z_t\}$ and estimation of the function f. If we have more than one functional coefficient, we suggest making them share the same prior uncertainty within each equation (state or observation), and treat these uncertainty levels as free parameters. Moreover, we set $m_{z_1}, s_{z_1}^2$ equal to the sample moments of $\{y_t\}$, since y_t is equal to z_t with added noise. In more general settings, we propose using a reasonably diffuse normal distribution, since the effect of the initial distribution dies off with more data in well behaved models. We apply the Kalman filter for calculating the marginal likelihood, which we maximize over $(\sigma, \omega, \alpha, \mu, \nu, h)$ using a gradient descent algorithm. For simplicity, we use a numerical approximation to the gradient, but for models with many parameters it might be more convenient to calculate it explicitly. The gradient can be obtained by straightforward, although tedious, differentiation of the Kalman recursions. Shumway and Stoffer [106] proposed an



(a) y_t



(b) z_t

Figure 3.5: Plot of simulated data from model (3.30-3.31), (a) observations, (b) state variables.

alternative way of selecting parameters in the LSS model which is based on the expectation maximization (EM) algorithm and is potentially more stable, but we have not experimented with it yet. Finally, we describe our suggestion for the initial values of the numerical gradient descent algorithm. First, we fit a LSS model to the data by substituting the function f with a constant ϕ and optimizing its marginal likelihood over the parameters. Similar to the FAR setting, we relate the selected parameters in the LSS model with the initial values

of the NLSS model. In particular, the parameters that appear in both models are equated, and for the kernel hyperparameters we set $\mu_{init} = \hat{\phi}$, $\nu_{init} = \sqrt{\hat{\sigma}^2 + \hat{\omega}^2}$ and $h_{init} = S_y$, where S_y is the sample standard deviation. Turning back to our simulated data, the selected parameters are given in Table 3.2 and the posterior distribution of the function is shown in Fig. 3.6. Both the model parameters and the estimated function are close to their true values. In Fig. 3.7 we also plot the true state variables $\{z_t\}$ versus the output of the Kalman smoother (i.e. $\mathrm{E}[z_t|y_1, \ldots, y_T]$, for $t = 1, \ldots, T$), which lie close to the 45-degree line as we would expect. Our proposed method seems to work well for the simulated data set, and can be useful for efficient nonparametric estimation in the class of NLSS models we consider. We give a real data example in section 6.3, where we apply our NLSS methodology to a stochastic volatility model.

Table 3.2: The parameters that maximize the marginal likelihood of the simulated data from model (3.30-3.31).

model		f	
α	0.55169	μ	0.16590
σ	0.52455	ν	0.50226
ω	0.26841	h	0.82661
ℓ = -934.2438			

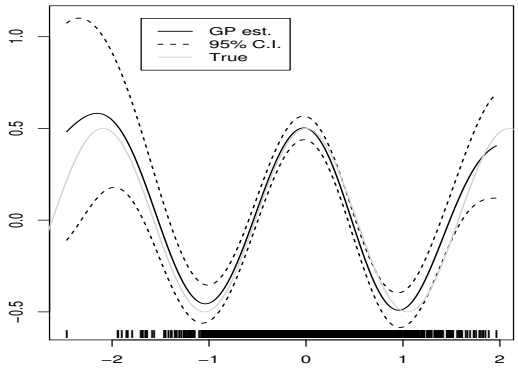


Figure 3.6: Plot of estimated and true functions from model (3.30-3.31).

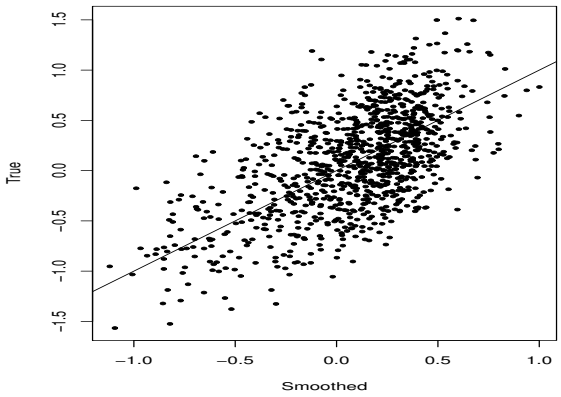


Figure 3.7: Plot of true versus Kalman smoothed state values from model (3.30-3.31).

Chapter 4

Theoretical Properties

In this chapter we explore some of the theoretical aspects of our proposed nonparametric estimation technique. Our main goal is to prove the consistency of our estimators from a frequentist perspective and, through that, get an understanding of the conditions under which our method is expected to perform best. We begin with an overview of relevant theoretical results for nonparametric estimation, and we then discuss the characteristics of our model which help in the development of its properties, in particular its correspondence to penalized regression in reproducing kernel Hilbert spaces. The basic result is Theorem 4.3 which proves the consistency of our functional coefficient estimates in such spaces under sufficient identifiability and ergodicity conditions, covering both the Markov and the more general time series regression setting. Finally, we describe the theoretical behavior of estimation with reduced rank approximations, because this is what we realistically use for large data sets and is, therefore, more fitting for asymptotic considerations.

4.1 Review of Nonparametric Estimation Theory

Many theoretical results from nonparametric estimation with independent data have been extended to a time series context. The monograph of Bosq [9] contains a detailed review of

kernel regression and kernel density estimation for dependent sequences of data. Specifically, he provides consistency, asymptotic normality and convergence rates for kernel estimators of the NLAR model under mixing conditions. Masry and Fan [80] give similar results for local polynomial regression, under the same type of conditions. We are more interested, however, in the FAR model, for which several results have appeared depending on the estimation procedure used. For the ALR method, Chen [20] proves mean square consistency of the functional coefficients evaluated at the observed data points, and Chen and Liu [21] show that the functional coefficient estimates, evaluated at a fixed point, converge to a normal distribution. For the LLR method, Cai, Fan and Yao [16] also look at fixed point evaluations, and give consistency and asymptotic normality of the coefficient estimates. For both ALR and LLR, the authors assume the coefficients are twice differentiable and provide standard nonparametric $T^{2/5}$ convergence rates. Finally, Huang and Shen [60] prove L_2 consistency of their spline regression estimated functions over a compact range. All the results for the FAR model require special identifiability and ergodicity conditions, similar to the ones we adopt later.

Turning to GP regression, the question of consistency can be interpreted in a Bayesian manner, by looking at the concentration of the posterior distribution around the true function. This approach has been adopted by Shen and Wasserman [105], and recently Ghosal and van der Vaart [43] extended it to non i.i.d. data, including a first order NLAR model. However, their results rely on technical entropy conditions that are difficult to verify. Choi [24] gives more intuitive conditions in terms of the GP prior specification, but his setting does not cover time series. We do not consider posterior consistency, but rather focus on the consistency of point estimates, i.e. the posterior means of the functional coefficients. Lin and Brown [73] look at GP regression from this perspective, using a periodic variation of the Gaussian prior kernel, and demonstrate that the method performs well, in a minimax

sense, when the true function is very smooth. Our treatment of the FAR model aims at consistency, but we do not provide convergence rates or adaptive results for smoothness classes of function, because our setting is more involved and the results also depend on the dynamic behavior of the data, not just the functional coefficients' smoothness.

4.2 GP regression and Reproducing Kernel Hilbert Spaces

We now establish a connection between GP estimation for the FAR model and reproducing kernel Hilbert spaces (RKHSs). We give a brief introduction to RKHSs, together with the relevant results for our purposes, in section A in the Appendix. The main point of this section is Lemma 4.2, which gives a convenient characterization of the posterior mean functions. Suppose we have the model

$$Y_t = f_1(U_t^{(1)})X_t^{(1)} + \ldots + f_p(U_t^{(p)})X_t^{(p)} + \epsilon_t \tag{4.1}$$

$$f_i \sim \mathcal{GP}(\mu_i, C_i), \quad i = 1, \ldots, p, \tag{4.2}$$

and each function's covariance kernel $C_i(\cdot,\cdot)$ defines its RKHS $\mathcal{K}_i$ respectively, see Theorem A.4 in the Appendix. For simplicity, we assume that all prior mean functions are constant, equal to zero, $\mu_i(\cdot) = 0$; from (2.10) the posterior mean of the vector of observed functional coefficients is $E[\boldsymbol{f}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}] = \boldsymbol{C}\boldsymbol{\alpha}$, where

$$\boldsymbol{\alpha} = \boldsymbol{X}(\boldsymbol{\Sigma} + \boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{X})^{-1}\boldsymbol{y} \tag{4.3}$$

Moreover, the posterior means of the functions $\{f_i\}$ evaluated at an arbitrary point are

$$\mathrm{E}[f_i(\cdot)|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}] = \sum_t \alpha_{i,t} C_i(\cdot, u_t^{(i)}); \quad i = 1, \ldots, p$$

where the coefficients $\alpha_{i,t}$ are such that $\boldsymbol{\alpha}_i^\top = [\ldots, \alpha_{i,t}, \ldots]$ and $\boldsymbol{\alpha}^\top = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p]^\top$. The posterior mean functions are linear combinations of their covariance kernels centered at the observed arguments, thus each $\mathrm{E}[f_i(\cdot)|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}]$ belongs to its corresponding RKHS $\mathcal{K}_i$. There is an equivalent way of obtaining these estimates through a penalized least squares problem in the RKHS. First, we need the following proposition which is an extension of the Representer Theorem of Kimeldorf and Wahba [68] to varying coefficient regression. Schölkopf, Herbrich and Smola [99] have generalized the Representer theorem to a larger class of regularizers and empirical risk functions, and our proof is adapted from theirs.

Proposition 4.1 (Representer Theorem for Varying Coefficient Models)**.** *Assume spaces $\mathcal{X}_i$ endowed with positive definite kernels $k_i : \mathcal{X}_i \times \mathcal{X}_i \to \mathbb{R}$ and their corresponding RKHSs $\mathcal{K}_i$, for $i = 1, \ldots, p$, and a sample $\left\{y_t, \{u_{i,t}\}_{i=1}^p, \{x_{i,t}\}_{i=1}^p\right\}_{t=1}^T$, where $y_t, x_{i,t} \in \mathbb{R}$ and $u_{i,t} \in \mathcal{X}_i$. Then any functions $\{f_i^\star \in \mathcal{K}_i\}_{i=1}^p$ that minimize*

$$\sum_t \left(y_t - f_1(u_{1,t})x_{1,t} - \ldots - f_p(u_{p,t})x_{p,t}\right)^2 + \sum_i \|f_i\|_{\mathcal{K}_i}^2 \tag{4.4}$$

admit a representation of the form

$$f_i^\star(\cdot) = \sum_t \alpha_{i,t} k_i(\cdot, u_{i,t}); \quad i = 1, \ldots, p \tag{4.5}$$

Proof: Consider linear subspaces $\mathcal{K}_{D,i} \subset \mathcal{K}_i$ spanned by the functions $\{k_i(\cdot, u_{i,t})\}_{t=1}^T$ and their orthogonal complements $\mathcal{K}_{D,i}^\perp \subset \mathcal{K}_i$, for $i = 1, \ldots, p$. Every $f_i \in \mathcal{K}_i$ has a unique decomposition $f_i(\cdot) = f_i^\parallel(\cdot) + f_i^\perp(\cdot) = \sum_t \alpha_{i,t} k_i(\cdot, u_{i,t}) + f_i^\perp(\cdot)$, where $f_i^\parallel \in \mathcal{K}_{D,i}$ and $f_i^\perp \in \mathcal{K}_{D,i}^\perp$. Using the reproducing property, for any $u_{i,s}$ in the sample we have

$$f_i(u_{i,s}) \quad = \quad \langle f_i(\cdot), k_i(\cdot, u_{i,s})\rangle_{\mathcal{K}_i}$$

$$
\begin{aligned}
&= \sum_t \alpha_{i,t} \langle k_i(\cdot, u_{i,t}), k_i(\cdot, u_{i,s}) \rangle_{\mathcal{K}_i} + \langle f_i^{\perp}(\cdot), k_i(\cdot, u_{i,s}) \rangle_{\mathcal{K}_i} \\
&= \sum_t \alpha_{i,t} k_i(u_{i,t}, u_{i,s})
\end{aligned}
$$

where the second term vanishes due to orthogonality. So, the values of any f_i evaluated at the data points only depend on $f_i^{\|}$ and not on $f_i^{\perp}$. Assume a solution $\{f_i^{\star}\}_{i=1}^p$ with decomposition $f_i^{\star\|} \in \mathcal{K}_{D,i}$ and $f_i^{\star\perp} \in \mathcal{K}_{D,i}^{\perp}$, for $i = 1, \ldots, p$. The first term in the objective function (4.4) is independent of the orthogonal complements and the second term is the sum of the function norms

$$
\sum_i \|f_i^{\star}\|_{\mathcal{K}_i}^2 = \sum_i \|f_i^{\star\|}\|_{\mathcal{K}_i}^2 + \|f_i^{\star\perp}\|_{\mathcal{K}_i}^2 \leq \sum_i \|f_i^{\star\|}\|_{\mathcal{K}_i}^2
$$

This implies that for each $f_i^{\star}$ we must have $\|f_i^{\star\perp}\|_{\mathcal{K}_i}^2 = 0$, thus $f_i^{\star} \in \mathcal{K}_{D,i}$ and the representation (4.5) is valid. □

Proposition 4.1 is easily extended to more general objective functions. Assume the following minimization problem

$$
\min_{f_i \in \mathcal{K}_i,\, i=1,\ldots,p} \left\{ \sum_t C\left(y_t, \{x_{i,t}, f_i(u_{i,t})\}_{i=1}^p\right) + \Omega\left(\{\|f_i\|_{\mathcal{K}_i}\}_{i=1}^p\right) \right\} \tag{4.6}
$$

where the loss function C is point-wise, i.e. it depends on the f_i's only through their values at the data points, and the penalty function $\Omega : \mathbb{R}_+^p \to \mathbb{R}$ that is strictly monotonically increasing. Then the same representation (4.5) applies to the solution of (4.6).

We have already seen that the posterior mean functions from our model belong to their corresponding RKHSs. The next step is to identify these functions as the solutions to a penalized least squares problem.

Lemma 4.2. *Let* $\left\{y_t, \{u_{i,t}\}_{i=1}^p, \{x_{i,t}\}_{i=1}^p\right\}_{t=1}^T$ *be a sample from (4.1). The posterior means*

of the functional coefficients from our GP estimation procedure are given as solutions to the following minimization problem

$$\min_{f_i \in \mathcal{K}_i,\, i=1,\ldots,p} \left\{ \frac{1}{\sigma^2} \sum_t \left(y_t - f_1(u_t^{(1)}) x_t^{(1)} - \ldots - f_p(u_t^{(p)}) x_t^{(p)} \right)^2 + \sum_i \|f_i\|_{\mathcal{K}_i}^2 \right\} \tag{4.7}$$

Proof: From Proposition 4.1, we can represent the functions evaluated at the observed arguments as $\boldsymbol{f} = \boldsymbol{C}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is the stacked vector of the representation coefficients. Substituting back into (4.7), we get the minimization problem

$$\min_{\boldsymbol{\alpha}} \left\{ (\boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{\alpha})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{C}\boldsymbol{\alpha}) + \boldsymbol{\alpha}^\top \boldsymbol{C}\boldsymbol{\alpha} \right\}$$

Differentiating w.r.t. $\boldsymbol{\alpha}$ and setting the gradient equal to zero, the solution for $\boldsymbol{\alpha}$ is equal to (4.3). Thus, the posterior mean functions from our procedure are exactly the solutions to problem (4.7). □

In our exposition, we fixed the prior mean functions to zero, but this can be relaxed. In particular, assuming that the prior mean for each f_i is given by an arbitrary μ_i, it is obvious from (2.10) that the posterior mean of each function is a linear combination of μ_i and an element of $\mathcal{K}_i$. The analogous semiparametric representer theorem involves a minimization over the spaces of functions $\mathcal{K}_i + \text{span}\{\mu_i\} = \{f_i : f_i = g_i + \beta\mu_i; \beta \in \mathbb{R}, g_i \in \mathcal{K}_i\}$, see e.g. Theorem 2 in [99].

4.3 Consistency

We look at the consistency of ours estimator from a frequentist perspective. We assume our data come from the true underlying model

$$Y_t = f_1(U_t^{(1)}) X_t^{(1)} + \ldots + f_p(U_t^{(p)}) X_t^{(p)} + \epsilon_t \tag{4.8}$$

with fixed but unknown functional coefficients f_i. The regressor and arguments variables $\{X^{(i)}\}_{i=1}^p$ and $\{U^{(i)}\}_{i=1}^p$, are allowed to be endogenous (lagged values of Y_t) or exogenous, in any combination. We estimate the functional coefficients based on data $\left\{y_t, \{x_t^{(i)}, u_t^{(i)}\}_{i=1}^p\right\}_{t=1}^T$, using an independent GP prior $\mathcal{GP}(0, C_i)$ for each f_i. We assume the prior mean is zero and the prior covariance C_i is fixed, and defines a RKHS $\mathcal{K}_i$, for each function. Before we present our consistency result for the posterior mean functions of our estimation procedure, we provide the required conditions. The first set of conditions C.1 must hold in any case, but we impose two alternative conditions to ensure ergodicity, depending on whether model (4.8) is Markovian or not.

Condition C.1. *Let* $\boldsymbol{X}_t^\top = [X_t^{(1)}, \ldots, X_t^{(p)}]$ *and* $\boldsymbol{U}_t^\top = [U_t^{(1)}, \ldots, U_t^{(p)}]$.

i. The process $\{\boldsymbol{X}_t, \boldsymbol{U}_t\}$ *is jointly strictly stationary, with stationary measure* π.

ii. There exists a compact set $C \in \mathbb{R}^p$ *with positive Lebesgue measure, such that the eigenvalues of* $\mathrm{E}_\pi[\boldsymbol{X}_t \boldsymbol{X}_t^\top | \boldsymbol{U}_t = \boldsymbol{u}_t]$ *are uniformly bounded away from zero and infinity for all* $\boldsymbol{u}_t \in C$, *and the density of* $\boldsymbol{U}_t$ *over* C *is bounded away from zero.*

iii. The functions $\{f_i\}_{i=1}^p$ *belong to the corresponding RKHSs* $\{\mathcal{K}_i\}_{i=1}^p$ *and are bounded.*

iv. The error terms $\{\epsilon_t\}$ *are an independent white noise sequence.*

We do not discuss conditions C.1.i and C.1.iv, since they are standardly assumed for time series. Condition C.1.iii requires the functional coefficients to belong to the RKHSs of the prior covariance functions, implying that we have knowledge the appropriate C_i's. For the Gaussian kernels that we typically use, the RKHSs are considerably smooth, containing infinitely differentiable functions. However, our consistency result also holds for more general covariance kernels. More important is Condition C.1.ii, which serves as an identifiability condition for the functions. The compact set C restricts the range over which consistency holds, in order to avoid problematic behavior at infinity. The condition can fail depending

on the specification of the model; for instance, when $U_t^{(i)} = X_t^{(i)}$ for all i, then the condition fails for any set C because this FAR model is equivalent to an additive model. However, the condition is not void; consider for example the following Markovian FAR model

$$X_t = f_1(X_{t-2})X_{t-1} + f_2(X_{t-2})X_{t-2} + \epsilon_t \tag{4.9}$$

with stationary measure π. Letting $\boldsymbol{X}_t^\top = [X_{t-1}, X_{t-2}]$, $\boldsymbol{U}_t = X_{t-2}$ and $\bar{X}_{t-1} = X_{t-1} - \epsilon_{t-1} = X_{t-2}f_1(X_{t-3}) + X_{t-3}f_2(X_{t-3})$, we have

$$\begin{aligned}
\mathrm{E}_\pi[\boldsymbol{X}_t\boldsymbol{X}_t^\top|X_{t-2}=x] &= \mathrm{E}_\pi\left\{\begin{bmatrix} X_{t-1}^2 & X_{t-1}X_{t-2} \\ X_{t-1}X_{t-2} & X_{t-2}^2 \end{bmatrix}|X_{t-2}=x\right\} \\
&= \mathrm{E}_\pi\left\{\begin{bmatrix} \bar{X}_{t-1}^2 & \bar{X}_{t-1}X_{t-2} \\ \bar{X}_{t-1}X_{t-2} & X_{t-2}^2 \end{bmatrix} + \begin{bmatrix} \bar{X}_{t-1}\epsilon_{t-1}+\epsilon_{t-1}^2 & \epsilon_{t-1}X_{t-2} \\ \epsilon_{t-1}X_{t-2} & 0 \end{bmatrix}|X_{t-2}=x\right\} \\
&= \underbrace{\begin{bmatrix} a_{11}(x) & a_{12}(x) \\ a_{21}(x) & x^2 \end{bmatrix}}_{\boldsymbol{A}(x)} + \begin{bmatrix} \sigma^2 & 0 \\ 0 & 0 \end{bmatrix}
\end{aligned}$$

where $a_{11}(x) = \mathrm{E}_\pi[\bar{X}_{t-1}^2|X_{t-2}=x]$, $a_{12}(x) = a_{21}(x) = \mathrm{E}_\pi[\bar{X}_{t-1}X_{t-2}|X_{t-2}=x]$. We assume that functional coefficients are continuous and bounded, and that the stationary measure is well behaved in the sense that it is absolutely continuous with respect to Lebesgue measure and expectations in $\boldsymbol{A}(x)$ are finite. An obvious, although uninteresting, example is the linear Gaussian AR model, but it is easy to imagine small, nonlinear variations thereof which retain these properties. The matrix $\boldsymbol{A}(x)$ is positive semi-definite by construction for any x. For some $x \neq 0$, and after the addition of σ^2 in its first element, both eigenvalues of the matrix are positive. Since these eigenvalues are continuous functions of x, we can find a compact set C, excluding zero, which satisfies Condition C.1.ii on the eigenvalues and the stationary measure. In general, this condition is quite strong and difficult to verify, but the

same identifiability requirements are always used for the theoretical development of FAR models.

In the case where all $\{X^{(i)}\}_{i=1}^{p}$ and $\{U^{(i)}\}_{i=1}^{p}$ variables are lagged values of Y_t, we require in addition the following conditions C.2.

Condition C.2.

i. The variables in $\boldsymbol{X}_t, \boldsymbol{U}_t$ depend only on a finite number of lagged values of Y_t.

ii. The Markov chain $\{\boldsymbol{X}_t, \boldsymbol{U}_t\}$ is positive Harris recurrent.

In the case where there is at least one exogenous variable serving as either a regressor or an argument, we alternatively require the more general mixing conditions C.3.

Condition C.3.

i. For some $r > 2$, the elements of $\boldsymbol{X}_t$ satisfy $\mathrm{E}_\pi \left[|X_t^{(i)} X_t^{(j)}|^r \right] \leq \infty$, for all $i, j = 1, \ldots, p$

ii. The process $\{\boldsymbol{X}_t, \boldsymbol{U}_t\}$ is α-mixing with coefficients $\alpha(t)$ such that $\sum_{t \geq 1} \alpha(t)^{\frac{r-2}{r}} < \infty$ (e.g. they decay exponentially as $\alpha(t) \leq Ct^{-a}$, at a rate $a > r/(r+2)$)

The importance of the last two conditions is illustrated in the following section. We are now in a position to state our main result.

Theorem 4.3. *Assume Condition C.1 and either Condition C.2 or Condition C.3 hold, and let $\{\hat{f}_i\}$ be the posterior means of the functional coefficients from our GP method, applied over the set C. Then*

$$\|\hat{f}_i - f_i\|^2_{\mathcal{K}_i(C)} \xrightarrow{P} 0; \;\; T \to \infty, \; \forall i = 1, \ldots, p$$

where $\mathcal{K}_i(C)$ is the restriction of $\mathcal{K}_i$ over C.

The coefficient estimates over C converge to the true functions in the RKHS metric, which also implies pointwise convergence by the properties of RKHSs (see Appendix).

4.3.1 Ergodicity

For proving Theorem 4.3 we first need to establish the ergodicity of the process in the sense that

$$\frac{1}{T}\sum_{t=1}^{T} f_i(u_t^{(i)})x_t^{(i)} f_j(u_t^{(j)})x_t^{(j)}\,\mathrm{I}(\boldsymbol{u}_t \in C) \xrightarrow{P} \mathrm{E}_\pi\left[f_i(U_t^{(i)})X_t^{(i)} f_j(U_t^{(j)})X_t^{(j)}\,\mathrm{I}(\boldsymbol{U}_t \in C)\right] \quad (4.10)$$

as $T \to \infty$, for bounded functions $\{f_i\}$ and any $i, j = 1, \ldots, p$. In the remainder we sometimes suppress the dependence on the compact set C to simplify notation, but it is implicitly assumed that the result holds over this set. We show that Condition C.1 together with either Condition C.2 or C.3 can guarantee (4.10) under the Markovian or mixing setting, respectively.

Markov Conditions

If the variables $\{X^{(i)}\}_{i=1}^p$ and $\{U^{(i)}\}_{i=1}^p$ in (4.8) consist only of lagged values of Y_t, then the process $\{\boldsymbol{Y}_t = [Y_t, \ldots, Y_{t-q+1}]^\top\}$ is a Markov chain, where q is the maximum between the autoregressive order and the maximum lag used in defining the arguments. Markov processes have been studied extensively under various frameworks, and there is a plethora of results regarding different types of ergodicity. The following theorem of Meyn and Tweedie [84], although quite general, is sufficient for our purposes.

Theorem 4.4. *(Meyn and Tweedie [84], Thm 17.1.7, p. 421) Let $\{X_t\}$, taking values in $\mathcal{X}$, be a positive Harris recurrent Markov chain with stationary measure π. Then, for any*

$f \in L_1\big(\mathcal{X}, \mathcal{B}(\mathcal{X}), \pi\big)$, we have almost surely that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} f(X_t) = \mathrm{E}_\pi[f(X_1)] \tag{4.11}$$

The next proposition establishes the desired result.

Proposition 4.5. *Under Conditions C.1 and C.2, equation (4.10) holds.*

Proof: Let $F^{(i,j)}(\boldsymbol{X}_t, \boldsymbol{U}_t) = f_i(U_t^{(i)}) X_t^{(i)} f_j(U_t^{(j)}) X_t^{(j)} \,\mathrm{I}(\boldsymbol{U}_t \in C)$ for bounded functions f_i, f_j and $i, j = 1, \ldots, p$. Using Condition C.2 we can directly apply Theorem 4.4, provided we can show that $F^{(i,j)} \in L_1\big(C, \mathcal{B}(C), \pi\big)$. Since the functions are bounded, we only need to show $X_t^{(i)} X_t^{(j)} \in L_1\big(C, \mathcal{B}(C), \pi\big)$. Letting $\|X\|p = \big(\int_C |X|^p d\pi\big)^{1/p}$ and using Hölder's inequality, we have $\|X_t^{(i)} X_t^{(j)}\|_1 \leq \|X_t^{(i)}\|_2 \|X_t^{(j)}\|_2$. Let $\boldsymbol{e}_i$ be a vector of zeros with one at the i^{th} coordinate, then

$$\begin{aligned} \|X_t^{(i)}\|_2^2 &= \mathrm{E}_\pi\left[(X_t^{(i)})^2 \,\mathrm{I}(\boldsymbol{U}_t \in C)\right] = \mathrm{E}_{\pi,C}\left[\boldsymbol{e}_i^\top \boldsymbol{X}_t \boldsymbol{X}_t^\top \boldsymbol{e}_i\right] \\ &= \mathrm{E}_{\pi,C}\left[\boldsymbol{e}_i^\top \,\mathrm{E}_{\pi,C}\left[\boldsymbol{X}_t \boldsymbol{X}_t^\top | \boldsymbol{U}_t\right] \boldsymbol{e}_i\right] \leq M \,\mathrm{E}_{\pi,C}\left[\boldsymbol{e}_i^\top \boldsymbol{e}_i\right] = M \end{aligned}$$

by using Condition C.1.ii. Thus, $\|X_t^{(i)}\|_2 \leq \sqrt{M}$ for any i, and $F^{(i,j)} \in L_1\big(C, \mathcal{B}(C), \pi\big)$, so we can apply of Theorem 4.4 to prove equation (4.10). □

In the Markovian case, it is possible to verify the ergodicity of the process in terms of the functional coefficients of the true model. The preferred approach in the literature for establishing the ergodicity of the FAR model is based on drift conditions on the dynamics of the process. The definitive reference on the subject is the book of Meyn and Tweedie [84] on the stochastic stability of Markov chains. Under conditions on the true functions $\{f_i\}$ and the error distribution, we can ensure that (4.10) holds. The following theorem, which

is due to Chen [20], provides such sufficient conditions for the FAR model based on Foster-Lyapunov type criteria (for more details see Meyn and Tweedie [83] and Tjøstheim [112]).

Theorem 4.6. *(Chen [20], Thm 2.3, p. 31) Consider the Markovian FAR model*

$$X_t \quad = \quad f_1(U_t^{(1)})X_{t-1} + \ldots + f_p(U_t^{(p)})X_{t-p} + \epsilon_t$$

and assume that each function can be written as $f_i(U^{(i)}) = g_i(U^{(i)}) + h_i(U^{(i)})$ *such that* $g_i(\cdot)$ *and* $h_i(\cdot)$ *are bounded, with* $|g_i(U^{(i)})| < c_i$ *and* $h_i(U^{(i)})X_{t-i}$ *being uniformly bounded, for* $i = 1, \ldots, p$*. Furthermore, let* ϵ_t *be a white noise sequence with an everywhere positive density with respect to Lebesgue measure, and the roots of the characteristic polynomial* $\lambda^p - c_1\lambda^{p-1} - \ldots - c_p = 0$ *all lie inside the unit circle. Then, the process is geometrically ergodic.*

As a consequence of this theorem, Theorem 4.4 also holds and we can establish the ergodicity result we need. We can therefore substitute the rather general Condition C.2.ii with the more concrete conditions of Theorem 4.6 on the true functions and the error distribution. The conditions in Theorem 4.6 can be easily verified for a given set of functional coefficients, but they are rather strict. For instance, it is possible to have geometric ergodicity for FAR models whose functions are arbitrarily big, as Chan et al. [19] have shown for the first order TAR model, even though these counter examples are quite extreme. In reality, the condition on the characteristic roots is the most restrictive, because we try to control the behavior of the process by that of a stationary AR model. Nevertheless, Theorem 4.6 provides a practical tool for checking the stability of a fitted model, which is especially useful when we want to perform simulation.

As an example of this approach, consider again model (4.9) where the error term has a continuous positive density. If f_1 is bounded by c_1, f_2 bounded c_2 and the roots of

$\lambda^2 - c_1\lambda - c_2 = 0$ lie inside the unit circle, then the process is geometrically ergodic. Assume now that we do not know the true functions but we estimate them using our GP method, based on a Gaussian kernel k, and we want to check if the model we got is stationary and ergodic. For a fixed sample T, our estimates will be of the form $\hat{f}_i(\cdot) = \alpha_{i,0} + \sum_t \alpha_{i,t} k(\cdot, x_{t-2})$. For the first coefficient f_1, the argument of the function X_{t-2} is different from the regressor X_{t-1}, so we cannot find any useful decomposition $\hat{f}_1 = g_1 + h_1$ such that $h_1(x_1)x_2$ is bounded over the range of (x_1, x_2). However, we can still find an overall bound c_1, because $\hat{f}_1$ is represented by a sum of bounded functions plus a constant. For the second coefficient, on the other hand, we can take the decomposition $c_2 = \alpha_{2,0} + \epsilon$ (where $\epsilon > 0$) and $h_2(\cdot) = \sum_t \alpha_{2,t} k(\cdot, x_{t-2})$. The convenience of this choice stems from the fact that $h_2(x)x$ is bounded, since the Gaussian kernel decays exponentially fast. The bound c_2 is sharper than if we tried to bound the entire function $\hat{f}_2$, and the characteristic roots are more likely to lie within the unit circle. Therefore, our model's estimates, using a Gaussian kernel, are usually well suited for verifying the conditions of Theorem 4.6. We note for comparison that if we used a spline basis for the representation of the functions it would be almost impossible to verify the same conditions, because splines extrapolate linearly outside the observed range and there would be practically no chance of bounding the functional coefficients.

Mixing Conditions

The second set of conditions is used in a much more general setting; it does not require a Markovian structure and it applies to autoregression, time series regression and hybrids thereof. However, these conditions cannot be easily verified based on the true model. We allow the regressors $\{X^{(i)}\}$ and the functional coefficient arguments $\{U^{(i)}\}$ to be arbitrary time series. In order to prove ergodicity, we require the auto-dependence of the process to decay fast enough, and a boundedness condition on the moments of the $X^{(i)}$'s. We use the

following theorem of Davydov [26], in the form appearing in Bosq [9].

Theorem 4.7. *(Bosq [9], Thm 1.5, p. 34) Let $\{X_t\}$ be a zero mean, real valued strictly stationary and α-mixing process such that for some $r > 2$, $\mathrm{E}_\pi[|X_t|^r] < \infty$ and the mixing coefficients satisfy $\sum_{t\geq 1} \alpha(t)^{\frac{r-2}{r}} < \infty$. Then the series $\sum_{t\in\mathbb{Z}} \mathrm{Cov}(X_0, X_t)$ is absolutely convergent to $\sigma^2 \geq 0$ and*

$$\lim_{n\to\infty} n \,\mathrm{Var}_\pi \left[\frac{1}{n}\sum_{t=1}^{n} X_t\right] = \sigma^2.$$

The result gives mean square convergence, from which convergence in probability follows. We use Theorem 4.7 to establish ergodicity in the next proposition

Proposition 4.8. *Under Conditions C.1 and C.3, equation (4.10) holds.*

Proof: Let $Z_t^{(i,j)} = X_t^{(i)} f_i(U_t^{(i)}) X_t^{(j)} f_j(U_t^{(j)}) \,\mathrm{I}(\boldsymbol{U}_t \in C)$. We will show that the process $\{Z_t^{(i,j)}\}$ is also α-mixing, with the mixing coefficients satisfying Condition C.3.ii for all $i, j = 1, \ldots, p$. The α-mixing coefficient of two sub σ-fields $\mathcal{B}, \mathcal{C}$ of a probability space $(\Omega, \mathcal{A}, P)$ is defined as $\alpha(\mathcal{B},\mathcal{C}) = \sup_{B\in\mathcal{B}, C\in\mathcal{C}} |P(B\cap C) - P(B)P(C)|$, and the α-mixing coefficients of a stationary series $\{X_t\}$ are given by $\alpha_t = \alpha(\sigma(X_s), \sigma(X_{s+t}))$, where $\sigma(X_s)$ is the σ-field generated by X_s. Let $\alpha_t^{(i,j)}$ be the coefficients for $\{Z_t^{(i,j)}\}$ and α_t be the original coefficients for $\{\boldsymbol{X}_t, \boldsymbol{U}_t\}$. It is straightforward to see that $\alpha_t^{(i,j)} \leq \alpha_t$, because $Z_t^{(i,j)}$ is a function of $(\boldsymbol{X}_t, \boldsymbol{U}_t)$ and therefore $\sigma(Z_t^{(i,j)}) \subset \sigma(\boldsymbol{X}_t, \boldsymbol{U}_t)$. Thus, $\{Z_t^{(i,j)}\}$ is α-mixing (i.e. $\alpha_t^{(i,j)} \to 0$) and its coefficients satisfy Condition C.3.ii. Moreover, $\mathrm{E}_\pi[|Z_t^{(i,j)}|^r] \leq \infty$ for some $r > 2$, because of Condition C.3.i and the boundedness of the functions. Defining $\mu^{(i,j)} = \mathrm{E}_\pi[Z_t^{(i,j)}]$, the variable $Z_t^{(i,j)} - \mu^{(i,j)}$ obviously satisfies the conditions of Theorem 4.7, which gives

$$\lim_{T\to\infty} T \,\mathrm{Var}_\pi \left[\frac{1}{T}\sum_{t=1}^{T} (Z_t^{(i,j)} - \mu^{(i,j)})\right] = \sigma^2$$

for some $\sigma^2 \geq 0$. The validity of (4.10) follows readily from the above. $\square$

Checking whether a hypothesized model satisfies α-mixing is not easy; the most common cases in which one can prove strong mixing are for Markov models, where the last defy the purpose since we can use the first set of conditions anyway. We must therefore content ourselves with assuming beforehand that the process is α-mixing and just be aware of obvious deviations. For example, we can check if the autocorrelation function of a series decays to zero, since this is a necessary condition for α-mixing.

4.3.2 Proof of Consistency

We now present the proof of our main theoretical result

Proof of Theorem 4.3 : Assume we have T observations $\{y_t, \{x_t^{(i)}, u_t^{(i)}\}_{i=1}^p\}_{t=1}^T$ from the true model (4.8), where the true functions $\{f_i\}_{i=1}^p$ belong to their corresponding RKHS $\{\mathcal{K}_i\}_{i=1}^p$, each one having a reproducing kernel k_i. Moreover, we put a GP prior on each function $f_i \sim \mathcal{GP}(0, C_i)$, with zero mean and covariance kernel C_i equal to the reproducing kernel of that function. The estimators $\{\hat{f}_i\}_{i=1}^p$ of the functional coefficients over C are the posterior mean functions of our method, which by Lemma 4.2 are given as the solutions of the minimization problem

$$\min_{g_i \in \mathcal{K}_i} \left\{ \frac{1}{\sigma^2} \sum_{t=1}^{T} \left(y_t - g_1(u_t^{(1)}) x_t^{(1)} - \ldots - g_p(u_t^{(p)}) x_t^{(p)} \right)^2 \mathrm{I}(\boldsymbol{u}_t \in C) + \sum_{i=1}^{p} \|g_i\|^2_{\mathcal{K}_i(C)} \right\} \quad (4.12)$$

Letting $f_{i,t} = f_i(u_t^{(i)})$ and $x_{i,t} = x_t^{(i)}$, substituting $y_t = \sum_{i=1}^p f_{i,t} x_{i,t} + \epsilon_t$ in (4.12) and dividing by T, we equivalently get

$$\min_{g_i \in \mathcal{K}_i} \frac{1}{T} \left\{ \sum_{t=1}^{T} \left(\epsilon_t + \sum_{i=1}^{p} (f_{i,t} - g_{i,t}) x_{i,t} \right)^2 \mathrm{I}(\boldsymbol{u}_t \in C) + \sigma^2 \sum_{i=1}^{p} \|g_i\|^2_{\mathcal{K}_i(C)} \right\} \Leftrightarrow$$

$$\min_{g_i \in \mathcal{K}_i} \left\{ \left[\overbrace{\frac{1}{T}\sum_{t=1}^{T} \epsilon_t^2}^{I} + \overbrace{\frac{2}{T}\sum_{t=1}^{T} \epsilon_t \left(\sum_{i=1}^{p} (f_{i,t} - g_{i,t}) x_{i,t} \right)}^{II} + \right. \right.$$

$$\left. \left. + \underbrace{\frac{1}{T}\sum_{t=1}^{T} \left(\sum_{i=1}^{p} (f_{i,t} - g_{i,t}) x_{i,t} \right)^2}_{III} \right] \mathrm{I}(\boldsymbol{u}_t \in C) + \frac{\sigma^2}{T} \sum_{i=1}^{p} \|g_i\|^2_{\mathcal{K}_i(C)} \right\}$$

The squared error term I converges to σ^2 and the cross-terms II converges to zero, due to independence. Moreover, this convergence is uniform over the functions because they are bounded. We turn attention to term III; under Condition C.2 or Condition C.3, the average converges uniformly in the functional coefficients to the expectation with respect to the stationary measure

$$\frac{1}{T}\sum_{t=1}^{T} \left(\sum_{i=1}^{p} (f_{i,t} - g_{i,t}) x_{i,t} \right)^2 \mathrm{I}(\boldsymbol{u}_t \in C) \xrightarrow{P} \mathrm{E}_{\pi,C} \left[\left(\sum_{i=1}^{p} X_t^{(i)} \left(f_i(U_t^{(i)}) - g_i(U_t^{(i)}) \right) \right)^2 \right] \quad (4.13)$$

Thus, the posterior estimates minimize the following objective function

$$\mathrm{E}_{\pi,C} \left[\left(\sum_{i=1}^{p} X_t^{(i)} \left(f_i(U_t^{(i)}) - g_i(U_t^{(i)}) \right) \right)^2 \right] + \frac{\sigma^2}{T} \sum_{i=1}^{p} \|g_i\|^2_{\mathcal{K}_i(C)} + o_P(1) \quad (4.14)$$

Heuristically, the smoothness penalty dies off as $T \to \infty$ and we end up minimizing only the expectation. Letting $\boldsymbol{F}_t^\top = [f_1(U_t^{(1)}) - g_1(U_t^{(1)}), \ldots, f_p(U_t^{(p)}) - g_p(U_t^{(p)})]$, and $M > 0$ be the upper bound on the eigenvalues from Condition C.1.ii, we have

$$\begin{aligned} \mathrm{E}_{\pi,C} \left[\left(\sum_{i=1}^{p} X_t^{(i)} \left(f_i(U_t^{(i)}) - g_i(U_t^{(i)}) \right) \right)^2 \right] &= \mathrm{E}_{\pi,C} \left[\boldsymbol{F}_t^\top \boldsymbol{X}_t \boldsymbol{X}_t^\top \boldsymbol{F}_t \right] \\ &= \mathrm{E}_{\pi,C} \left[\mathrm{E}_{\pi,C} \left[\boldsymbol{F}_t^\top \boldsymbol{X}_t \boldsymbol{X}_t^\top \boldsymbol{F}_t | \boldsymbol{U}_t \right] \right] \\ &= \mathrm{E}_{\pi,C} \left[\boldsymbol{F}_t^\top \, \mathrm{E}_{\pi,C} \left[\boldsymbol{X}_t \boldsymbol{X}_t^\top | \boldsymbol{U}_t \right] \boldsymbol{F}_t \right] \end{aligned}$$

$$\begin{aligned} &\leq M\,\mathrm{E}_{\pi,C}\left[\boldsymbol{F}_t^\top \boldsymbol{F}_t\right] \\ &= M\sum_{i=1}^{p}\mathrm{E}_{\pi,C}\left[\left(f_i(U_t^{(i)}) - g_i(U_t^{(i)})\right)^2\right] \end{aligned}$$

We can, therefore, bound the objective function (4.14) from above by

$$M\sum_{i=1}^{p}\mathrm{E}_{\pi,C}\left[\left(f_i(U_t^{(i)}) - g_i(U_t^{(i)})\right)^2\right] + \frac{\sigma^2}{T}\sum_{i=1}^{p}\|g_i\|^2_{\mathcal{K}_i(C)} + o_P(1) \tag{4.15}$$

We apply an eigendecomposition in the RKHS in order to minimize (4.15), which allows us to translate the estimation problem into an infinite basis regression. We use Mercer's theorem (Theorem A.6 in the Appendix) to express the functions $f_i, g_i \in \mathcal{K}_i(C)$ in terms of their eigendecomposition with respect to the stationary measure, giving $f_i(\cdot) = \sum_{j=1}^{\infty} f_{i,j}\phi_{i,j}(\cdot)$ and $g_i(\cdot) = \sum_{j=1}^{\infty} g_{i,j}\phi_{i,j}(\cdot)$. Equation (4.15) becomes

$$\begin{aligned} &M\sum_{i=1}^{p}\int_C \left(f_i(u_i) - g_i(u_i)\right)^2 d\pi(\boldsymbol{u}_i) + \frac{\sigma^2}{T}\sum_{i=1}^{p}\|g_i\|^2_{\mathcal{K}_i(C)} + o_P(1) \\ &\qquad = \sum_{i=1}^{p}\left\{\sum_{j=1}^{\infty}\left\{M(f_{i,j} - g_{i,j})^2 + \frac{\sigma^2 g_{i,j}^2}{T\lambda_{i,j}}\right\}\right\} + o_P(1) \end{aligned}$$

We can readily minimize the above expression by differentiating it with respect to each $f_{i,j}$ and setting the derivative to zero. From the uniformity of convergence we can assume the $o_P(1)$ term is independent of the functions, so the solutions are given by

$$g^\star_{i,j} = \frac{\lambda_{i,j}}{\lambda_{i,j} + \sigma^2/(TM)} f_{i,j} \tag{4.16}$$

The upper bound (4.15) converges to zero and the solutions converge to the true functions, as $T \to \infty$. To verify this, note that for any $\varepsilon > 0$ there exist $n_1, n_2 > 0$ such that

$$\|f_i - g_i^\star\|^2_{\mathcal{K}_i(C)} = \sum_{j=1}^{\infty}\frac{(f_{i,j} - g^\star_{i,j})^2}{\lambda_{i,j}} = \sum_{j=1}^{\infty}\frac{f_{i,j}^2}{\lambda_{i,j}}\left(\frac{\sigma^2}{TM\lambda_{i,j} + \sigma^2}\right)^2$$

$$< \sum_{j=1}^{n_1} \frac{f_{i,j}^2}{\lambda_{i,j}} \left(\frac{\sigma^2}{TM\lambda_{i,n_1} + \sigma^2} \right)^2 + \varepsilon/2 < \varepsilon, \quad \forall T > n_2, \;\; i = 1, \ldots, p$$

where we use the fact that the series giving the norm is convergent, so we can bound its tail. Thus, the objective function (4.14) is also minimized by (4.16). Moreover, our estimators $\{\hat{f}_i\}$ must satisfy $\|f_i - \hat{f}_i\|^2_{\mathcal{K}_i(C)} = o_P(1)$, as $T \to \infty$ and for $i = 1, \ldots, p$. This is readily proved by contradiction, using the lower bound on the eigenvalues of $\mathrm{E}_\pi[\boldsymbol{X}_t \boldsymbol{X}_t^\top | \boldsymbol{U}_t = \boldsymbol{u}_t]$ and following the same steps as before. If any one of the estimates does not converge to its true coefficient function, then the objective function (4.14) will not attain its minimum as given by (4.16). □

4.4 Properties of Reduced Rank Approximation

In this section we examine the theoretical properties of the reduced rank approximation method we introduced in Chapter 3. The main characteristic in this setting is that we use a fixed number of bases in the representation of the functional coefficients. Therefore, we look at estimation as the number of observations increases and the function space basis remains constant, in contrast to the exact case where we have as many basis functions as observations. We assume that the estimated coefficient functions are represented by $\hat{f}_i(u) = \sum_{j=1}^{m_i} \alpha_{i,j} \phi_{i,j}(u)$, where the basis functions $\phi_{i,j}(u)$ are kernels $C_i(u, u_j^{(i)})$ centered at the basis points $u_j^{(i)}$. The resulting RKHS is degenerate, in the sense that it has a finite dimensional eigendecomposition. Therefore, we relax the requirement that the true functions belong to these RKHSs and substitute Condition C.1 with the following

Condition C.4. *Let* $\boldsymbol{X}_t^\top = [X_t^{(1)}, \ldots, X_t^{(p)}]$ *and* $\boldsymbol{U}_t^\top = [U_t^{(1)}, \ldots, U_t^{(p)}]$.

i. The process $\{\boldsymbol{X}_t, \boldsymbol{U}_t\}$ *is jointly strictly stationary, with stationary measure* π.

ii. There exists a compact set $C \in \mathbb{R}^p$ *with positive Lebesgue measure and such that the*

eigenvalues of $\mathrm{E}_\pi[\boldsymbol{X}_t\boldsymbol{X}_t^\top|\boldsymbol{U}_t = \boldsymbol{u}_t]$ *are uniformly bounded away from zero and infinity for all* $\boldsymbol{u}_t \in C$ *and the density of* $\boldsymbol{U}_t$ *over* C *is bounded away from zero.*

iii. The functions $\{f_i\}_{i=1}^p$ *are measurable and bounded.*

iv. The error term $\{\epsilon_t\}$ *is a white noise series.*

Our result for the approximate setting shows that the estimated coefficient functions converge to the projection of the true functions on the approximation space under a particular norm. The following proposition defines this norm

Proposition 4.9. *Consider coefficient functions* f_i *belonging to the space of bounded, measurable functions* $\mathcal{B}_i$*, for* $i = 1\ldots, p$*, and define the corresponding product space* $\mathcal{B} = \otimes_{i=1}^p \mathcal{B}_i$ *of p-tuples* $\boldsymbol{f} = [f_i, \ldots, f_p]$ *of such functions. Assume Condition C.4 holds and let* $\mathcal{B}(C)$ *be the restriction of* $\mathcal{B}$ *to functions over* C*. Then, for* $\boldsymbol{f}, \boldsymbol{g} \in \mathcal{B}(C)$*, the function*

$$< \boldsymbol{f}, \boldsymbol{g} >_{\mathcal{B}(C)} = \mathrm{E}_{\pi,C}\left[\sum_{i,j=1}^{p} X_t^{(i)} f_i(U_t^{(i)}) X_t^{(j)} g_j(U_t^{(j)})\right]$$

defines an inner product in $\mathcal{B}(C)$*, with the induced norm being equivalent to Lebesgue norm.*

Proof: The symmetry, linearity and non-negativity of the function is obvious. We just need to prove non-degeneracy, but this will be a byproduct of the equivalence to Lebesgue norm. We have

$$\begin{aligned}
\|\boldsymbol{f}\|^2_{\mathcal{B}(C)} =< \boldsymbol{f}, \boldsymbol{f} >_{\mathcal{B}(C)} &= \mathrm{E}_{\pi,C}\left[\left(\sum_{i=1}^{p} X_t^{(i)} f_i(U_t^{(i)})\right)^2\right] \\
&= \mathrm{E}_{\pi,C}\left[\boldsymbol{F}_t^\top \boldsymbol{X}_t\boldsymbol{X}_t^\top \boldsymbol{F}_t\right] = \mathrm{E}_{\pi,C}\left[\mathrm{E}_{\pi,C}\left[\boldsymbol{F}_t^\top \boldsymbol{X}_t\boldsymbol{X}_t^\top \boldsymbol{F}_t|\boldsymbol{U}_t\right]\right] \\
&= \mathrm{E}_{\pi,C}\left[\boldsymbol{F}_t^\top \mathrm{E}_{\pi,C}\left[\boldsymbol{X}_t\boldsymbol{X}_t^\top|\boldsymbol{U}_t\right]\boldsymbol{F}_t\right] \geq m\,\mathrm{E}_{\pi,C}\left[\boldsymbol{F}_t^\top \boldsymbol{F}_t\right] \\
&= m\sum_{i=1}^{p}\mathrm{E}_{\pi,C}\left[f_i^2(U_t^{(i)})\right] \geq m'\sum_{i=1}^{p}\int_C f_i^2(U_t^{(i)})d\mu
\end{aligned}$$

where the last two inequalities follow from Condition C.1.ii on the lower bound of the eigenvalues and the stationary density. By using the upper bound on the eigenvalues, we can similarly bound the norm from above, with respect to Lebesgue norm. As a result, the two norms are equivalent and define the same topology in $\mathcal{B}(C)$. □

As a consequence of Proposition 4.9 we can define projections on $\mathcal{B}(C)$. We are now able to state the result for approximate inference

Theorem 4.10. *Assume Condition C.4 and either Condition C.2 or Condition C.3 hold. Also assume that, in the reduced rank setting, each functional coefficient f_i is represented in terms of distinct basis functions $\{\phi_{i,j}\}_{j=1}^{m_i}$. Let $\mathcal{S}_i = \text{span}\{\phi_{i,j}; j = 1, \ldots, m_i\}$ be the span of the bases, $\mathcal{S} = \otimes_{i=1}^{p} \mathcal{S}_i$ be their product space and $\boldsymbol{f}^{\parallel}$ be the projection of $\boldsymbol{f}$ on $\mathcal{S}(C)$ under the metric of Proposition 4.9. Then, the posterior means $\{\hat{f}_i\}$ of the functional coefficients satisfy*

$$\|f_i^{\parallel} - \hat{f}_i\|^2_{L^2(C)} \xrightarrow{P} 0; \;\; T \to \infty, \; \forall i = 1, \ldots, p$$

where T is the number of observations and convergence is in the restriction of L^2 over C.

Proof: Following the same reasoning as in the proof of Theorem 4.3, and replacing $f_i = f_i^{\parallel} + f_i^{\perp}$, the estimators must minimize

$$\mathrm{E}_{\pi,C}\left[\left(\sum_{i=1}^{p} X_t^{(i)} \left(f_i^{\parallel}(U_t^{(i)}) + f_i^{\perp}(U_t^{(i)}) - g_i(U_t^{(i)})\right)\right)^2\right] + \frac{\sigma^2}{T}\sum_{i=1}^{p} \|g_i\|^2_{\mathcal{K}_i(C)} + o_P(1)$$

We can split the expectation as

$$\mathrm{E}_{\pi,C}\left[\left(\sum_{i=1}^{p} X_t^{(i)} \left(f_i^{\parallel}(U_t^{(i)}) + f_i^{\perp}(U_t^{(i)}) - g_i(U_t^{(i)})\right)\right)^2\right] =$$

$$= \mathrm{E}_{\pi,C}\left[\left(\sum_{i=1}^{p} X_t^{(i)}\left(f_i^{\parallel}(U_t^{(i)}) - g_i(U_t^{(i)})\right)\right)^2\right] + \mathrm{E}_{\pi,C}\left[\left(\sum_{i=1}^{p} X_t^{(i)} f_i^{\perp}(U_t^{(i)})\right)^2\right] +$$
$$+2\,\mathrm{E}_{\pi,C}\left[\left(\sum_{i=1}^{p} X_t^{(i)} f_i^{\perp}(U_t^{(i)})\right)\left(\sum_{i=1}^{p} X_t^{(i)}\left(f_i^{\parallel}(U_t^{(i)}) - g_i(U_t^{(i)})\right)\right)\right]$$

The cross-term is zero due to orthogonality, and the second term is independent of the estimates. Therefore, we can equivalently minimize

$$\mathrm{E}_{\pi,C}\left[\left(\sum_{i=1}^{p} X_t^{(i)}\left(f_i^{\parallel}(U_t^{(i)}) - g_i(U_t^{(i)})\right)\right)^2\right] + \frac{\sigma^2}{T}\sum_{i=1}^{p}\|g_i\|^2_{\mathcal{K}_i(C)} + o_P(1)$$

Using Condition C.4.ii, we minimize the upper bound

$$M\sum_{i=1}^{p}\mathrm{E}_{\pi,C}\left[\left(f_i^{\parallel}(U_t^{(i)}) - g_i(U_t^{(i)})\right)^2\right] + \frac{\sigma^2}{T}\sum_{i=1}^{p}\|g_i\|^2_{\mathcal{K}_i(C)} + o_P(1)$$

We substitute the functions $f_i^{\parallel}, g_i$ with their basis expansion $f_i^{\parallel}(u) = \sum_{j=1}^{m_i}\alpha_{i,j}\phi_{i,j}(u)$, $g_i(u) = \sum_{j=1}^{m_i}\beta_{i,j}\phi_{i,j}(u)$ to get

$$M\sum_{i=1}^{p}\mathrm{E}_{\pi,C}\left[\left(\sum_{j=1}^{m_i}(\alpha_{i,j} - \beta_{i,j})\phi_{i,j}(U_t^{(i)})\right)^2\right] + \frac{\sigma^2}{T}\sum_{i=1}^{p}\|g_i\|^2_{\mathcal{K}_i(C)} + o_P(1) \tag{4.17}$$

Let $\boldsymbol{\Phi}_i^{\top} = [\phi_{i,1}(U_t^{(i)}), \ldots, \phi_{i,m_i}(U_t^{(i)})]$ and $\boldsymbol{F}_i = \mathrm{E}_{\pi,C}[\boldsymbol{\Phi}_i\boldsymbol{\Phi}_i^{\top}]$, where $\boldsymbol{F}_i$ is positive definite. Also, let $\boldsymbol{\beta}_i^{\top} = [\beta_{i,1}, \ldots, \beta_{i,m_i}]$ and similarly for $\boldsymbol{\alpha}_i$, and let $\|g_i\|^2_{\mathcal{K}_i(C)} = \boldsymbol{\beta}_i^{\top}\boldsymbol{E}_i\boldsymbol{\beta}_i$. We need to minimize

$$\sum_{i=1}^{p}\left(M(\boldsymbol{\alpha}_i - \boldsymbol{\beta}_i)^{\top}\boldsymbol{F}_i(\boldsymbol{\alpha}_i - \boldsymbol{\beta}_i) + \frac{\sigma^2}{T}\boldsymbol{\beta}_i^{\top}\boldsymbol{E}_i\boldsymbol{\beta}_i\right) + o_P(1)$$

The solutions are given by $\boldsymbol{\beta}_i^{\star} = \left(\boldsymbol{F}_i + \sigma^2/(TM)\boldsymbol{E}_i\right)^{-1}\boldsymbol{F}_i\boldsymbol{\alpha}_i$, $i = 1, \ldots, p$, which converge to the coefficients of the projection as $T \to \infty$. Moreover, the bound on the objective

function (4.17) goes to zero, so the estimates $\hat{f}_i$ converge to the projections $f_i^{\|}$ under the norm of Proposition 4.9 and, consequently, in the L_2 space. □

Note that the projection of each function depends on the other functions through the dependence of the regressors $X_t^{(i)}$ and the arguments $U_t^{(i)}$. As a result, they can be quite different from the usual L^2 projections, although we expect them to be close when the regressors and arguments are loosely dependent. Moreover, even though the projections are dependent on the set C, they are consistent in the sense that if C' is another set satisfying Condition C.4.ii, the projections of the true functional coefficients will agree on $C \cap C'$. Theorem 4.10 is comparable to the consistency result of Huang and Shen [60] for spline regression.

4.5 Comments

We now provide some general comments on the results we have presented. First, we point out that all of the theory carries over from the FAR to the varying coefficient regression setting. In particular, when the regressor and argument variables are independent we can to remove the ergodicity requirements and still get the same results. Throughout this chapter, we assumed zero prior means for the functional coefficients, contrary to what we propose in practice. We did this for simplicity, but also because the prior mean is not important for our consistency result. The consistency of our estimators is proved over a compact set with positive probability, and the effect of any prior bias over this set diminishes asymptotically. On the other hand, our proposal of non-zero prior means is motivated by the need to control the behavior of the estimates for finite data and outside their observed range. Moreover, in Theorem 4.3 we also make the simplifying assumption that we have perfect knowledge of the RKHSs of the true functions. Essentially, we assume that we are using the correct prior covariance kernel for each function, but in practice we use the data to select the kernel

hyperparameters, i.e. the parameters which control smoothing such as the characteristic lengthscales. Therefore, it would be interesting to investigate the behavior of our procedure from the viewpoint of optimal adaptive estimation. This task is quite involved, however, since it would also require establishing convergence rates for different classes of functions, which we have not addressed here. Finally, we look at Theorem 4.10 for reduced rank estimation. The theorem establishes the convergence of our estimators to the projection of the true functions on the finite dimensional RKHSs of the approximate kernels. In most cases, we expect the true functional coefficients to be smooth, so that the projection error will be small. In particular, if we allow the number of bases in the approximation to increase with the number of data in such a way that the projection error goes to zero, we will end up with the original consistency result for the exact method.

Chapter 5

Identification and Inference

In this chapter we address further issues regarding the application of our proposed model, our goal being to provide a complete and integrated statistical procedure for modeling and inference. In particular, we look at model comparisons and how they can be applied for model identification, by proposing a greedy model selection algorithm. We also look at diagnostic checking for a given model specification; we examine the universal residuals for our method and describe how they can be computed efficiently and used for goodness of fit tests. Finally, we suggest simulation and graphical methods for revealing the dynamical structure of a model.

5.1 Model Comparisons

We propose a method for comparing different models, which we also use for our model selection procedure. Our estimation technique provides us with the marginal model likelihood, which is the basic quantity we use for these purposes. This is a distinguishing characteristic of our model compared to other nonparametric estimation techniques which rely on least squares. Suppose we want to compare two different model specifications, $M1$ and $M2$, with marginal likelihoods given by $\mathcal{L}_{M_1}(\boldsymbol{y})$ and $\mathcal{L}_{M_2}(\boldsymbol{y})$, respectively. Our empirical Bayes esti-

mation selects the hyperparameters by maximizing $\mathcal{L}(\boldsymbol{y})$, and this will always favor more complex models. Therefore, we need to measure the complexity of competing models and impose a penalty on their marginal likelihood based on it, in order to account for the extra flexibility we allow for the coefficients. In nonparametric regression, this measure of complexity or flexibility appears under the name of effective degrees of freedom (edf). For the class of linear smoothers, in particular, there are at least three alternative definitions of the edf, all of which rely on the smoother or hat matrix $\boldsymbol{H}$, given by the formula for the fitted values $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$. These definitions are explicitly $\mathrm{tr}(\boldsymbol{H})$, $\mathrm{tr}(\boldsymbol{H}\boldsymbol{H}^\top)$ and $\mathrm{tr}(2\boldsymbol{H} - \boldsymbol{H}\boldsymbol{H}^\top)$, and they are motivated by analogies to linear regression, for more details see chapter 3.5 of Hastie and Tibshirani [56]. In practice, the most popular is the trace of the hat matrix, and for smoothing splines its edf values lie between those of the other two definitions.

We define the edf of our model in a similar manner. As in the nonparametric regression setting, the rationale behind our definition comes from the formula for the fitted values, which is given in eqn (2.21) and which we reproduce below

$$\hat{\boldsymbol{y}} \;=\; \boldsymbol{X}^\top\boldsymbol{\mu} + \boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}(\boldsymbol{\Sigma} + \boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X})^{-1}(\boldsymbol{y} - \boldsymbol{X}^\top\boldsymbol{\mu}) \tag{5.1}$$

The matrix $\boldsymbol{H} = \boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}(\boldsymbol{\Sigma}+\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X})^{-1}$ plays the role of the hat matrix. We can interpret $\hat{\boldsymbol{y}}$ as a combination of the fitted values $\boldsymbol{X}^\top\boldsymbol{\mu}$ from a linear model (since the prior means are constant) and the fitted values of a nonparametric smoother $\boldsymbol{H}$ applied to the deviations $(\boldsymbol{y}-\boldsymbol{X}^\top\boldsymbol{\mu})$ from the linear model. Based on this observation, we define the effective degrees of freedom DF as

$$DF = p + \mathrm{tr}(\boldsymbol{H}) \tag{5.2}$$

The first term p is just the order of the model, i.e. the number of functional coefficients.

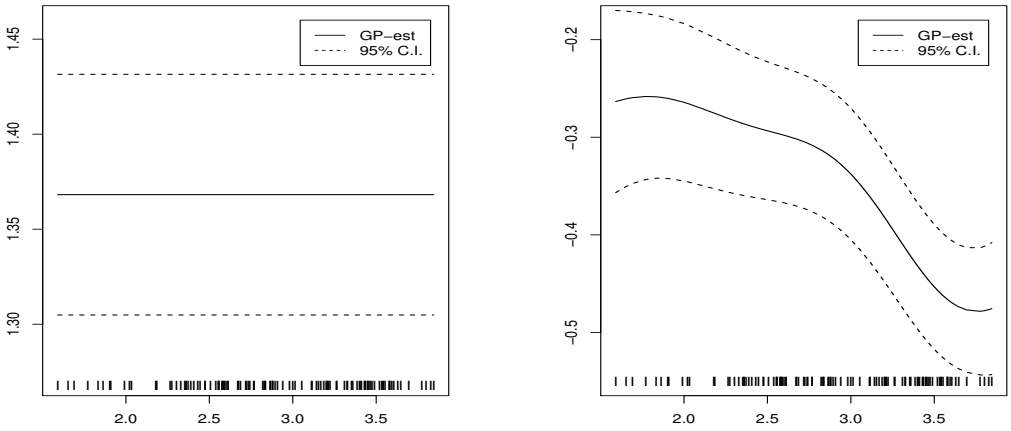
However, a better way to think of this term is as the number of free parameters required for describing the prior mean functions, and this is what we would use if we allowed more complex parametric forms for each μ_i. The second term is the trace of the hat matrix, the definition we adopt for measuring the complexity of the nonlinear part. We add the two terms because our framework does not impose a smoothness penalty on the prior mean functions and the hat matrix $\boldsymbol{H}$ is independent of $\boldsymbol{\mu}$. We choose the trace of the hat matrix over alternative definitions for two reasons: first, it is easier to calculate and second, its values tend to lie between those of the other two definitions. We point out that our model, as well as other nonparametric estimation techniques for time series, do not classify as linear smoothers. When working with autoregressive models, the observations also serve as regressor or argument variables and this invalidates much of the theoretical reasoning behind using the trace of $\boldsymbol{H}$. Lin and Pourahmadi [72] point out the dangers of treating the lags of the series as fixed design variables and propose simulation alternatives. Nevertheless, our definition in (5.2) has intuitively the right behavior, the edf decrease when any one of the smoothing parameters h_i increase, and it also works reasonably well in practice. It would be instructive to try and develop more robust theoretical arguments for measuring edf in time series autoregression, but we do not pursue this direction.

Based on our edf definition, we use standard methods for comparing different models such as Akaike's information criterion (AIC) given by $AIC = -2\ell(\boldsymbol{y}) + 2DF$ and introduced by Akaike [1], or the Bayesian information criterion (BIC) given by $BIC = -2\ell(\boldsymbol{y}) + \log(T)DF$ and proposed by Schwartz [100]. We give an example of our model comparison method applied to the Canadian lynx data. We look at the second order FAR model (2.30) that we fit in section 2.6 and we compare it to the same model, restricted so that the first functional coefficient is constant. The hyperparameters that maximize the marginal likelihood of the restricted model are given in Table 5.1 and the posterior of the coefficients is presented

Table 5.1: The hyperparameters that maximize the marginal likelihood of the Canadian lynx series, with constant f_1.

		f_1		f_2	
σ	0.2096703	μ_1	1.3681759	μ_2	-0.3458263
		h_1	∞	h_2	0.736213
ℓ=9.170201					

in Fig. 5.1, and they are comparable to Table 2.1 and Fig. 2.2, respectively. The two specifications are very close in every respect, with the likelihood of the unrestricted model being greater, as we would expect. However, the restricted model has lower edf than the original model (6.66 vs. 6.78) and also a lower AIC (-5.02 vs. -4.95) and BIC (13.08 vs. 13.48) score. This suggests that, based on any of the two criteria, the loss in flexibility from constraining the first functional coefficient to be constant does not surpass the gain in parsimony.



(a) (b)

Figure 5.1: Plots of GP estimated functional coefficients for (a) f_1 and (b) f_2 for the Canadian lynx series, with constant f_1.

5.2 Model Selection

In this section we describe the use of the previously presented criteria for selecting a model specification. Given a time series data set with no further information about the generating process, we need to build a model that adequately describes its dynamics. In the context of the FAR model we are working with, this amounts to specifying the regressor and functional coefficient argument variables in $Y_t = f_1(U_t^{(1)})X_t^{(1)} + \ldots + f_p(U_t^{(p)})X_t^{(p)} + \epsilon_t$. We formalize the model selection procedure by only considering regressors from a fixed set of possible variables $\{X^{(i)}\}_{i=1}^{P}$, and arguments from a fixed set of possible variables $\{U^{(j)}\}_{j=1}^{Q}$. The two sets can be arbitrary, provided they include $\mathcal{F}_{t-1}$ measurable variables so that the conditional analysis carries through. The natural choice for autoregressive models is to use lagged values of the series itself, without excluding other possibilities such as exogenous variables. The total number of models we are facing in this case is 2^{PQ}, allowing the same regressors to appear more that once under different functional coefficient arguments. An exhaustive search in this setting soon becomes prohibitive, since the possible number of models we have to fit grows exponentially in both P and Q. There is an obvious need for a faster model selection scheme, therefore we look at greedy model selection procedures. In our case, the backward selection procedure is substantially more complicated because the full model is very big. In order not to exclude any possibility, we would have to start from a model with PQ terms; each of the P regressors $X^{(i)}$ must be multiplied with Q different functional coefficients, one for each possible argument $U^{(j)}$. Moreover, the backward selection procedure, tends to give bigger models and requires significantly more computations. For this reason, we propose a forward selection procedure: we start with an empty (zero mean) model and at each step n of the algorithm we include the new term $X^{(i_n)} f_n(U^{(j_n)})$ that minimizes a criterion C, until there is no further improvement. More specifically, assuming we want to minimize some criterion C, either AIC or BIC, we use Algorithm 1.

Algorithm 1 Forward model selection algorithm.

1. Initialize the index set $I = \{(i,j)|i = 1,\ldots,P; j = 1,\ldots,Q\}$ and set $n = 0$
2. Let M_0^* be the empty model and set C_0^* equal to the criterion value for it.
3. For every $k = (l,m) \in I$, fit the model $M_k : \quad Y = \sum_{r=1}^{n} f_r(U^{(j_r)})X^{(i_r)} + f_{n+1}(U^{(m)})X^{(l)}$ and calculate its criterion value C_k.
4. Find the optimal k^* such that $C_{k^*} \leq C_k$, $\forall k \in I$. Set $i_{n+1} = l^*, j_{n+1} = m^*$, $C_{n+1}^* = C_{k^*}$, and let the optimal model at step $n+1$ be $M_{n+1}^* = M_{k^*}$. Remove $k^* = (l^*, m^*)$ from the index set I.
5. If $C_{n+1}^* \geq C_n^*$ or I is empty, stop the algorithm and set the selected model equal to M_n^*. Otherwise, set $n = n+1$ and go to 3.

At the n^{th} step of this algorithm, we have to fit $PQ-n$ different models, so the total number of models we fit is approximately PQ times the order of the resulting specification. The resulting model can have a regressor variable being multiplied with an additive function of many arguments. By this, we mean that we can get specifications of the form

$$Y = f_1(U^{(1)})X^{(1)} + f_2(U^{(2)})X^{(1)} = \left[f_1(U^{(1)}) + f_2(U^{(2)})\right]X^{(1)}$$

where the coefficient of $X^{(1)}$ is essentially an additive function of $(U^{(1)}, U^{(2)})$. This can be generalized to more complex settings. In particular, if we want to allow for a possibly varying mean in the model, we can take $X^{(1)}$ to be a constant unity variable. Thus, the selection procedure can produce additive models as well; we can get specifications such as $Y = f_1(U^{(1)}) + f_2(U^{(2)})$ or hybrids between FAR and additive models, such as $Y = f_1(U^{(1)})X^{(1)} + f_2(U^{(2)})$.

The basic forward selection scheme can be modified in different ways. We propose an important adjustment that replaces very smooth functional coefficients with constants, as in the example of the previous section. This helps substantially in avoiding overfitting, speeding

up the selection process and improving interpretability through simplifying the resulting model. We introduce this modification at the third step of the selection procedure: for each candidate model that we fit, we test whether any of its coefficients are constant. We do this in a stepwise manner, we choose the flattest coefficient and refit a restricted model with that coefficient being constant, where we identify the flattest coefficient as the one with the minimum average deviation from its mean at the observed points. We compare the criterion values of the two models and keep the best one, and we repeat this until there is no further improvement. As a result, each of the candidate models at step three can have some of its coefficients being constant. At step four we select the optimal model, and any of its coefficients that is identified as constant at this point remains so for the subsequent steps of the selection procedure. This means that all candidate models will have at least the same constant coefficient as the most recent optimal model. The details are presented more formally in Algorithm 2.

Although this modification seems to complicate the procedure, it can actually offer significant speed gains. These occur when we fit the candidate models, because we reduce the parameter space by restricting some coefficients to be constant. As a result, the numerical maximization of the marginal likelihood is faster and more stable. The basic forward model selection algorithm can easily accommodate further modifications, an obvious one being the inclusion of a backward step. For additional speed-up, we could allow regressors to appear only once in the model, or we could allow only specific regressor-argument combinations based on contextual knowledge.

5.2.1 Examples

We now give an example of our forward selection procedure applied to the Canadian lynx data. For both regressors and arguments we consider lagged values of the series of order

Algorithm 2 Forward model selection algorithm with constant coefficients.

1. Initialize the index set $I = \{(i,j)|i = 1,\dots,p; j = 1,\dots,q\}$, the constant coefficient set $J = \emptyset$, and set $n = 0$
2. Let M_0^* be the empty model and set C_0^* equal to the criterion value for it.
3. For every $k = (l,m) \in I$:
   (i) Set $J^k = J$, fit the model $M_k : Y = \sum_{r=1}^{n} f_r(U^{(j_r)})X^{(i_r)} + f_{n+1}(U^{(m)})X^{(l)}$ with f_r being constant for $r \in J^k$ and calculate its criterion value C_k.
   (ii) Find $r^c \in (\{1,\dots,n+1\} - J^k)$ such that it minimizes $\sum_{t=1}^{T}[f_r(U_t^{(j_r)}) - \bar{f}_r]^2$, where $\bar{f}_r = \sum_{t=1}^{T} f_r(U_t^{(j_r)})/T$
   (iii) Fit the model $M_k^c : Y = \sum_{r=1}^{n} f_r(U^{(j_r)})X^{(i_r)} + f_{n+1}(U^{(m)})X^{(l)}$ with f_r being constant for $r \in J^k \cup \{r^c\}$ and calculate its criterion value C_k^c.
   (iv) If $C_k^c \le C_k$, set $C_k = C_k^c$, $M_k = M_k^c$, $J^k = J^k \cup \{r^c\}$, and if there are non-constant terms left, go to (ii). Otherwise proceed with a different k.
4. Find the optimal k^* such that $C_{k^*} \le C_k$, $\forall k \in I$. Set $i_{n+1} = l^*, j_{n+1} = m^*$, $C_{n+1}^* = C_{k^*}$, $J = J_{k^*}$, and let the optimal model at step $n+1$ be $M_{n+1}^* = M_{k^*}$. Remove $k^* = (l^*, m^*)$ from the index set I.
5. If $C_{n+1}^* \ge C_n^*$ or I is empty, stop the algorithm and set the selected model equal to M_n^*. Otherwise, set $n = n+1$ and go to 3.

up to five, and we also include a unit regressor variable to allow for additive terms. The selected model specification using AIC has three terms and is given by

$$X_t = f_1(X_{t-3})X_{t-1} + f_2(X_{t-2})X_{t-2} + f_3(X_{t-4}) + \epsilon_t$$

were f_1 is a constant and f_3 is very smooth. Fig. 5.2 shows the optimal model at the end of each step of the model selection procedure and Table 5.2 gives information on the optimal model at different steps. Notice that the first term $f_1(X_{t-3})X_{t-1}$ enters in the model with a varying coefficient, but from the second step onward it becomes constant, the second term absorbing most of the nonlinearity. We also performed the model selection procedure with BIC as a criterion, and the resulting model includes only the first two terms of the AIC

selected model. The plots of the coefficients are given in the second row of Fig. 5.2 and the maximum BIC is presented in Table 5.2, for the second step. The BIC criterion generally favors more parsimonious models because it imposes a heavier complexity penalty. We also point out that the specification selected by BIC is almost the same as the one in section 2.6, the difference in the first argument variable being negligible because the function is almost constant.

Table 5.2: Information on forward model selection steps for the Canadian lynx data, with 1-5 candidate regressor lags.

step	$\ell(\boldsymbol{y})$	DF	AIC	BIC
1	-3.1180	4.1800	14.5960	25.8460
2	7.4584	6.5994	-1.7178	16.0436
3	9.8201	8.3587	-2.9227	19.5734

We also apply the forward selection procedure on a bigger set of regressors which contains lagged series of up to order 15. We do this because the lynx series demonstrates an approximate ten year cycle, and also because the best linear AR model, selected by AIC, is of order 11 (see Tong [113]). For the coefficient arguments we still use up to five lags. The results of the selection based on AIC are shown in Fig. 5.3, and information on the criteria are given in Table 5.3. This time the AIC criterion selects a bigger model of order four, where the first term is the same as before, but the rest have regressors at higher lags. This is not so surprising given that the series exhibits cyclic behavior, so higher lags can still be informative. The BIC criterion again selects a smaller model with three terms, which is identical to that of the third row of Fig. 5.3. Notice that the model in the third row has less DF than the model in the second row, despite having more regressors. This is due to the fact that the three term model has two constant functions, compared to none of the two term model. Even though the selected models for the two different candidate regressor sets seem very different, their fits are close. Moreover, the lynx series has only 114 observations in total, and by conditioning on lags of up to 15 we reduce the actual number to 99. Besides

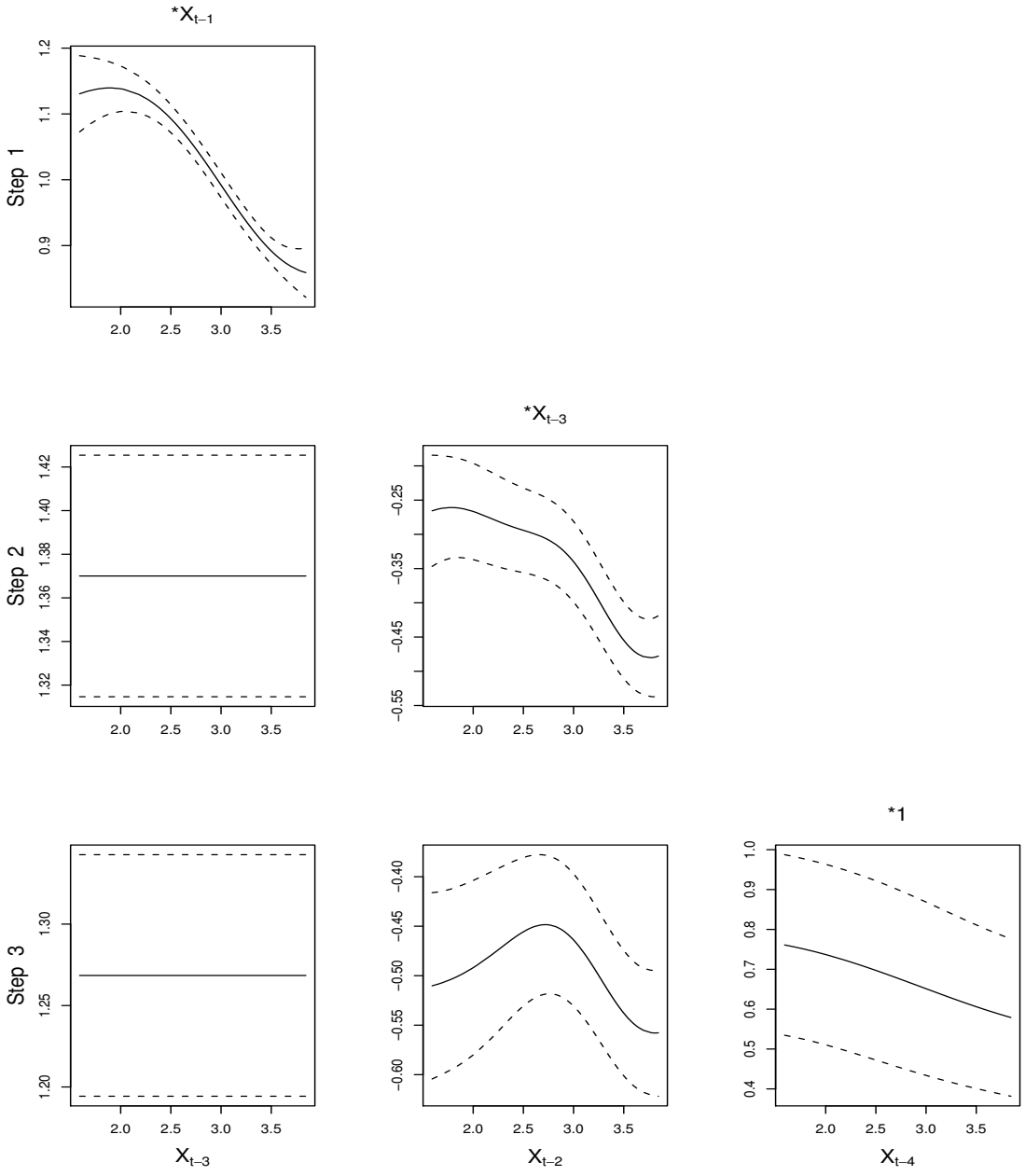


Figure 5.2: Plots of forward model selection steps for the Canadian lynx data using AIC; labels at the top of the plots indicate regressor variables and labels at the bottom indicate argument variables.

the fact that the procedures use slightly different data sets, there are not enough data to provide robust selection results.

Table 5.3: Information on forward model selection steps for the Canadian lynx data, with 1-11 candidate regressor lags.

step	$\ell(\boldsymbol{y})$	DF	AIC	BIC
1	-1.4819	4.2826	11.5289	22.6427
2	13.1126	8.7879	-8.6493	14.1565
3	17.1014	7.9212	-18.3604	2.1960
4	21.7126	11.0749	-21.2752	7.4656

We also demonstrate the performance of our model selection procedure on two simulated data sets, a nonlinear FAR model and a simple linear AR model. The dynamics of the nonlinear model are given by the following equation

$$X_t = \mu + f_1(X_{t-2})X_{t-1} + f_2(X_{t-1})X_{t-3} + \epsilon_t \tag{5.3}$$

where $\mu = 2$, $\epsilon_t \overset{i.i.d}{\sim} \mathcal{N}(0, .5^2)$, $f_1(x) = .5\sin(2\pi(x-\mu)/.3)\exp\{-2(x-\mu)^2\}$ is an exponentially decaying sinusoidal and $f_2(x) = -.5\tanh(x-\mu) + .3$ is a sigmoid; Fig. 5.4 shows a realization of 500 observations from this model. We apply our model selection procedure using the PP approximation with 10 equally spaced bases for each function, under both BIC and AIC. For the regressors we allow up to seven lags plus a unity variable for additive terms, and for the arguments we allow up to five lags. The BIC criterion manages to identify the correct model specification and the resulting estimates are shown in Fig. 5.5, together with the true coefficient functions. The mean of the series is represented as a constant function multiplied by the unity variable. The other two functions are actually varying, and describe the shape of the true functions accurately. For the first two estimates, the true functions fall slightly outside the 95% pointwise confidence intervals, but notice that they do so in opposite ways. The mean level is higher and the first coefficient is more negative, but since the series is always positive these effects partially cancel out for the conditional

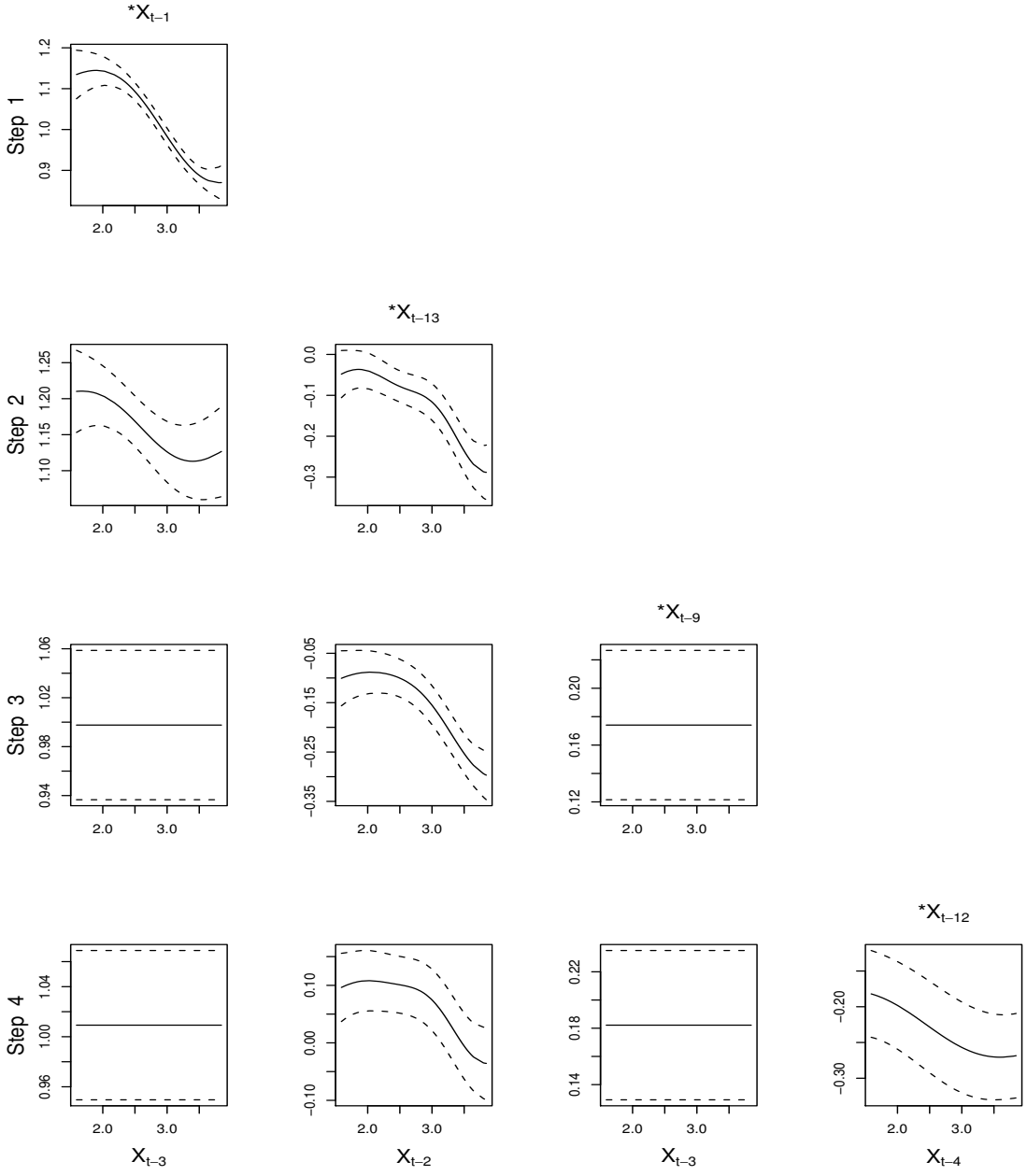


Figure 5.3: Plots of forward model selection steps for the Canadian lynx data using AIC; labels at the top of the plots indicate regressor variables and labels at the bottom indicate argument variables.

mean of the process.

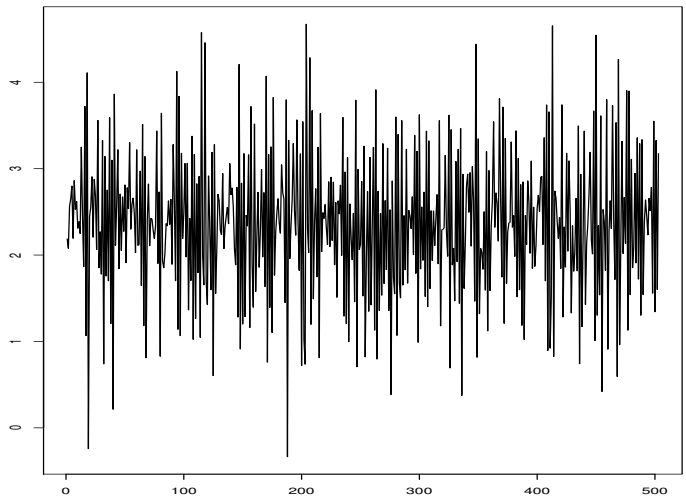


Figure 5.4: Plots of 500 simulated observations from model (5.3).

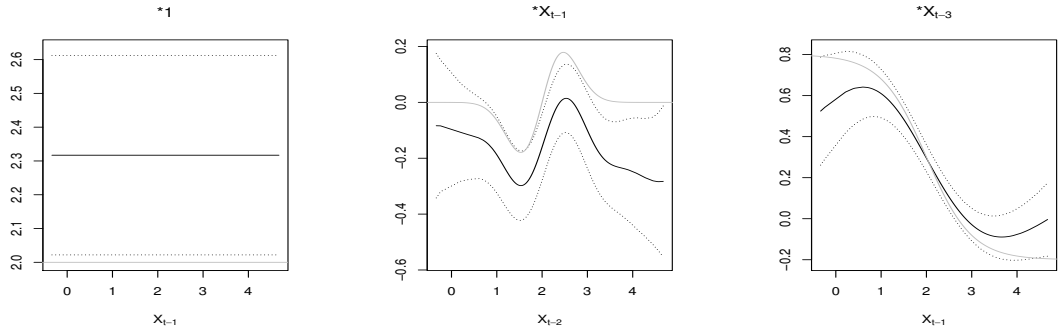


Figure 5.5: Estimated functional coefficients for the simulated data from model (5.3) using BIC selected specification; grey lines represent true functions, labels at the top of the plots indicate regressor variables and labels at the bottom indicate argument variables.

We present the results for the AIC criterion, which selected a four term specification of the form $X_t = f_1 + f_2 X_{t-3} + f_3(X_{t-2})X_{t-3} + f_4(X_{t-2})X_{t-1} + \epsilon_t$. The function estimates, in order of appearance in the previous formula, are shown in Fig. 5.6. Note that the second and third terms can be combined in one term $[f_2 + f_3(X_{t-2})]X_{t-3} = f_{2,3}(X_{t-2})X_{t-3}$, since

both coefficients are multiplied with the same regressor X_{t-3} and one is constant. At the second step of the selection procedure the regressor X_{t-3} was included in the model with a function having X_{t-2} as an argument, which was then set to constant and so became independent of the argument. We can merge the two functions to get the correct specification, which actually gives a lower AIC. We present this example in order to stress that we should always check the results of the selection procedure to identify potential mergers or simplifications of functions. Such redundancies can occur when a wrong combination of regressor and argument is added to the model, whose coefficient becomes constant at later steps. However, this problem is less frequent when we have more data. In fact, when we rerun the selection procedure on a simulated set of 1000 observations from the same model, we get the correct specification under both criteria and confidence intervals which include the true functions.

Finally, we present a simulation experiment from a simple AR model. We generate 500 observations from the linear model

$$X_t = \mu + \alpha_1 X_{t-1} + \alpha_3 X_{t-3} + \epsilon_t \tag{5.4}$$

where $\mu = 2$, $\alpha_1 = .5$, $\alpha_3 = -.3$ and $\epsilon_t \overset{i.i.d}{\sim} \mathcal{N}(0, .5^2)$. We use the same possible regressors and arguments, and the same reduced rank approximation as in the previous example. This time, both AIC and BIC select the correct model specification with constant coefficient functions for all terms. The resulting estimates are shown in Fig. 5.7, together with the true coefficients which are captured by the confidence bands. Our model selection procedure is capable of identifying a wide range of models with additive or multiplicative terms which are constant or varying. Since the procedure can identify linear models, it is not imperative to perform a nonlinearity test before applying it. If the selection algorithm gives a linear specification, we can either keep the fitted model or reestimate it using conventional

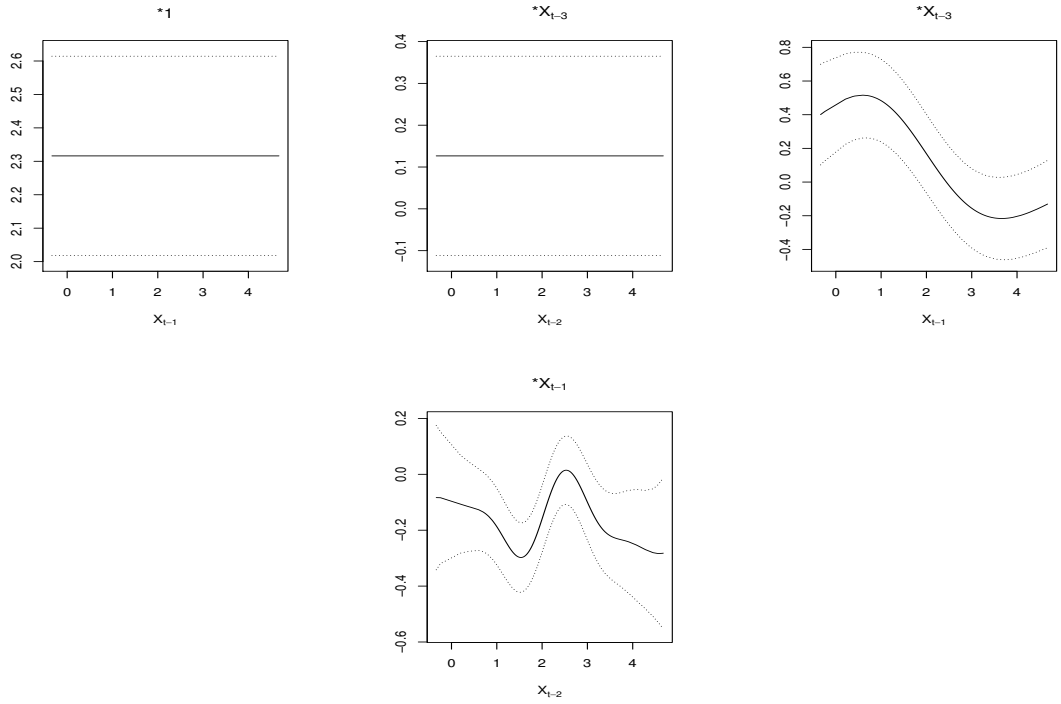


Figure 5.6: Estimated functional coefficients for the simulated data from model (5.3) using AIC selected specification; labels at the top of the plots indicate regressor variables and labels at the bottom indicate argument variables.

frequentist procedures. For large data sets all methods give similar results, but for small data sets it is better to avoid conditioning on initial values, as does our empirical Bayes approach, and use the exact MLEs. Under the approximate inference scheme, our model selection procedure becomes not just feasible but also practical to implement, even for large data sets. This makes it an attractive alternative to other parametric and nonparametric methods.

5.3 Residuals

We now discuss the specification of our model's residuals and their use as a diagnostic tool. The simple fitted residuals $e_t = y_t - \hat{y}_t$ are defined as the difference between the observed

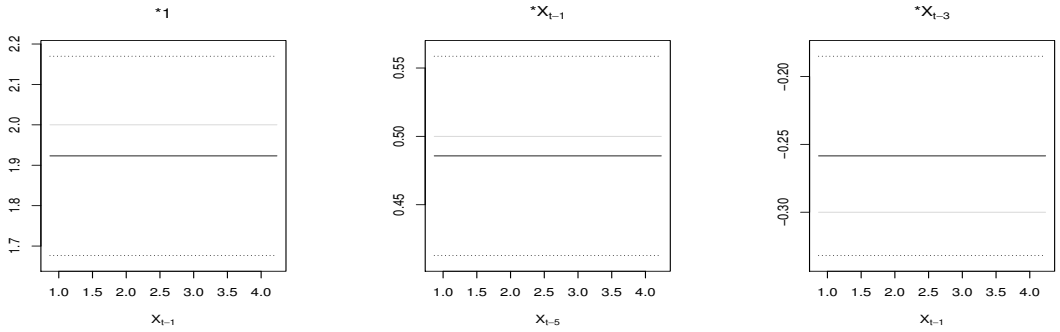


Figure 5.7: Estimated functional coefficients for the simulated data from model (5.4) using either AIC or BIC selected specification; grey lines represent true functions, labels at the top of the plots indicate regressor variables and labels at the bottom indicate argument variables

series and the model's fitted values, given by (5.1). However, these residuals disregard estimation uncertainty, which can be significant since our model assumes the coefficients are random. We can account for this by using the normalized residuals

$$\tilde{e}_t = \frac{y_t - \hat{y}_t}{\sqrt{\sigma^2 + \text{Var}[Z_t|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{u}]}}, \quad t = 1, \ldots, T$$

where the normalization uses the posterior variance of the conditional mean Z_t of the observations. Under the assumption that the model is correct, these residuals should approximate a white noise sequence. Nevertheless, for nonlinear and/or non-Gaussian time series models the use of the universal residuals is more appropriate; e.g. see Smith [108] and Gerlach et al. [40]. These are defined as the quantiles of the recursive one-step-ahead predictive distributions, according to the transformation originally proposed by Rosenblatt [97]. Letting $\boldsymbol{y}_{<t}, \boldsymbol{x}_{<t}, \boldsymbol{u}_{<t}$ denote all the data up to time $t-1$, the universal residuals are given by

$$v_t = P(Y_t \leq y_t | \boldsymbol{y}_{<t}, \boldsymbol{x}_{<t}, \boldsymbol{u}_{<t}), \quad t = 1, \ldots, T$$

For our model, the one-step-ahead predictive distribution is normal with moments that can be computed efficiently using online estimation, for either the exact or the approximate method. Under the correct model for the data, the universal residuals constitute an i.i.d. sample from a uniform distribution and goodness of fit tests can be performed based on this result. The result relies on the assumption that the true parameters of the model are known, but it should hold approximately even if they are to be estimated from the data. For practical purposes, we can equivalently use the inverse normal probability transform $r_t = \Phi^{-1}(v_t)$ of the universal residuals to create a sample of i.i.d. $\mathcal{N}(0,1)$ variables. These are called recursive residuals and are commonly used for conditionally Gaussian models, where r_t can be easily computed in terms of the one-step-ahead predictive moments as

$$r_t = \frac{y_t - \mathrm{E}[Y_t|\boldsymbol{y}_{<t}, \boldsymbol{x}_{<t}, \boldsymbol{u}_{<t}]}{\sqrt{\mathrm{Var}[Y_t|\boldsymbol{y}_{<t}, \boldsymbol{x}_{<t}, \boldsymbol{u}_{<t}]}}, \quad t = 1, \ldots, T$$

In particular, this is the standard definition of residuals for Kalman filtering and more general state-space models, e.g. see section 5.4 in Harvey [54] or Frühwirth-Schnatter [39]. For the remainder we focus on the use of the recursive residuals.

As a preliminary graphical method for assessing fit, we always plot the residuals of a model versus time, as well as versus the different regressor and argument variables. This helps to identify outliers and systematic patterns which can suggest improvements to the model. Furthermore, we use the recursive residuals to perform various quantitative tests for the model's fit. Essentially, all diagnostic tests which are useful for linear time series are also applicable in our case. There are various assumptions we can look at: for normality we can use the Jarque-Bera [63] or Shapiro-Wilk [104] tests, for uncorrelatedness we can use the portmanteau test of Ljung and Box [77] and for independence we can use the test of McLeod and Li [82] on the square residuals. We can also apply general goodness of fit tests or Q-Q plots against the exact $\mathcal{N}(0,1)$ distribution assumption of the residuals.

When working with time series, it is important to ensure that a model remains valid across time and the recursive residuals also serve in detecting structural changes through the use of CUSUM procedures. The basic CUSUM test relies on the rescaled cumulative sum process of the residuals $W_t = \sum_{s=1}^{t} r_s/\sqrt{T}$ for $t = 1, \ldots, T$, which approximates a Brownian motion under the assumption that $r_t \overset{iid}{\sim} \mathcal{N}(0, 1)$. In order to identify a systematic shift in the mean of the series, Brown, Durbin and Evans [15] suggest plotting the process and checking whether it crosses certain linear bounds. The linear bounds are derived from the crossing probabilities of Brownian motion; for .05 significance level, in particular, the authors suggest slopes of $\pm 1.89/T$ and intercepts of $\pm .94$. Petruccelli and Davies [93] use the same idea to perform nonlinearity tests for TAR model identification. First, they reorder the data according to a potential threshold variable and compute the recursive residuals by fitting a linear model along this ordering, instead of the natural time ordering. They thus construct the CUSUM process, and if its maximum exceeds a certain level the linear model is deemed inadequate, which implies the presence of nonlinearity with respect to the ordering variable. As a refinement, Petruccelli [92] suggests summing the residuals in reverse order (because the highest deviations should appear later) and comparing the maximum of the reverse CUSUM process to linear bounds. Although Petruccelli gives slightly different bounds, we can still use the previous bounds of Brown, Durbin and Evans, since both tests rely on the same Brownian motion asymptotics. We apply the two CUSUM procedures to our model for detecting structural changes through time or excess nonlinearity with respect to lagged variables. If the Petruccelli and Davies test is rejected when the residuals are computed along a variable that already appears as an argument to a nonlinear function, this would suggest that the function is too smooth; if the variable does not appear in the model, it would suggest that it could be included as an argument in a nonlinear term. However, an important limitation of such CUSUM tests is that they can only pick up systematic departures

from the assumptions. For example, they will have small power against situations where the mean of the residuals oscillates between positive and negative values of the same magnitude.

Finally, we use a variation of the CUSUM procedure for detecting heteroskedasticity or more general lack of fit. We look at the cumulative sum of squares (CUSUM-SQ) test, originally proposed in Brown, Durbin and Evans [15]. As the name suggests, the CUSUM-SQ test relies on the cumulative sum of square residuals, $V_t = \left(\sum_{s=1}^{t} r_s^2\right)/\left(\sum_{s=1}^{T} r_s^2\right) - t/T$ for $t = 1, \ldots, T$. Inclán and Tiao [62] further develop the procedure, and show that under the usual assumptions on the recursive residuals, V_t converges to a Brownian bridge. They propose the maximum deviation test statistic $Q = \sqrt{T/2}\max_t |V_t|$ for which they provide simulated quantiles and the theoretical asymptotic distribution

$$P(Q \leq c) = 1 + 2\sum_{k=1}^{\infty}(-1)^k \exp\left\{-2k^2b^2\right\} \tag{5.5}$$

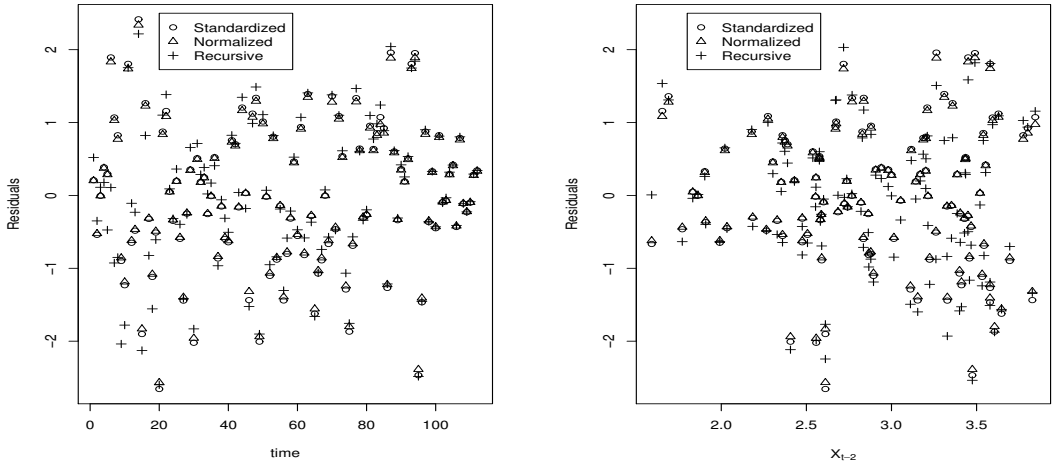
The limiting distribution is the same as for the Kolmogorov-Smirnov test and the .95 quantile, in particular, is 1.358. Similar to the CUSUM procedure, it is instructive to plot the CUSUM-SQ process V_t versus time and lagged variables, in order to identify departures.

We demonstrate the use of these diagnostic tools on the residuals of the second order FAR model (2.30) that we fit to the Canadian lynx data in section 2.6. Fig. 5.8 presents plots of the (standardized) fitted, normalized and recursive residuals versus the time and versus the common coefficient argument X_{t-2}. As we would expect from their definition, the recursive residuals are more diffused in the beginning, when they are based on less data, and converge to the fitted and normalized residuals toward the end. Moreover, the second plot implies that the residuals might increase in variance with X_{t-2}. Table 5.4 gives p-values for normality and whiteness tests applied to the recursive residuals, which do not show significant departures from the assumptions. We also present plots of the CUSUM and

CUSUM-SQ processes in Fig. 5.9 and Fig. 5.10, respectively. Again, we plot the processes against time and X_{t-2}, and we give 95% confidence bands based on their approximating distributions. We do not witness significant departures in any of the processes, although we point out that we have too few observations for the procedures to be powerful.

Table 5.4: P-values of diagnostics tests applied to the recursive residuals of model (2.30) fit to the Canadian lynx data.

Normality		Whiteness	
Jarque-Bera	0.4795	Ljung-Box	0.1142
Shapiro-Wilk	0.5287	McLeod-Li	0.1236

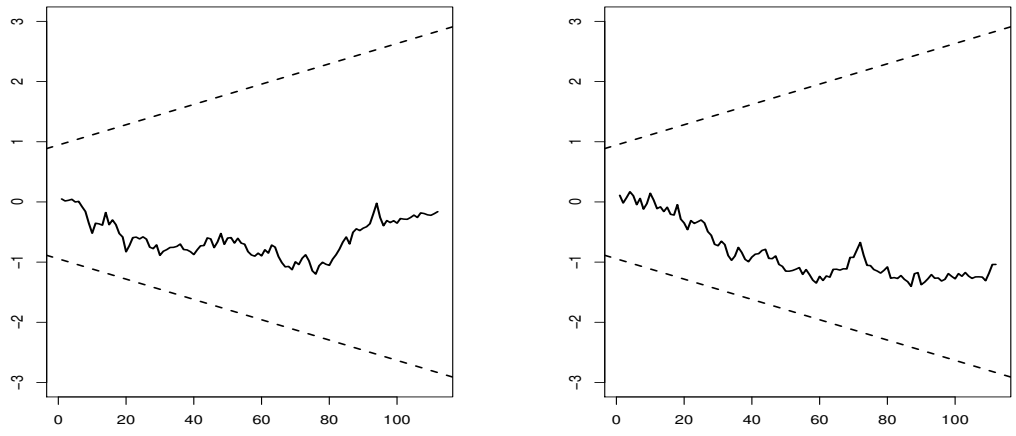


(a) (b)

Figure 5.8: Plots of standardized, normalized and recursive residuals of model (2.30) fit to the Canadian lynx data versus (a) time, and (b) X_{t-2}.

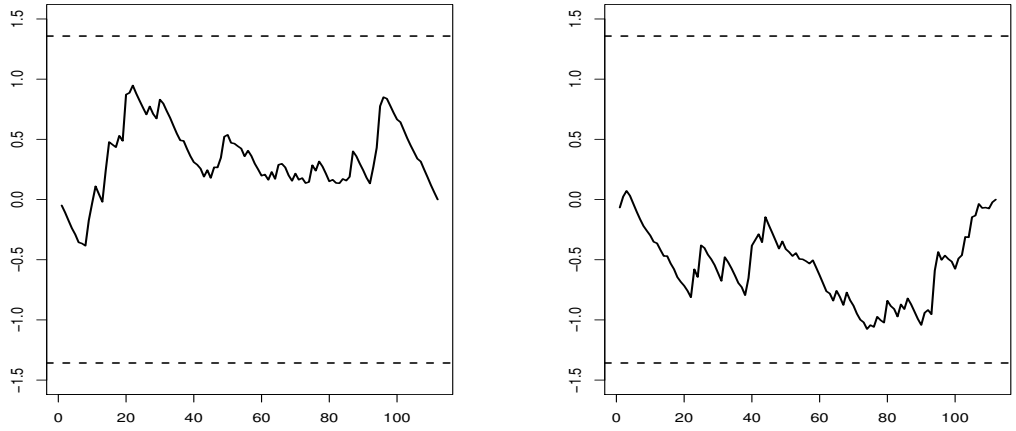
5.4 Dynamics and Stability

The estimates of the coefficient functions define the properties of the fitted model in ways which are not always obvious. When dealing with time series it is important to understand and characterize a model's dynamic behavior and, in particular, its stability. There are vari-



(a) (b)

Figure 5.9: Plots of CUSUM process of model (2.30) fit to the Canadian lynx data, versus (a) time, and (b) X_{t-2}; (- -) 95% confidence bands.



(a) (b)

Figure 5.10: Plots of CUSUM-SQ process of model (2.30) fit to the Canadian lynx data, versus (a) time, and (b) X_{t-2}; (- -) 95% confidence bands.

ous flavors of stability but we will be interested in stationarity and ergodicity, which roughly imply that the process does not change its statistical properties with time. These two prop-

erties afford a sensible use of the model for making future predictions and for applying simulation to study its behavior. Naturally, we only look at models that are self-contained and do not include exogenous variables, which for the main part means Markovian models. Time series regression on exogenous variables describes the cross-sectional relation of the data, so there is not much we can say about the evolution of the process. Specifically, such models cannot be used for forecasting or simulation without additional information about the evolution of the exogenous variables.

For Markov models, we have already presented in Theorem 4.6 conditions on the functional coefficients and the error term which guarantee geometric ergodicity. As we pointed out, these conditions are only sufficient and rather strict, so they are often expected not to hold. As an alternative way of examining stability we look at the deterministic part of the conditional mean equation, what Tong [114] calls the skeleton of the model. This is essentially a nonlinear difference equation, also known as a recurrence relation, which is fundamental for the properties of its stochastic counterpart. For example, the stationarity of a linear Gaussian time series is uniquely determined by the stability of its corresponding homogeneous linear difference equation. However, nonlinear difference equations exhibit extremely diverse and frequently chaotic behavior and there is no systematic way of solving them analytically. Tong suggests plotting trajectories for different plausible starting values from a fitted model's deterministic skeleton as a graphical diagnostic tool. These trajectories can be plotted against time, but it is also instructive to plot them in phase space, i.e. the space in which all possible states of a system are represented, which in our case is the space of all regressor and argument variables. When the phase space has more than three dimensions, we can still plot the trajectories in two dimensional planes of important variables, for example the space of the current and lag-one values. These plots are very informative for the dynamical structure of the nonlinear system and especially for revealing

periodic patterns such as limit cycles. We have already given an example of such a plot in Fig. 2.9 for the Canadian lynx data. If the trajectories of the system are not explosive, this is a good indication that the stochastic model is also well behaved. For this last point, the distribution of the error term plays an important role. There are examples of linear AR(1) models with $\alpha > 1$ and exponential errors which are stationary (see Grunwald et al. [48]) and with $0 < \alpha \leq 1/2$ and Bernoulli errors which are not strong-mixing (see Andrews [2]), where α is the autoregressive coefficient. However, such situations can be avoided when the error term distribution is absolutely continuous with respect Lebesgue measure on $\mathbb{R}$, and this is what we typically require for our model as well.

An alternative way of examining the dynamic properties of a model is through impulse response functions. This approach is common in control theory and econometrics, and is especially usefully when dealing with multivariate data. The impulse response function quantifies the effect of a shock on the evolution of a system. Formally, the impulse response function of a Markov model $Y_t = F(\boldsymbol{U}_t)\boldsymbol{X}_t + \epsilon_t$ is defined as

$$\begin{aligned} IRF(n;\delta,\mathcal{F}_{t-1}) &= \mathrm{E}[Y_{t+n}|\epsilon_t = \delta, \epsilon_{t+1} = 0, \ldots, \epsilon_{t+n} = 0, \mathcal{F}_{t-1}] - \\ &\qquad \mathrm{E}[Y_{t+n}|\epsilon_t = 0, \epsilon_{t+1} = 0, \ldots, \epsilon_{t+n} = 0, \mathcal{F}_{t-1}]; \quad n = 1, 2, \ldots \end{aligned}$$

where δ is the shock size and $\mathcal{F}_{t-1}$ is the filtration of all the information up to time $t-1$. For nonlinear models, Koop et al. [70] suggest using the generalized impulse response function, defined as

$$GIRF(n;\delta,\mathcal{F}_{t-1}) = \mathrm{E}[Y_{t+n}|\epsilon_t = \delta, \mathcal{F}_{t-1}] - \mathrm{E}[Y_{t+n}|\mathcal{F}_{t-1}]; \quad n = 1, 2, \ldots$$

where the effect of the future errors is averaged out instead of fixed to zero, and the conditional means can be calculated with Monte Carlo methods. The two definitions are

equivalent for linear Gaussian models, but for more general models the authors claim the latter approach is better in treating the future. In either case, plots of the impulse response function demonstrate different aspects of the model such as shock persistence and asymmetric effects, and they are a useful diagnostic tool. We give an example of their use in the application of section 6.3.

Finally, we can use simulation for exploring the properties of a fitted model. Similar to the trajectories of the deterministic skeleton, we can create paths from the stochastic version of the model and plot them versus time or in phase space. We also suggest looking at estimates of the autocorrelation function (ACF) and/or the spectral density of the simulated process, and comparing it to the that of the original data. The spectral density summarizes information on second order properties, so it is helpful for describing the autocorrelation structure of a fitted model. Gaussian processes are uniquely determined by their second order properties, but for more general processes this is not true and one can also use higher order periodograms or spectra, see for example Brillinger [13]. Lastly, simulation serves in investigating the probabilistic aspects of the fitted model, by looking at histograms or density estimates of its stationary distribution and/or multivariate distributions of lagged variables. We demonstrate the use of such methods in the application of section 6.1.

Chapter 6

Applications

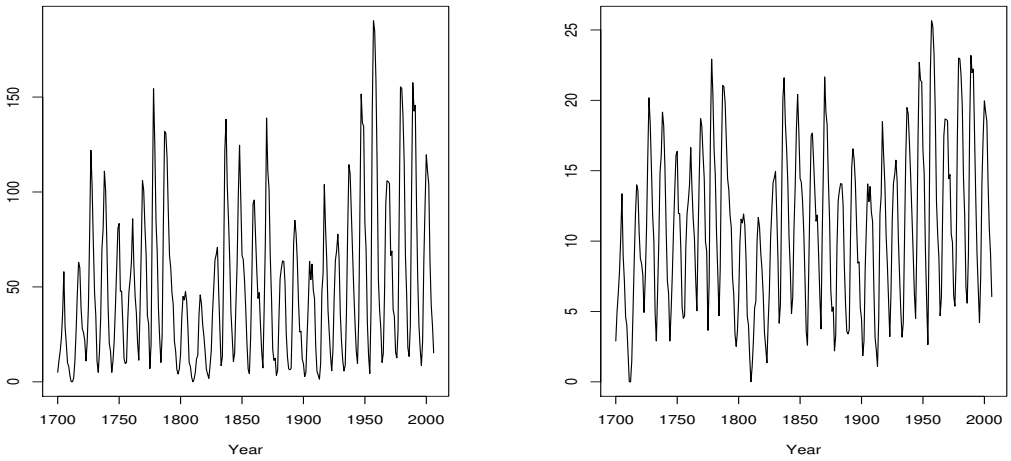
6.1 Wölf's Sunspot Numbers

6.1.1 Introduction and Review

Our first application concerns the classic Wölf sunspot data which, like the Canadian lynx data, has attracted a lot of attention in the nonlinear and nonparametric time series literature. The data consist of an annual measure of solar activity that was devised by Rudolph Wölf in 1848, and which is based on the number of spots on the face of the sun. Even though much more sophisticated measures exist today, they are still of value because no other index of the sun's activity reaches into the past as far and as continuously. There are, however, some issues about the quality and consistency of the data through time, because earlier measurements were based on observations from the Earth and were to a certain extent subjective. The particular data set we work with spans the years 1700 to 2006, giving a total of 307 observations, and is available online at the website of the National Geophysical Data Center [1]. The sunspot numbers are shown in Fig. 6.1(a), and they exhibit an approximate 11.1 year cycle which is asymmetrical. It takes around 4.8 years to rise from

[1] `ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA/SUNSPOT_NUMBERS/YEARLY.PLT`

a minimum to a maximum and another 6.2 years for the opposite.



(a) Original (b) Transformed

Figure 6.1: Wölf's annual sunspot numbers, 1700 to 2006; (a) original and (b) square root transformed data.

An early nonlinear analysis of these data was carried out by Tong [114], followed by non-parametric analyses by Chen and Tsay [22] and Cai, Fan and Yao [16]. In all of these, the square root transform $2(\sqrt{1+x}-1)$ was applied to stabilize the variance, and the transformed series is presented in Fig. 6.1 (b). Moreover, only the first 280 observations (from 1700 to 1979) were used in these analyses. We give a brief description of the models involved. Tong [114] suggests the following two regime TAR model

$$X_t = \begin{cases} 1.92 + .84X_{t-1} + .07X_{t-2} - .32X_{t-3} + .15X_{t-4} \\ \quad -.2X_{t-5} + .19X_{t-7} - .27X_{t-8} + .21X_{t-9} \\ \quad +.01X_{t-10} + .09X_{t-11} + \epsilon_t^{(1)}, & \text{if } X_{t-8} \le 11.93 \\ 4.27 + 1.44X_{t-1} - .84X_{t-2} + .06X_{t-3} + \epsilon_t^{(2)}, & \text{if } X_{t-8} > 11.93 \end{cases} \tag{6.1}$$

Chen and Tsay [22] first employ various nonlinearity tests which suggest that the process is

indeed nonlinear. From these tests they identify X_{t-3} as the argument variable and use lags 1, 2 and 8 to fit the model $X_t = f_1(X_{t-3}) + f_2(X_{t-3})X_{t-1} + f_3(X_{t-3})X_{t-2} + f_4(X_{t-3})X_{t-8}$ with arranged local regression (ALR). After visually inspecting the form of the estimated functions, they specify a parametric model which combines a threshold effect with varying coefficients. Specifically, they use conditional least squares to fit the model

$$X_t = \begin{cases} 1.23 + (1.75 - .17|X_{t-3} - 6.6|)X_{t-1} \\ \quad +(-1.28 + .27|X_{t-3} - 6.6|)X_{t-2} + .2X_{t-8} + \epsilon_t^{(1)}, & \text{if } X_{t-3} < 10.3 \\ .92 + .87X_{t-1} + .17X_{t-2} - .24X_{t-3} \\ \quad +.06X_{t-6} + .04X_{t-8} + \epsilon_t^{(2)}, & \text{if } X_{t-3} \geq 10.3 \end{cases} \tag{6.2}$$

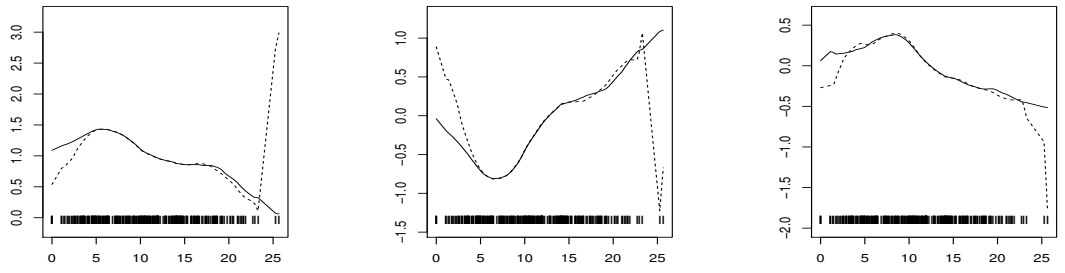
Cai, Fan and Yao [16] apply local linear regression (LLR) for nonparametric estimation of the functional coefficients. They use multifold cross validation to select both the bandwidth and the model specification, i.e. the argument and regressor variables. For the weighting scheme, they use the Epanechnikov kernel $K_h(u) = (1 - (u/h)^2)_+$, and their selection procedure looks through all models with argument lags and autoregressive order from one to 11. They pick X_{t-3} as the argument and X_{t-1} to X_{t-8} as regressors, but they combine their results with the previous model of Chen and Tsay to reduce the fitted model to

$$\begin{aligned} X_t &= f_1(X_{t-3})X_{t-1} + f_2(X_{t-3})X_{t-2} + f_3(X_{t-3})X_{t-3} \\ &\quad + f_4(X_{t-3})X_{t-6} + f_5(X_{t-3})X_{t-8} + \epsilon_t \end{aligned} \tag{6.3}$$

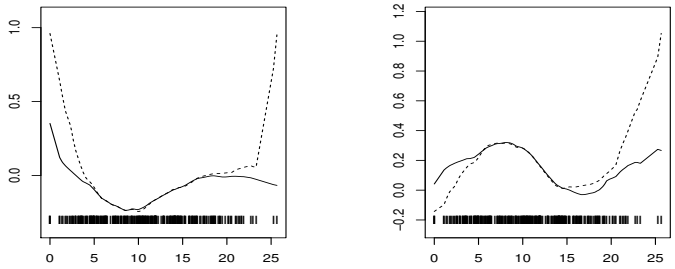
For this specification, the optimal bandwidth is $h = 4.75$, and the estimated functional coefficients are presented in Fig. 6.2 with dashed lines. The estimates exhibit an obvious instability near the endpoints of the argument's observed range, especially to the right. The reason is that very little data falls within the kernel's support, and the resulting local linear regression is ill conditioned. At each point in the argument space, we need to fit a 10

dimensional linear system, since each of the five coefficient functions contributes a constant and a linear term from its Taylor expansion. To correct this instability, we make a slight modification of the kernel. We use $K_h(u) = .01 + (1 - (u/h)^2)_+$, so that every local linear system is stable by assigning a minimal weight to all observations. This correction is also important for extrapolating the function, a need that arises in simulation. The Epanechnikov kernel has compact support, so any evaluation beyond a bandwidth's length from the observed range is undefined. In contrast, our adjustment offers stable predictions in this case, through equally weighting all of the data. The estimated functions under our modified kernel are presented in Fig. 6.2 with solid lines. Where the main body of data lies, the two estimates are practically indistinguishable. However, the modified kernel gives more stable results at the endpoints, were it converges to the global linear estimates of the functional coefficients. For the rest of the example, we use the modified kernel for LLR.

For completeness, we also look at the regression splines method of Huang and Shen [60]. We apply their suggested procedure for selecting the number of knots and the model specification using AIC. First, we fix a number of knots and argument lag for all coefficients, and then create a sequence of models by stepwise addition and deletion of regressor lags. In more detail, we begin with an empty model and start adding terms from a regressor candidate set in the model by minimizing the MSE in a stepwise fashion. When all possible terms have been added, we start a backward procedure of stepwise deletion of terms till we reach the empty model, again using MSE. In the end, we are left with a sequence of models from which we select the one with the smallest AIC. We repeat this procedure for all possible numbers of knots and argument lags, and the final model is the one that minimizes the overall AIC. For this application, we use cubic B-splines with quantile spaced internal knots. The candidate set for the number of internal knots is $\{1, \ldots, 6\}$ and the candidate set for both argument and regressor lags is $\{1, \ldots, 11\}$. The procedure selects a model with



(a) f_1 (b) f_2 (c) f_3



(d) f_4 (e) f_5

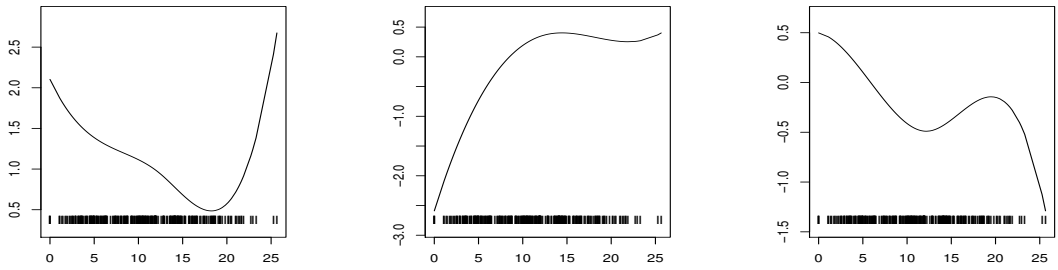
Figure 6.2: Functional coefficient estimates for model (6.3) using LLR, (—) modified kernel, (- -) Epanechnikov kernel.

one internal knot X_{t-2} as the argument and five regressor variables

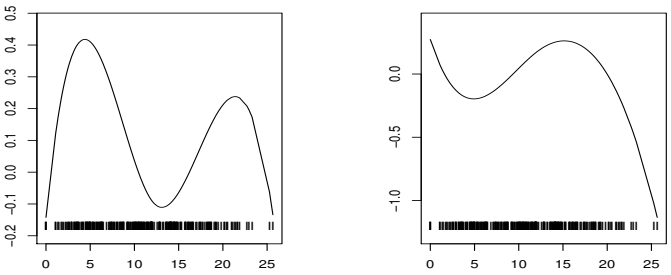
$$\begin{aligned} X_t &= f_1(X_{t-2})X_{t-1} + f_2(X_{t-2})X_{t-2} + f_3(X_{t-2})X_{t-3} \\ &\quad + f_4(X_{t-2})X_{t-9} + f_5(X_{t-2})X_{t-11} + \epsilon_t \end{aligned} \tag{6.4}$$

The estimated functions using splines are presented in Fig. 6.3. Finally, we also provide the simple linear AR model selected by AIC

$$\begin{aligned} X_t &= 10.79 + 1.22X_{t-1} - .48X_{t-2} - .16X_{t-3} + .28X_{t-4} - .25X_{t-5} \\ &\quad + .02X_{t-6} + .17X_{t-7} - .22X_{t-8} + .3X_{t-9} + \epsilon_t \end{aligned} \tag{6.5}$$



(a) f_1 (b) f_2 (c) f_3



(d) f_4 (e) f_5

Figure 6.3: Functional coefficient estimates for model (6.4) using splines.

6.1.2 Estimation Using GPs

We apply our GP methodology to the sunspot data using the first 280 observations, in accordance with the other models. We choose the model specification with our forward selection algorithm, also testing for constant coefficients. The candidate set of argument and regressor variables contains lags one to 11, and we also include a unit regressor variable to allow for additive terms. To speed up model selection, we use a PP reduced rank approximation with 10 bases. However, the resulting model we present is estimated with the exact method. Both AIC and BIC give similar specifications with three terms, where the BIC model is simpler, having two constant coefficients. Nevertheless, we prefer the AIC model because it gives substantially better residual behavior. Our selected model

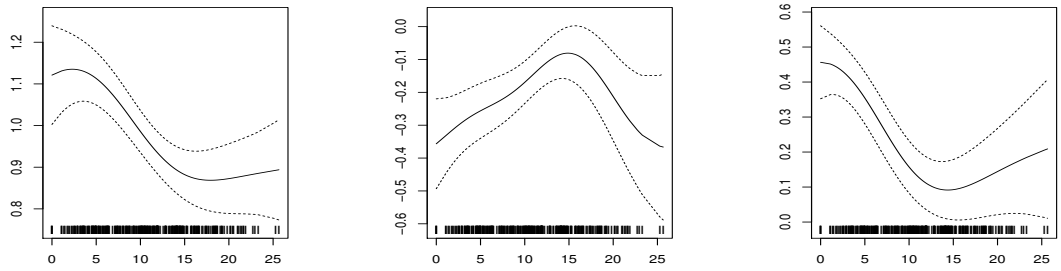
specification is

$$X_t = f_1(X_{t-3})X_{t-1} + f_2(X_{t-1})X_{t-5} + f_3(X_{t-2})X_{t-7} + \epsilon_t \tag{6.6}$$

and the optimal parameters, i.e. the error variance σ and each coefficient's prior mean μ and characteristic lengthscale h, are given in Table 6.1. The estimated functional coefficients for model (6.6) are presented in Fig. 6.4. Compared to the previous nonparametric estimates, our estimates are less variable, taking on a smaller range of values. Moreover, our model has different arguments for the functional coefficients and fewer terms.

Table 6.1: Selected parameters for model (6.6) fitted to the sunspot data.

		f_1		f_2		f_3	
σ	1.906	μ_1	0.972	μ_2	0.261	μ_3	-0.282
		h_1	11.051	h_2	9.818	h_3	7.985
ℓ = -582.0164, AIC = 1194.188							



(a) f_1 (b) f_2 (c) f_3

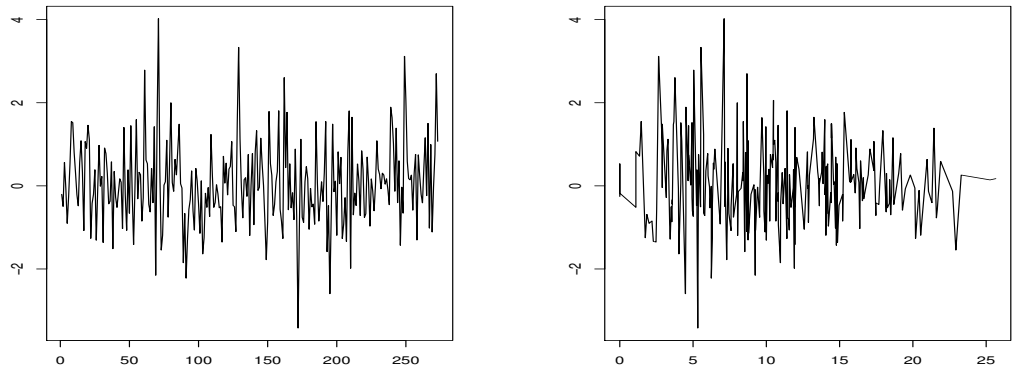
Figure 6.4: Functional coefficient estimates of model (6.6) using GP regression.

We look at the model residuals and present several diagnostic checks. The recursive residuals are plotted in Fig. 6.5 versus time and X_{t-1}. There are some obvious outliers in the data with values beyond the (-2,2) band, and they occur for lower values of X_{t-1}, suggest-

ing there is greater variability when the series moves at the minimum of its cycle. We also present the p-values for several diagnostic tests on the recursive residuals in Table 6.2. The normality assumption is strongly rejected in the presence of the outliers, but there is not significant evidence of residual correlation. We also plot in Fig. 6.6 the CUSUM processes versus time and X_{t-1}, which do not exhibit departures from their hypothetical bounds. We do the same for the CUSUM-SQ processes in Fig. 6.7, and here there is clear indication that the squared residuals are systematically higher for lower levels of X_{t-1}. This heteroskedastic pattern appears for the other models' residuals as well, so a possible way to address it would be to apply a different transformation to the data. Finally, we use simulation to assess the stability and dynamics of the model. The conditions of Theorem 4.6 on the bounds of the functional coefficient are not satisfied, because the bound on f_1 is greater than one. This does not mean that the model is explosive, though, since the conditions are only sufficient. We check for stability by simulating paths from the fitted FAR model, treating the posterior mean functions as our fixed functional coefficient estimates and using normal errors with the selected variance. We generate 1000 paths of 1000 observations each, and none of them is explosive, suggesting that the fitted model is stationary. In Fig. 6.8 we present estimates of the ACF and the spectral density of one of the simulated paths and the sunspot numbers, which we use for comparing their second order properties. Even though our model has only three terms, its ACF seems to follow that of the data for lags up to 30. Moreover, our model's periodic behavior is in accordance to the 11 year cycle of the sunspot data, confirmed by its spectral density's peak around the .09 frequency.

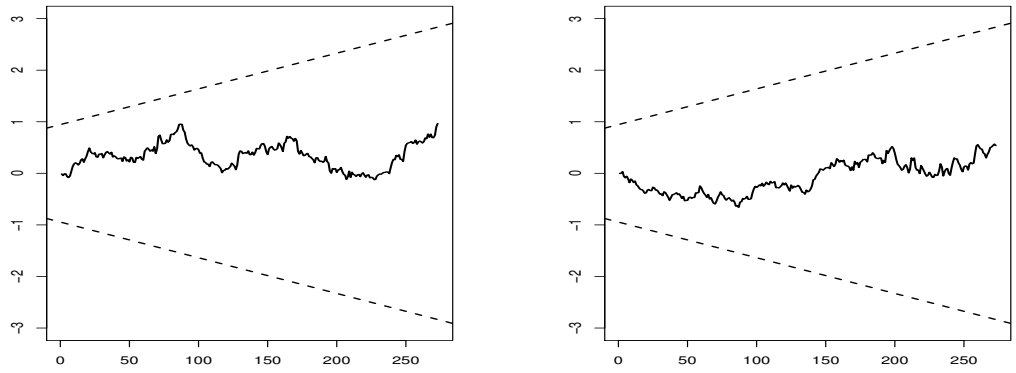
Table 6.2: P-values of diagnostics tests applied to the recursive residuals of model (6.6).

Normality		Whiteness	
Jarque-Bera	8.47×10^{-7}	Ljung-Box	0.14800
Shapiro-Wilk	0.00165	McLeod-Li	0.08425



(a) (b)

Figure 6.5: Recursive residuals of model (6.6) versus (a) time, (b) X_{t-1}.

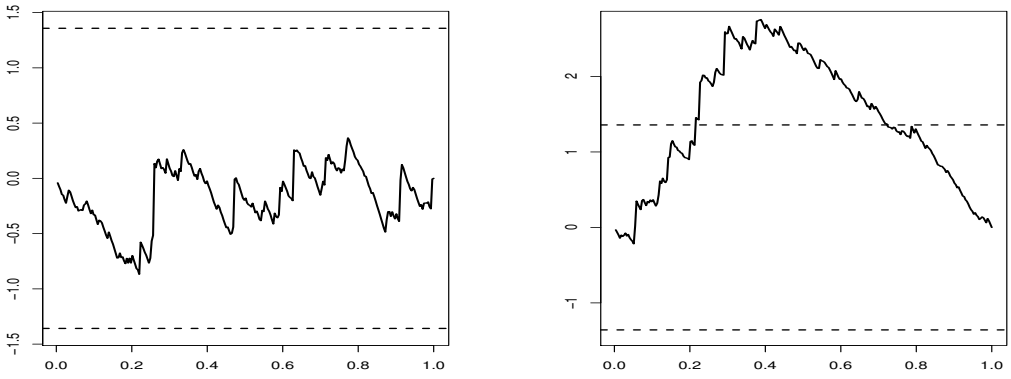


(a) (b)

Figure 6.6: Plots of CUSUM process of model (6.6) versus (a) time, and (b) X_{t-1}; (- -) 95% confidence bands.

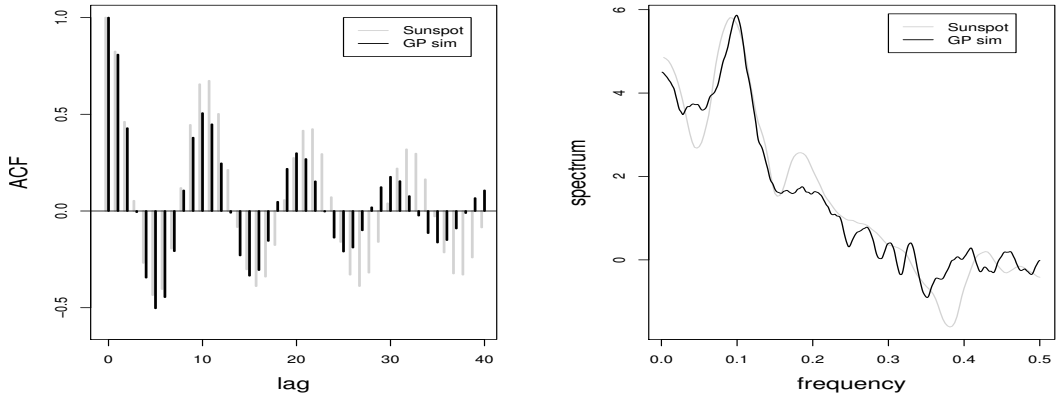
6.1.3 Model Comparisons

The six alternative models we have presented so far appear very different. They have a wide range of argument/threshold variables, regressor variables and functional coefficient forms. However, as can be seen from Fig. 6.9, in terms of fit all models are quite close, the biggest differences arise at the peaks and troughs of the series. To better distinguish their



(a) (b)

Figure 6.7: Plots of CUSUM-SQ process of model (6.6) versus (a) time, and (b) X_{t-1}; (- -) 95% confidence bands.



(a) (b)

Figure 6.8: Simulation diagnostic plots for model (6.6): (a) acf, and (b) spectral density.

dynamics, we present in Fig. 6.10 the skeleton plot of each one in (X_t, X_{t-1})-space. The initial state is that of the data in 1979, and the length of the trajectories is 200 years. All models exhibit cyclical behavior, but only the TAR, LLR and our GP model have sustained cycles, where the TAR model seems to have a smaller circumference. The AR and spline models converge to a stable point, and the ALR model has a strange type of periodic be-

havior with irregular cycles. Of course the behavior depends on the initial state, but trials with other reasonable starting points give similar results.

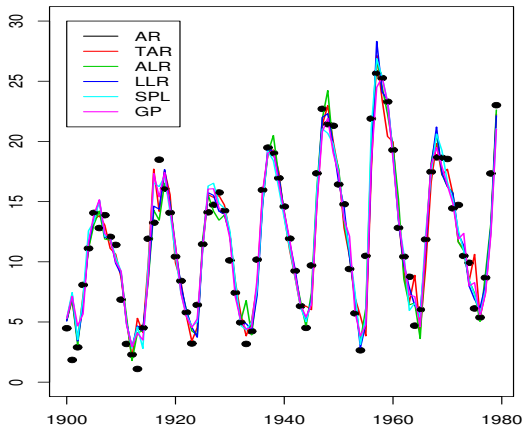
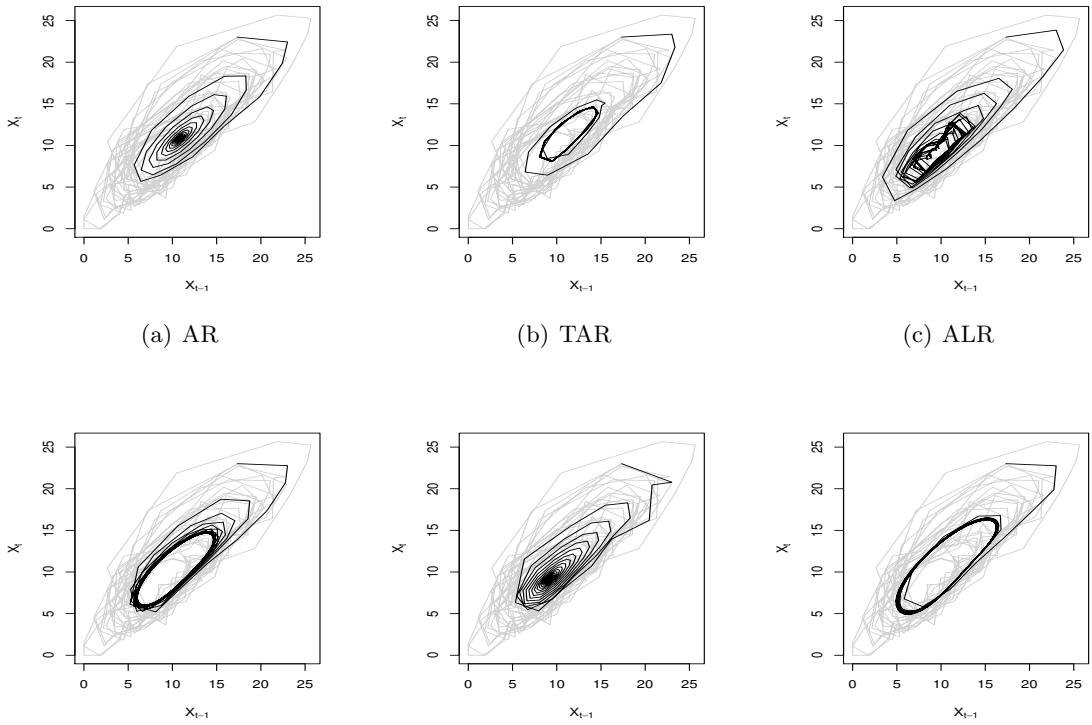


Figure 6.9: Fitted values from all six models, 1900 to 1979; dots represent true values.

Next, we test the predictive performance for each model. We refit each of the six models to the first 200 data points, and use the results to make iterative predictions for the remaining 80. We look at one- to 25-step-ahead predictions on a rolling basis, i.e. starting from X_{200} we predict X_{201} to X_{225}, starting form X_{201} we predict X_{202} to X_{226} and so on. When refitting the models to the initial 200 observations, we just reestimate the coefficient functions. In particular, the model specification and the hyperparameters or threshold values are unchanged. We calculate the mean absolute prediction error (MAPE) for each model and lead time, which are presented in Fig. 6.11. Some characteristics are similar across models, they are all very close in one-step-ahead predictions and they seem to deteriorate after the passing of a period, i.e. at lead times 11 and 22. For one- to 11-step-ahead predictions, there is no model that clearly dominates, with GP, ALR and splines giving lower MAPE at different times. Moreover, the simple AR model provides good prediction for



(a) AR (b) TAR (c) ALR

(d) LLR (e) Splines (f) GP

Figure 6.10: Plots of phase space trajectories for the dynamics of the six models fitted to the sunspot data; initial state is that of 1979, gray lines represent the observed trajectory.

this range. For longer term predictions, however, the GP model gives lower errors, with TAR and LLR being second closely together. The AR model deteriorates the most after the first period, probably because its iterative predictions converge faster to its stable point.

Finally, we make predictions for the 27 observations of the period 1980-2006, which were not used in the model fitting. Besides iterative predictions, we also employ simulation to approximate the distribution of future sunspot numbers. We use normal errors with variances given by the RSS (in the TAR and ALR models we use a separate variance for each regime), except for GP where we use the fitted variance parameter. We generate 10000

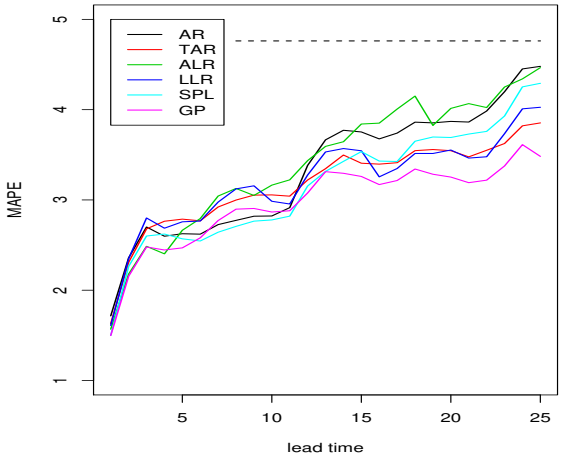
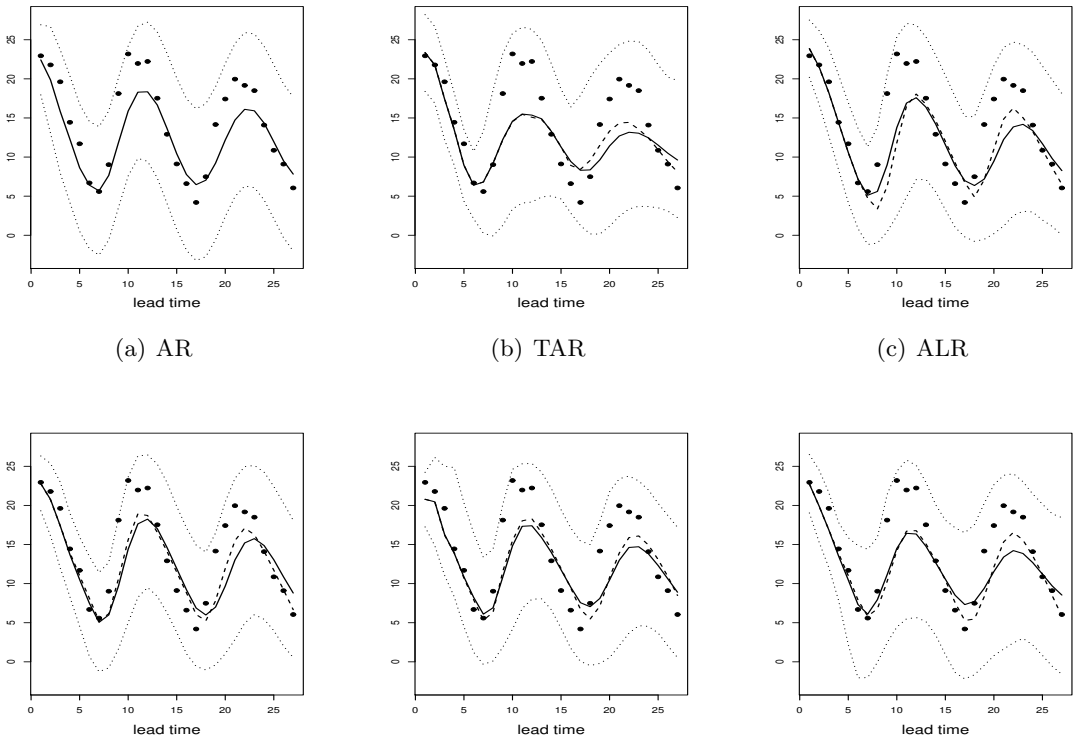


Figure 6.11: Mean absolute error of iterative predictions versus lead time for the observations from 1900 to 1979, where all six models were fit over the period from 1700 to 1899; dashed line represents the mean absolute error from using the mean of the data as the prediction.

paths for each model, and calculate the median and 95% bands of the simulated predictive distributions. These are presented in Fig. 6.12, together with the iterative predictions and the true data points. For the splines model, about 5.8% of the paths were explosive, in the sense that they went beyond 50 in absolute value, and were removed. Thus, the simulated confidence intervals for splines have a downward bias. This problem was caused by large coefficient values outside the observed range of the data, where the splines extrapolate linearly. In the AR model, the iterative predictions are the same as the median of the predictive distribution, which is normal and known explicitly. For all other models, the medians of the distributions seem shrunk toward the series mean, and do a little worse than the iterative predictions. Another common characteristic is that the lower part of the confidence bands seems too wide. In particular, the predictive distributions allow for negative values, which are impossible for the data at hand. Our use of normal errors is a simplification, and a different data transformation, or models which are faithful to this

asymmetry, might be preferable. Overall, all models capture the observations within their confidence bands, but the LLR and AR model give the closest point predictions in terms of absolute error, followed by our model.



(a) AR (b) TAR (c) ALR

(d) LLR (e) Splines (f) GP

Figure 6.12: Monte Carlo predictive distributions for the sunspot data from 1980 to 2006, for all six models; (—) median, $(\cdots)$ 95% confidence bands,(- -) iterative predictions, dots represent true data .

We conclude this example with some remarks. Although the sunspot series exhibits obvious nonlinearity, the simple linear AR model does relatively well for short term predictions. For long-term predictions and for capturing the dynamics of the series, however, there is a need for nonlinear models. Our method provides a parsimonious specification, with only three terms and simple forms for the estimated functional coefficients. Additionally, our

model seems to perform slightly better than the alternatives in describing the dynamics and for making long term predictions. In terms of fit and of stochastic behavior, there is room for improvement, not least because the data are non-negative but are treated as real. Nevertheless, our model gives good results and avoids some of the pitfalls of the other nonparametric methods. Specifically, our method is well suited for extrapolating the coefficients and for giving stable models.

6.2 Nonlinear Vector Error Correction Model

6.2.1 Introduction and Data Description

In this section we give an example of our methodology in a multivariate setting. Our goal is to model a bivariate series of a stock index and a futures contract on the index. A standard approach is to use a linear vector error correction model (VECM) for representing the dynamics of the series. We briefly describe this model and the closely related concept of cointegration, which were formally introduced in the seminal paper of Engle and Granger [35]. Two time series $\{Y_{1,t}, Y_{2,t}\}$ are called cointegrated of order one, if each of them is integrated of order one (i.e. their first difference $\Delta Y_{\cdot,t} = Y_{\cdot,t} - Y_{\cdot,t-1}$ is stationary) and there exists a linear combination $Z_t = a_1 Y_{1,t} + a_2 Y_{2,t}$ that is stationary. The linear combination $a_1 Y_{1,t} + a_2 Y_{2,t}$ expresses an equilibrium relationship between the two variables. This interpretation follows from the fact that $\{Z_t\}$ will revert to its mean infinitely often due to stationarity, and for this reason the variable Z_t is called the cointegration error term. Letting $\boldsymbol{Y}_t = [Y_{1,t}, Y_{2,t}]^\top$, and under some regularity conditions, the Granger representation theorem states that we can represent the dynamics of $\Delta \boldsymbol{Y}_t$ as

$$\Delta \boldsymbol{Y}_t = \boldsymbol{a}_0 + \sum_{i \geq 1} \boldsymbol{A}_i \Delta \boldsymbol{Y}_{t-i} + \boldsymbol{b} Z_{t-1} + \boldsymbol{\epsilon}_t$$

where the $\{\boldsymbol{A}_i\}_{i \geq 1}$ are 2×2 matrices, $\boldsymbol{a}_0, \boldsymbol{b} \in \mathbb{R}^2$, and $\{\boldsymbol{\epsilon}_t\}$ is a two-dimensional white noise sequence. This is similar to a vector autoregressive model for $\{\Delta \boldsymbol{Y}_t\}$, with the difference that the error correction term $\boldsymbol{b} Z_{t-1}$ forces the two series to revert to equilibrium.

A standard example of cointegrated time series is an asset's spot and futures log-prices. The theoretical price of a futures contract is given by a no arbitrage argument using the cost-of-carry model. This refers to the strategy of buying and holding the asset until the futures contract expires. The absence of arbitrage requires that the current price of the

futures contract is equal to the discounted cost of the strategy. Brenner and Kroner [12] give a more detailed discussion of the cost-of-carry model together with empirical results in support of cointegration. Let S_t denote the asset's price at time t, and $F_{t,\tau}$ denote the the futures price at time t for a contract expiring at some later time τ. The cost-of-carry model we use states that

$$F_{t,\tau} = S_t \exp\{(r_{t,\tau} - q_{t,\tau})(t - \tau)\}$$

where $r_{t,\tau}$, $q_{t,\tau}$ are the risk-free interest and asset dividend rates over the period (t, τ). For real data the relationship is not exact, so we define the mispricing error term as

$$Z_t = \ln F_{t,\tau} - \ln S_t - (r_{t,\tau} - q_{t,\tau})(t - \tau) \tag{6.7}$$

The deviations of Z_t from zero can be attributed to transaction costs, short-selling restrictions and interest rate risks, among others. The implicit assumption leading to the VECM is that $\{Z_t\}$ is stationary. Conceptually, if the error term moves away from zero, arbitrage opportunities will present themselves and market participants will offset the discrepancy by taking advantage of these opportunities. In the simple VECM, the error correcting intensity, as represented by the coefficient $\boldsymbol{b}$ of the error term, is constant. In practice, however, arbitrage positions are entered only when the price discrepancy is significantly large, because of the costs and risks entailed in such positions. This observation suggests that the error correcting intensity should not be constant, in particular it should be weaker when the error term is close to zero and stronger when it is far from zero. We try to capture this type of behavior with a nonlinear VECM.

The data for this application come from Martens, Kofman and Vorst [79], and are available

online at the data archive of the Journal of Applied Econometrics [2]. They consist of the S&P 500 index throughout May and November 1993, and the matching futures prices for contracts maturing in June and December 1993, respectively. Minute-by-minute log-returns were calculated from records of the exchanges these assets are traded in, and the first ten observations within each day are discarded, giving approximately 379 observations per day. This was done because of sparse trading activity in the beginning of the day, and in order to avoid overnight returns. For the error term, the daily US discount rate between banks was used, and the dividend rate was calculated using the daily realized dividends, as reported by Standard and Poors. The resulting time series consist of 7060 observations for May and 7693 observations for November. In Fig. 6.13 we present May's data for all three variables. In what follows, these series are treated as if trading was uninterrupted. By this we mean that, when we condition on the past of an observation, we make no distinction as to whether the previous observations come from the same trading day or from the previous one. This is not the best course of action, but it is what other authors did when analyzing the data. Moreover, this problem arises only when dealing with the first few observations out of around 379 within each trading day, so we hope its effect is limited.

6.2.2 Threshold Vector Error Correction Models

The most common approach in the literature to model nonlinear error correcting behavior is through a threshold VECM (TVECM). The general framework is given by Balke and Fomby [4], but we present the particular estimation procedure of Martens, Kofman and Vorst [79]. Letting $\boldsymbol{Y}_t = [\ln F_{t,\tau}, \ln S_t]^\top$, the TVECM is given by

$$\Delta \boldsymbol{Y}_t = a_0^{(j)} + \sum_{i=1}^{p} \boldsymbol{A}_i^{(j)} \Delta \boldsymbol{Y}_{t-i} + \boldsymbol{b}^{(j)} Z_{t-1} + \boldsymbol{\epsilon}_t^{(j)}; \text{ if } c_{j-1} < Z_{t-d} \leq c_j, \ j = 1, \ldots, J \quad (6.8)$$

[2] http://qed.econ.queensu.ca/jae/1998-v13.3/martens-kofman-vorst/

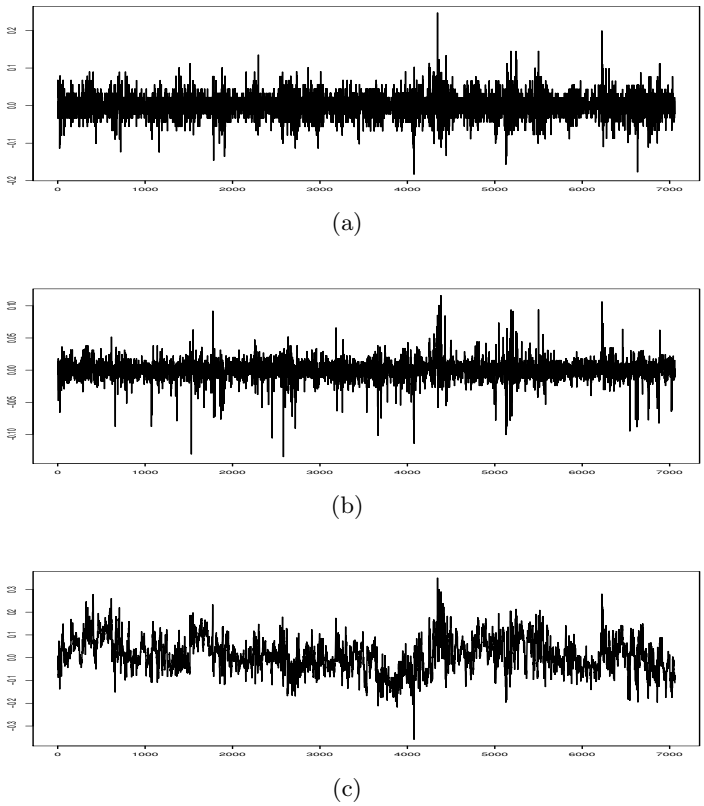


Figure 6.13: Plots of (a) $\Delta \ln F_{t,\tau}$ in % points, (b) $\Delta \ln S_t$ in % points and (c) Z_t, for May 1993.

The TVECM allows for different error correcting behavior in different regimes. Fitting the model involves estimating the threshold lag d for the threshold variable Z_{t-d}, the number of regimes J, the value of the thresholds c_j, and the values of the parameters $\boldsymbol{a}_0^{(j)}$, $\boldsymbol{A}_i^{(j)}$, $\boldsymbol{b}^{(j)}$ and $\boldsymbol{\Sigma}^{(j)} = \mathrm{Cov}[\boldsymbol{\epsilon}_t^{(j)}]$ for each regime. We do not discuss estimation of the cointegrating vector, i.e. the coefficients in the stationary linear combination $Z_t = a_1 \ln F_{t,\tau} + a_2 \ln S_t - (r_{t,\tau} - q_{t,\tau})(t - \tau)$. The authors show that when they are estimated in the linear VECM, they are very close to the theoretical values of $a_1 = 1, a_2 = -1$, so they use the error term as defined in (6.7). Their estimation procedure is roughly divided in two parts, each one involving different data. The first part estimates all parameters related to the thresholds by examining the mispricing series, while the second part selects the linear autoregression

parameters within each regime, using all of the data. In the first part, the authors fit the TAR model

$$Z_t = \phi_0^{(j)} + \sum_{i=1}^{p} \phi_i^{(j)} Z_{t-i} + \epsilon_t^{(j)}; \text{ if } c_{j-1} < Z_{t-d} \leq c_j, \quad j = 1, \ldots, J \tag{6.9}$$

to the error series, using the four step methodology proposed by Tsay [116]. First, they select the AR order p using the PACF and propose the set $S = \{1, 2, 3, 4\}$ for candidate values of the delay parameter d. Then, they fit an arranged autoregression for a given p and every element of S, for each of whose residuals they perform an F-type nonlinearity test. From these tests they keep the delay parameter $d^\star$ with the smallest p-value. The third step is about selecting the number and location of the thresholds. This is a difficult task because possible criteria (likelihood, sum of squared errors) are piecewise discontinuous with respect to the threshold values. A nonparametric arranged autoregression for the chosen $d^\star$ is used to establish a collection of potential thresholds. This is done by visual inspection of the plots of the estimated coefficients, as well as their t-ratios, against the threshold variable $Z_{t-d^\star}$. Potential thresholds are identified by the presence of abrupt changes in these plots. The last step is a refinement step, where the authors consider two competing specifications, a three and a five regime model. For each model, the thresholds are chosen using a grid search among the potential thresholds from part three, by minimizing the aggregate sum of squares across regimes. The final choice is given by a likelihood ratio test in favor of the five regime model. At the second part, the same regime specification is used for the TVECM model, and the parameters are estimated through least squares. The autoregressive order within each regime and for each coordinate of $\Delta \boldsymbol{Y}_t$ is variable. It is decided based on the maximum significant lag at the 10% level, with a minimum lag of one. The constant and error correction coefficients are also included by default, irrespective of their significance level. The fitted models for May and November, using the procedure of Martens, Kofman and Vorst, are presented in Table 6.3 and Table 6.4 respectively.

Table 6.3: TVECM estimates by Martens, Kofman and Vorst [79]; May 1993 data

	Regime 1 $-\infty < Z_{t-1} \leq -0.158$	Regime 2 $-0.158 < Z_{t-1} \leq -0.073$	Regime 3 $-0.073 < Z_{t-1} \leq 0.072$	Regime 4 $0.072 < Z_{t-1} \leq 0.204$	Regime 5 $0.204 < Z_{t-1} < \infty$
Futures equation					
Constant	0.000242	0.000003	0.000001	-0.000049	0.000726
$\Delta \ln F_{t-1,\tau}$	0.109	-0.151[a]	-0.0431	-0.0383	0.318
$\Delta \ln F_{t-2,\tau}$			0.0229	0.0141	
$\Delta \ln F_{t-3,\tau}$			0.0389[b]	-0.00349	
$\Delta \ln F_{t-4,\tau}$			0.0258[c]	-0.0571[c]	
$\Delta \ln S_{t-1}$	0.203	-0.0768	-0.0233	0.0159	0.064
$\Delta \ln S_{t-2}$				-0.137[b]	
Z_{t-1}	-0.00193	0.000281	-0.000129	0.000728[b]	-0.00473
Adj. R^2	-0.04	0.023	0.003	0.007	0.028
Index equation					
Constant	0.000165	0.00106[a]	0.000001	-0.000084[a]	0.000201
$\Delta \ln F_{t-1,\tau}$	0.0287	0.0856[a]	0.041[a]	0.04[a]	0.0786
$\Delta \ln F_{t-2,\tau}$		0.23[a]	0.101[a]	0.117[a]	
$\Delta \ln F_{t-3,\tau}$		0.164[a]	0.104[a]	0.0883[a]	
$\Delta \ln F_{t-4,\tau}$		0.144[a]	0.0744[a]	0.0917[a]	
$\Delta \ln F_{t-5,\tau}$		0.141[a]	0.0587[a]	0.0474[a]	
$\Delta \ln F_{t-6,\tau}$		0.134[a]	0.0422[a]	0.0108	
$\Delta \ln F_{t-7,\tau}$		0.0714[a]	0.0484[a]	0.0328[a]	
$\Delta \ln F_{t-8,\tau}$		0.0597[a]	0.0253[a]	0.0095	
$\Delta \ln F_{t-9,\tau}$		0.0471[b]	0.0277[a]	0.0287[b]	
$\Delta \ln F_{t-10,\tau}$		0.0408[c]	0.0187[a]		
$\Delta \ln F_{t-11,\tau}$		0.0391[c]	0.0111[c]		
$\Delta \ln F_{t-12,\tau}$		0.0573[a]			
$\Delta \ln S_{t-1}$	0.551[a]	-0.0308	-0.0508[a]	0.0738[a]	0.426[b]
$\Delta \ln S_{t-2}$	0.819[a]	0.0493			
$\Delta \ln S_{t-3}$		-0.134[a]			
Z_{t-1}	0.00215	0.00106[a]	0.000138[a]	0.000999[a]	0.00176
Adj. R^2	0.424	0.263	0.120	0.208	0.181
Observations	45	772	4809	1391	36

[a] Significant at 1% level
[b] Significant at 5% level
[c] Significant at 10% level

Table 6.4: TVECM estimates by Martens, Kofman and Vorst [79]; November 1993 data

	Regime 1 $-\infty < Z_{t-1} \leq -0.186$	Regime 2 $-0.186 < Z_{t-1} \leq -0.09$	Regime 3 $-0.09 < Z_{t-1} \leq 0.062$	Regime 4 $0.062 < Z_{t-1} \leq 0.212$	Regime 5 $0.212 < Z_{t-1} < \infty$
Futures equation					
Constant	0.000162	0.000068	0.000007^{c}	0.000039^{b}	0.000283
$\Delta \ln F_{t-1,\tau}$	-0.455^{a}	-0.256^{a}	0.00406	-0.0393^{c}	0.0558
$\Delta \ln F_{t-2,\tau}$		0.107		0.0169	
$\Delta \ln F_{t-3,\tau}$		-0.142^{b}		0.0154	
$\Delta \ln F_{t-4,\tau}$				0.0519^{a}	
$\Delta \ln S_{t-1}$	0.781	0.0382	0.118^{a}	0.0889^{b}	-0.144
$\Delta \ln S_{t-2}$		0.315	0.0104		
$\Delta \ln S_{t-3}$			0.0676^{b}		
$\Delta \ln S_{t-4}$			0.0865^{a}		
Z_{t-1}	0.00048	0.000712	-0.000209^{c}	-0.000473^{a}	-0.00132
Adj. R^2	0.185	0.119	0.008	0.008	0.002
Index equation					
Constant	-0.000515^{b}	0.000079	-0.00012^{a}	-0.000034^{a}	0.000028
$\Delta \ln F_{t-1,\tau}$	0.157^{c}	0.0812^{b}	0.0473^{a}	0.0305^{a}	0.0926
$\Delta \ln F_{t-2,\tau}$		0.121^{a}	0.0514^{a}	0.0698^{a}	
$\Delta \ln F_{t-3,\tau}$		0.123^{a}	0.0561^{a}	0.0664^{a}	
$\Delta \ln F_{t-4,\tau}$			0.0533^{a}	0.0423^{a}	
$\Delta \ln F_{t-5,\tau}$			0.047^{a}	0.0344^{a}	
$\Delta \ln F_{t-6,\tau}$			0.0204^{a}	0.0224^{b}	
$\Delta \ln F_{t-7,\tau}$			0.0347^{a}		
$\Delta \ln F_{t-8,\tau}$			0.0307^{a}		
$\Delta \ln F_{t-9,\tau}$			0.0126^{c}		
$\Delta \ln F_{t-10,\tau}$			0.0232^{a}		
$\Delta \ln F_{t-11,\tau}$			0.0178^{a}		
$\Delta \ln F_{t-12,\tau}$			0.0152^{b}		
$\Delta \ln F_{t-13,\tau}$			0.0137^{b}		
$\Delta \ln S_{t-1}$	0.637	-0.0108	0.0221	0.083^{a}	0.488^{a}
$\Delta \ln S_{t-2}$		0.281^{a}	0.0303^{b}	0.0743^{a}	
$\Delta \ln S_{t-3}$		0.181^{b}	0.0453^{a}	0.0437^{b}	
$\Delta \ln S_{t-4}$		0.156^{c}	0.0278^{c}	0.0438^{b}	
Z_{t-1}	-0.00194	0.00115^{b}	0.000386^{a}	0.000441^{a}	0.000251
Adj. R^2	0.314	0.238	0.119	0.137	0.172
Observations	45	772	4809	1391	36

[a] Significant at 1% level
[b] Significant at 5% level
[c] Significant at 10% level

There are at least two more approaches in the literature for modeling the same data set. They are both based on the TVECM and they both try to address estimation in an operationally more convenient and integrated way. The first approach is Bayesian, and was proposed by Forbes, Kalb and Kofman [37]. Initially, the authors fix some aspects of the model specification. Specifically, they decide to work with three regimes and an autoregressive order of eight for both index and futures log returns within each regime. The remaining parameters of the threshold lag d, threshold values $\boldsymbol{c} = (c_1, c_2)$, and the linear autoregression coefficients within each regime are given prior distributions. The parameter d is assigned a uniform prior over a possible range S, and the threshold vector $\boldsymbol{c}$ is assigned a bivariate normal, truncated so that it satisfies $c_1 < c_2$. The parameters within each regime follow a non-informative prior, the same as for the treatment of the linear VAR model given by Zellner [123]. In order to sample from the posterior distribution, the authors provide the marginal posterior of $(d, \boldsymbol{c})$ given the data, and the conditional posteriors of the parameters within each regime given $(d, \boldsymbol{c})$ and the data. The latter are analogous to the linear model, with coefficients following normal and covariance matrices following inverse Wishart distributions, but the former is a non-standard distribution. The authors propose a numerical integration scheme for normalizing $p(d, \boldsymbol{c}|Data)$, and then sample $(d, \boldsymbol{c})$ from the resulting discrete approximation to the distribution. Finally, using a Monte Carlo sampling scheme, they construct Rao-Blackwellized estimates of the parameters within each regime. The prior for $\boldsymbol{c}$ is the only informative prior in the model, where the means and variances of the truncated normals are specified based on knowledge of market behavior. Table 6.5 is reproduced from Forbes, Kalb and Kofman [37], and it presents the means and standard deviations of the Monte Carlo sample for the linear coefficients within each regime. Moreover, the authors note that the posterior of d overwhelmingly supports a threshold lag $d = 1$, and that the threshold pair $(c_1, c_2) = (-0.01039, 0.1278)$ holds a marginal posterior posterior probability of 0.727. So, for interpretation purposes, it can be assumed that the

thresholds are defined by these particular values.

The second approach is frequentist and was proposed by Tsay [117]. It is actually an extension of his technique for fitting univariate threshold models to a multivariate setting. First, he proposes a nonlinearity test, based on arranged autoregression. Then, he advocates the use of conditional least squares estimates for estimation. Specifically, he assumes the number of regimes and the order within each regime are known, so that the conditional sum of squares depends only on the threshold delay d, the threshold values $\boldsymbol{c}$, and the coefficients. In support of this approach, he shows that, under certain conditions, the estimators that maximize the conditional sum of squares are strongly consistent. For selecting all the remaining parameters and the model specification, he suggests using AIC. Tsay then applies his method to our current data, after replacing 10 extreme observations by the average of their two nearest neighbors, in order to reduce the influence of outliers. Then, he fixes $p = 8$ and $J = 3$, and allows $d \in \{1, 2, 3, 4\}$, and c_1 and c_2 to assume discrete values in the two intervals $[-0.115, -0.2]$ and $[0.025, 0.145]$, respectively. The later are chosen based on the empirical range of Z_t and not on the plots of nonparametric coefficient estimates. The author then selects the threshold values c_1, c_2, as well as d, by minimizing AIC using a grid search. The remaining coefficient of the regimes are chosen by conditional least squares and are presented in Table 6.6.

6.2.3 Estimation Using GPs

We apply our nonparametric estimation procedure to the same problem. We fit a bivariate functional coefficient VAR model, given by

$$\Delta \boldsymbol{Y}_t = \boldsymbol{X}_t^\top F(\boldsymbol{U}_t) + \boldsymbol{\epsilon}_t \tag{6.10}$$

Table 6.5: Bayesian TVECM estimates (posterior standard deviations) by Forbes, Kalb and Kofman [37]; May 1993 data.

	Lower regime		Middle Regime		Upper Regime	
	Futures equation	Index equation	Futures equation	Index equation	Futures equation	Index equation
Constant	0.00005	0.00015	0.0	0.0	0.0002	-0.00011
	(0.00009)	(0.00006)	(<0.00001)	(<0.00001)	(0.00011)	(0.00005)
$\Delta \ln F_{t-1,\tau}$	-0.12345	0.15863	-0.05624	0.04065	0.04751	0.04766
	(0.05609)	(0.03542)	(0.01495)	(0.00636)	(0.06256)	(0.03112)
$\Delta \ln F_{t-2,\tau}$	-0.03807	0.29188	0.01931	0.10681	-0.00922	0.17734
	(0.05869)	(0.03645)	(0.01497)	(0.00622)	(0.06353)	(0.0316)
$\Delta \ln F_{t-3,\tau}$	-0.01158	0.17084	0.03534	0.10477	0.00829	0.11326
	(0.06327)	(0.04043)	(0.01491)	(0.00624)	(0.06734)	(0.03395)
$\Delta \ln F_{t-4,\tau}$	0.06247	0.12117	0.01013	0.08486	-0.02001	0.05436
	(0.06711)	(0.04263)	(0.01547)	(0.00633)	(0.06567)	(0.03414)
$\Delta \ln F_{t-5,\tau}$	0.02033	0.13589	0.01195	0.06417	0.02491	-0.00905
	(0.06444)	(0.04142)	(0.01539)	(0.00632)	(0.06874)	(0.03503)
$\Delta \ln F_{t-6,\tau}$	-0.06893	0.08754	-0.00914	0.04905	0.04929	-0.01143
	(0.701)	(0.04436)	(0.01479)	(0.00613)	(0.06822)	(0.03292)
$\Delta \ln F_{t-7,\tau}$	-0.00372	0.09263	0.00044	0.04049	0.11949	0.02589
	(0.07338)	(0.04537)	(0.01421)	(0.00614)	(0.0701)	(0.3398)
$\Delta \ln F_{t-8,\tau}$	-0.11563	0.05837	-0.01029	0.02002	0.00399	-0.01271
	(0.06702)	(0.04273)	(0.01361)	(0.00596)	(0.06436)	(0.03191)
$\Delta \ln S_{t-1}$	0.08683	0.07322	-0.05646	-0.04276	0.10756	-0.22998
	(0.10325)	(0.06888)	(0.02935)	(0.0123)	(0.10745)	(0.05474)
$\Delta \ln S_{t-2}$	-0.18424	0.05204	-0.04384	-0.01549	-0.02106	-0.04501
	(0.11179)	(0.07429)	(0.02813)	(0.01166)	(0.11678)	(0.05714)
$\Delta \ln S_{t-3}$	-0.05493	-0.21202	0.02244	-0.0131	-0.19519	-0.04291
	(0.12567)	(0.08068)	(0.02762)	(0.0117)	(0.11776)	(0.05894)
$\Delta \ln S_{t-4}$	0.14323	0.1821	0.00833	-0.00196	0.00402	0.00934
	(0.11565)	(0.07471)	(0.02731)	(0.01189)	(0.11519)	(0.05794)
$\Delta \ln S_{t-5}$	0.43599	-0.11881	0.01723	0.02187	-0.06118	-0.0131
	(0.1222)	(0.07836)	(0.02805)	(0.01157)	(0.10984)	(0.05489)
$\Delta \ln S_{t-6}$	0.07386	0.04281	0.03839	0.0017	0.06791	0.07497
	(0.1099)	(0.06971)	(0.02626)	(0.01102)	(0.11368)	(0.05546)
$\Delta \ln S_{t-7}$	-0.22131	-0.04934	-0.00421	0.01792	-0.24175	0.01335
	(0.1084)	(0.06868)	(0.02552)	(0.01079)	(0.11237)	(0.05524)
$\Delta \ln S_{t-8}$	-0.07016	0.03165	0.00519	0.01695	0.17988	0.03239
	(0.11353)	(0.07282)	(0.02585)	(0.01049)	(0.10286)	(0.05191)
Z_{t-1}	0.00054	0.0013	0.00002	0.00017	-0.00124	0.00109
	(0.00076)	(0.00047)	(0.00007)	(0.00003)	(0.00077)	(0.00038)
Thresholds	$c_1 = -0.1039,\ c_2 = 0.1278$ (maximum a-posteriori estimates).					
Observations	365		6269		418	

Table 6.6: TVECM estimates (absolute t-ratios) by Tsay [117]; May 1993 data.

	Lower regime		Middle Regime		Upper Regime	
	Futures equation	Index equation	Futures equation	Index equation	Futures equation	Index equation
Constant	0.00002	0.00005	0.0	0.0	-0.00001	-0.00005
	(1.47)	(7.64)	(0.07)	(0.53)	(0.74)	(6.37)
$\Delta \ln F_{t-1,\tau}$	-0.8468	0.07098	-0.03861	0.04037	0.04102	0.02305
	(3.83)	(6.15)	(1.53)	(3.98)	(1.72)	(1.96)
$\Delta \ln F_{t-2,\tau}$	-0.0045	0.15899	0.04478	0.08621	-0.02069	0.09898
	(0.2)	(13.36)	(1.85)	(8.88)	(0.87)	(8.45)
$\Delta \ln F_{t-3,\tau}$	0.2274	0.11911	0.07251	0.09752	0.00365	0.08455
	(0.95)	(9.53)	(3.08)	(10.32)	(0.15)	(7.02)
$\Delta \ln F_{t-4,\tau}$	0.02429	0.08141	0.01418	0.06827	-0.02759	0.07699
	(0.99)	(6.35)	(0.6)	(7.24)	(1.13)	(6.37)
$\Delta \ln F_{t-5,\tau}$	0.0034	0.08936	0.01185	0.04831	-0.00638	-0.05004
	(0.14)	(7.1)	(0.51)	(5.13)	(0.26)	(4.07)
$\Delta \ln F_{t-6,\tau}$	-0.00098	0.07291	0.01251	0.0358	-0.03941	0.02615
	(0.04)	(5.64)	(0.54)	(3.84)	(1.62)	(2.18)
$\Delta \ln F_{t-7,\tau}$	-0.00372	0.05201	0.02989	0.04837	-0.023031	0.02293
	(0.15)	(4.01)	(1.34)	(5.42)	(0.85)	(1.95)
$\Delta \ln F_{t-8,\tau}$	0.00043	0.00954	0.01812	0.02196	-0.04422	0.00462
	(0.02)	(0.76)	(0.85)	(2.57)	(1.90)	(0.4)
$\Delta \ln S_{t-1}$	-0.08419	0.00264	-0.07618	-0.05633	0.06664	0.11143
	(2.01)	(0.12)	(1.7)	(3.14)	(1.49)	(5.05)
$\Delta \ln S_{t-2}$	-0.05103	0.00256	-0.1092	-0.01521	0.04099	-0.01179
	(1.18)	(0.11)	(2.59)	(0.9)	(0.92)	(0.53)
$\Delta \ln S_{t-3}$	0.07275	-0.03631	-0.00504	0.01174	-0.01948	-0.01829
	(1.65)	(1.58)	(0.12)	(0.71)	(0.44)	(0.84)
$\Delta \ln S_{t-4}$	0.04706	0.01438	0.02751	0.0149	0.01646	0.00367
	(1.03)	(0.6)	(0.71)	(0.96)	(0.37)	(0.17)
$\Delta \ln S_{t-5}$	0.08118	0.02111	0.03943	0.0233	-0.0343	-0.00462
	(1.77)	(0.88)	(0.97)	(1.43)	(0.83)	(0.23)
$\Delta \ln S_{t-6}$	0.0439	0.04569	0.0169	0.01919	0.06084	-0.00392
	(0.96)	(1.92)	(0.44)	(1.25)	(1.45)	(0.19)
$\Delta \ln S_{t-7}$	-0.03033	0.02051	-0.08647	0.0027	-0.00491	0.03597
	(0.7)	(0.91)	(2.09)	(0.16)	(0.13)	(1.9)
$\Delta \ln S_{t-8}$	-0.0292	0.03018	0.01887	-0.00213	0.0003	0.02171
	(0.68)	(1.34)	(0.49)	(0.14)	(0.01)	(1.14)
Z_{t-1}	0.00024	0.00097	-0.0001	0.00012	0.00025	0.00086
	(1.34)	(10.47)	(0.3)	(0.86)	(1.41)	(9.75)
Thresholds	$c_1 = -0.022574,\ c_2 = 0.037673.$					
Observations	2234		2410		2408	

where

$$\boldsymbol{X}_t^\top = \begin{bmatrix} X_{11t} & \cdots & X_{1p_1t} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & X_{21t} & \cdots & X_{2p_2t} \end{bmatrix} \tag{6.11}$$

$$F(\boldsymbol{U}_t)^\top = \begin{bmatrix} f_{11}(U_{11t}) & \cdots & f_{1p_1}(U_{1p_1t}) & 0 & \cdots & 0 \\ 0 & \cdots & 0 & f_{21}(U_{21t}) & \cdots & f_{2p_2}(U_{2p_2t}) \end{bmatrix} \tag{6.12}$$

The relevance with the VECM comes from our choice of regressors $\boldsymbol{X}_t$ and functional coefficient arguments $\boldsymbol{U}_t$. We perform a model search through candidate regressor variables $(\Delta \ln F_{t-1,\tau}, \ldots, \Delta \ln F_{t-12,\tau} \Delta \ln S_{t-1}, \ldots, \Delta \ln S_{t-12}, Z_{t-1})$ and a unity variable to allow for a varying mean level, and with functional coefficient argument set to Z_{t-1}. The argument variable is fixed in order to capture the error-correcting behavior and to make the results comparable with those of the TVECM, but more flexibility could be allowed. For example, the candidate arguments could be $(Z_{t-1}, \ldots, Z_{t-4})$, corresponding to the candidate delays in the threshold variable for the TVECM. We use our forward model selection procedure to incrementally add new terms for either coordinate of $\Delta \boldsymbol{Y}_t$, starting from a null model without mean and trying to minimize BIC. For the functional coefficients, we use a projected process approximation based on 10 kernels, centered at the observed quantiles of the argument variable, and we also test for constant coefficients. The resulting model specification for May is

$$\Delta \ln F_{t,\tau} = \epsilon_{1t} \tag{6.13}$$

$$\begin{aligned} \Delta \ln S_t = {} & f_1 Z_{t-1} + f_2(Z_{t-1}) \Delta \ln F_{t-1,\tau} + f_3(Z_{t-1}) \Delta \ln F_{t-2,\tau} + f_4 \Delta \ln F_{t-3,\tau} + \\ & f_5 \Delta \ln F_{t-4,\tau} + f_6 \Delta \ln F_{t-5,\tau} + f_7 \Delta \ln F_{t-6,\tau} + f_8 \Delta \ln S_{t-9} + \epsilon_{2t} \end{aligned} \tag{6.14}$$

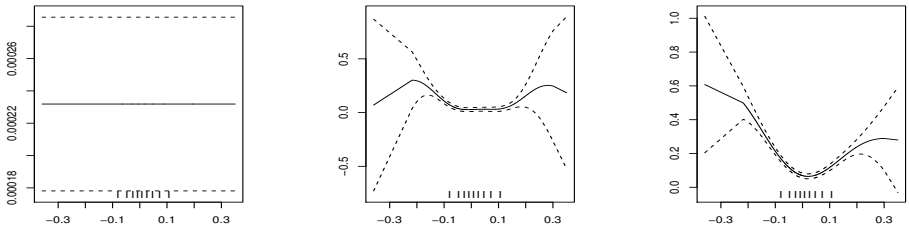
and all estimated coefficient functions are plotted in Fig. 6.14. The resulting model specification for November is

$$\begin{aligned} \Delta \ln F_{t,\tau} &= \epsilon_{1t} && (6.15) \\ \Delta \ln S_t &= f_1(Z_{t-1}) + f_2 \Delta \ln F_{t-1,\tau} + f_3 \Delta \ln F_{t-2,\tau} + f_4 \Delta \ln F_{t-3,\tau} + \\ & \quad f_5 \Delta \ln F_{t-4,\tau} + f_6 \Delta \ln F_{t-5,\tau} + f_7(Z_{t-1}) \Delta \ln S_{t-1} + f_8 \Delta \ln S_{t-2} + \\ & \quad f_9 \Delta \ln S_{t-3} + f_{10} \Delta \ln S_{t-4} + \epsilon_{2t} && (6.16) \end{aligned}$$

and all estimated coefficient functions are plotted in Fig. 6.15. The hyperparameters for the two models include the prior means, the smoothing parameters of the Gaussian kernels, and the error variances and correlation. They are given in Table 6.7, together with the maximum marginal likelihood, the edf and the BIC.

Table 6.7: Hyperparameters for May's model (6.14-6.14) and November's model (6.14-6.14).

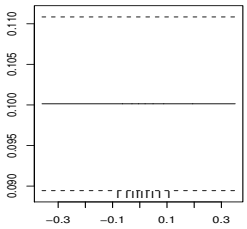
	May		November	
	μ	h	μ	h
f_1	0.02319	∞	0.00902	0.19024
f_2	-0.05467	0.17768	0.06999	∞
f_3	0.25104	0.30457	0.05618	∞
f_4	0.10014	∞	0.05328	∞
f_5	0.07572	∞	0.06816	∞
f_6	0.05526	∞	0.06031	∞
f_7	0.03790	∞	0.34313	0.11656
f_8	0.01921	∞	0.03536	∞
f_9	-	-	0.03686	∞
f_{10}	-	-	0.05376	∞
σ_1^2	8.51267×10^{-4}		7.42622×10^{-4}	
σ_2^2	1.69936×10^{-4}		1.59587×10^{-4}	
ρ	0.26334		0.17031	
ℓ	-18554.81		-20854.65	
DF	26.75843		32.64562	
BIC	37365.26		42024	



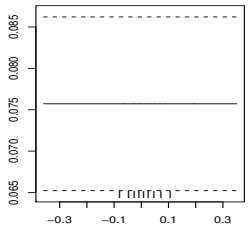
(a) f_1

(b) f_2

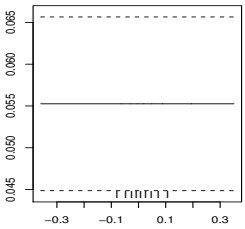
(c) f_3



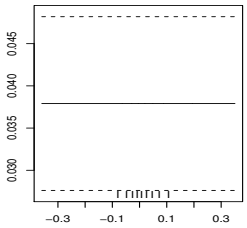
(d) f_4



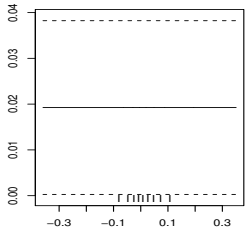
(e) f_5



(f) f_6

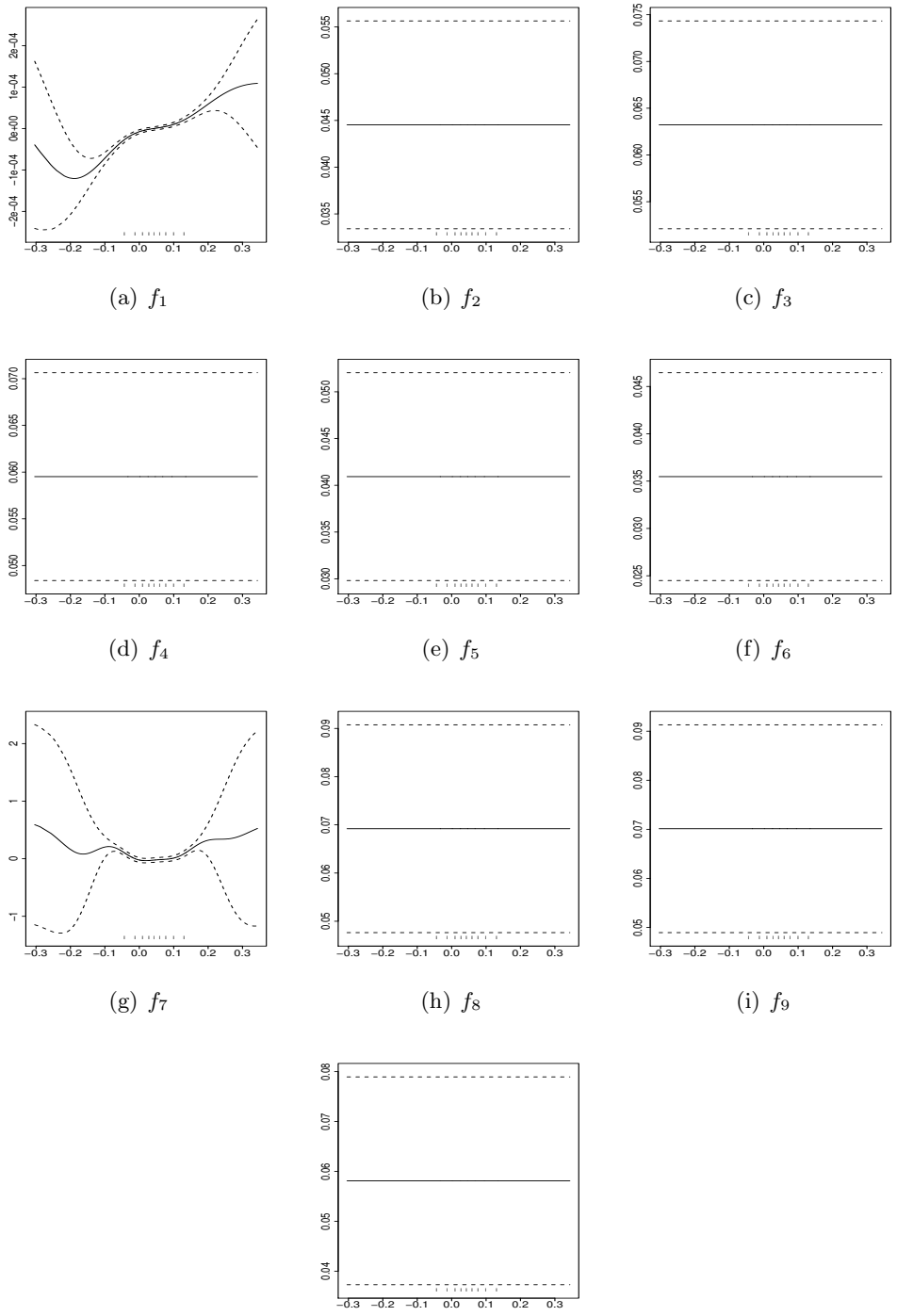


(g) f_7



(h) f_8

Figure 6.14: Plots of functional coefficients in eq. (6.14) for the dynamics of $\Delta \ln S_t$ in May 1993: (—) posterior mean; (- -) pointwise 95% confidence band. Vertical bars at the bottom indicate 1^{st} to 9^{th} observed deciles of the argument variable Z_{t-1}.



(a) f_1 (b) f_2 (c) f_3

(d) f_4 (e) f_5 (f) f_6

(g) f_7 (h) f_8 (i) f_9

(j) f_{10}

Figure 6.15: Plots of functional coefficients in eq. (6.16) for the dynamics of $\Delta \ln S_t$ in November 1993: (—) posterior mean; (- -) pointwise 95% confidence band. Vertical bars at the bottom indicate 1^{st} to 9^{th} observed deciles of the argument variable Z_{t-1}.

Both fitted models are nonlinear, where the nonlinearity comes from only two terms, the remaining having constant coefficients. Moreover, they have explanatory/predictive power only for the index returns, the futures returns being conditionally independent of their past. For both models, the first terms depend on Z_{t-1} alone, and lead to error correcting behavior. To verify this, we plot in Fig. 6.16 the net additive effects of these terms on the index log-returns $\Delta \ln S_t$, i.e. we plot $f_1 Z_{t-1}$ for May's model, and $f_1(Z_{t-1})$ for November's model. From the definition of the error term in eq. (6.7), we expect $\Delta \ln S_t$ to drop when Z_{t-1} is below zero and vice versa, and this is exactly what happens in both cases. For November's model, the effect is nonlinear and more pronounced away from zero. For May's model, the error correcting rate with respect to Z_{t-1} is constant, but if we consider other terms the overall behavior is nonlinear. In particular, the U-shaped form of the coefficient functions for the lagged futures log-returns increases the error correction intensity when the errors move away from zero. Notice that when futures returns are negative, $\ln F_{t,\tau}$ and consequently the error Z_t drop, so Z_t is positively correlated with $\Delta \ln F_{t,\tau}$. Thus, terms like $f(Z_{t-1})\Delta \ln F_{t-1,\tau}$, with f being U-shaped, will give higher error correction rates when Z_{t-1} is away from zero. We demonstrate this graphically in Fig. 6.17, where we plot May's fitted values for $\Delta \ln S_t$ versus Z_{t-1}. It is evident that the error correcting effect is weaker around zero and stronger near the boundaries of the range. As a final confirmation of the nonlinear error correcting behavior, we present in Fig. 6.18 plots of the impulse response function of the mispricing error for different shocks and histories. The shocks are equal to plus/minus one and two standard deviations of the mispricing error. They are introduced only in $\Delta \ln F$, because our model implies that futures returns are random and are, therefore, leading the index returns and shaping the mispricing error. For plots (a) and (b), the pre-shock mispricing errors are close to zero, whereas for plots (c) and (d) they are substantially positive and negative, respectively. The impulse response functions are asymmetric in the last two cases exactly because the error correction intensity is variable. When the error

is already substantially positive/negative, the introduction of a further positive/negative shock will disproportionately increase the error correction strength.

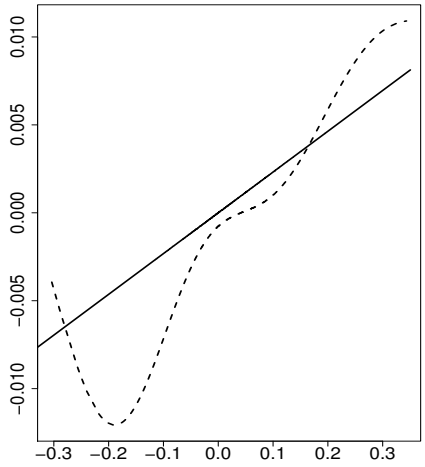


Figure 6.16: Plots of additive effect on $\Delta \ln S_t$ from the error correcting term: (—) $f_1 Z_{t-1}$ for May's model (6.14) and (- -) $f_1(Z_{t-1})$ for November's model (6.16).

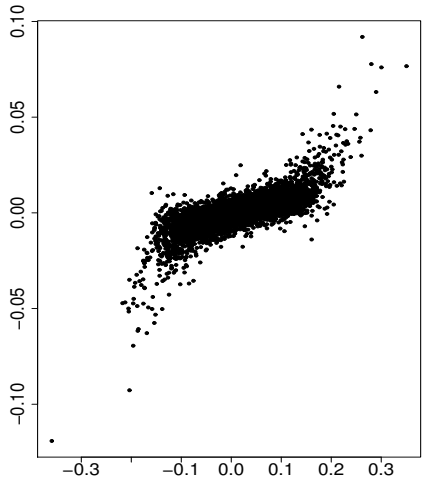


Figure 6.17: Plots of fitted values of $\Delta \ln S_t$ versus Z_{t-1} for May's model (6.14).

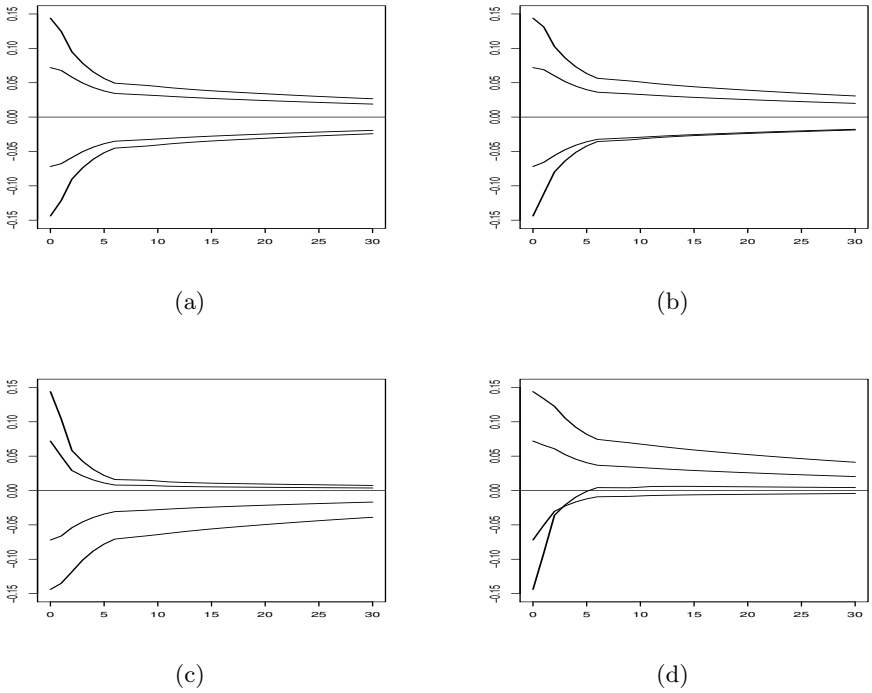


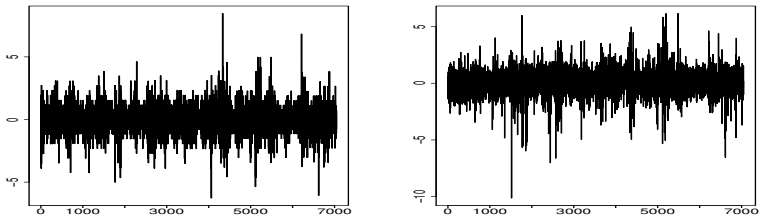
Figure 6.18: Impulse response functions of the mispricing error Z for May's model (6.14). Values at zero give the initially introduced impulse, and subsequent values give the difference between the iterated dynamics of Z with and without the impulse. The functions are conditional on the history of the series at times (a) $t = 100$ ($Z_{100} = 0.028$), (b) $t = 200$ ($Z_{150} = 0.0169$), (c) $t = 324$ ($Z_{324} = 0.176$), (d) $t = 7056$ ($Z_{7056} = -0.089$).

We also check the fit of our model by inspecting the residuals. We only present May's model, since the results were similar to those for November. The recursive residuals for both the index and futures are plotted in Fig. 6.19, versus time and the error variable Z_{t-1}. As is typically the case for financial time series, the residuals are over-dispersed, having heavy tails and several extreme outliers. Moreover, the over-dispersion seems more pronounced when the error term is higher in absolute value. We also performed normality tests on the residuals, and they all strongly rejected the null hypothesis, as expected. On the other hand, the CUSUM process plots in Fig. 6.20 are more optimistic. There are no deviations besides those of the index returns versus time, which means there could be a mean shift

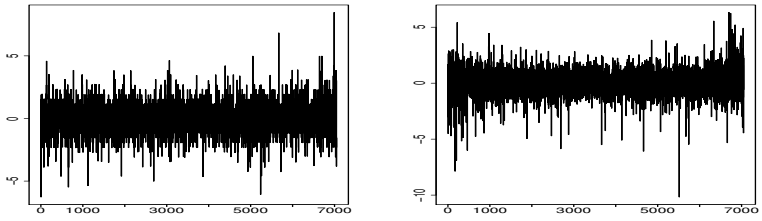
within the month. We could address this issue by subtracting a moving average from the data, but it is not sufficient reason to worry about the model's fit. A more serious problem is evident from the CUSUM-SQ process plot, presented in Fig. 6.21. There are significant departures from the assumptions with respect to both time and Z_{t-1}. The shapes of the plots versus the error term verify our previous remark that the variance is higher when the futures and the index are in disequilibrium. Moreover, the CUSUM-SQ process makes excursions outside the confidence bands across time. The residual ACF plots in Fig. 6.22 give a reasonable explanation for this behavior. Evidently, there is significant autocorrelation in the squared residuals, which clearly suggests that the data exhibit volatility clustering. We also performed Ljung-Box and McLeod-Li tests on the usual and squared residuals and they both rejected the whiteness assumption, the latter ones overwhelmingly. It is obvious that the i.i.d. normal error assumption is inadequate for our data, and other error specifications would be more realistic. For financial time series, in particular, it is common to assume GARCH-type models with more dispersed distributions, such as t or alpha-stable. Nevertheless, we will use the model as such for making point predictions using the conditional mean function, and we will not focus so much on its stochastic or second order properties.

6.2.4 Model Comparisons

We compare the different TVECMs and our model, which we refer to as GP-VECM from now on. First we point out that all models exhibit nonlinear error correcting behavior, and that this behavior depends primarily on the error term Z_{t-1}, as we would expect. However, the models have quite different forms depending on the estimation method used and the time period they are applied to. It is difficult to identify a data generating mechanism that is consistent through time and robust to changes in the model specification. For this reason, we concentrate on predictive ability for comparison purposes. In Table 6.8 we give the mean

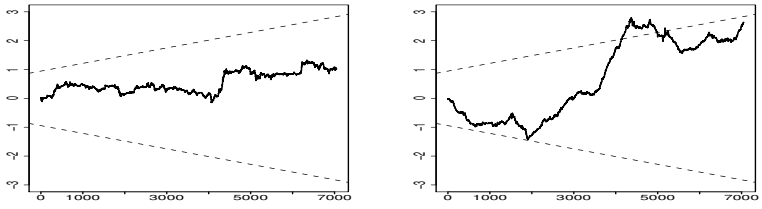


(a) Futures vs. time (b) Index vs. time

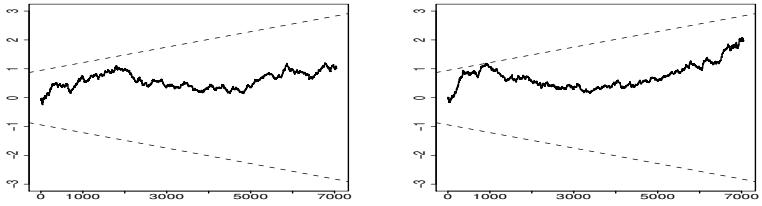


(c) Futures vs. Z_{t-1} (d) Index vs. Z_{t-1}

Figure 6.19: Plots of recursive residuals for May's model (6.14-6.14).



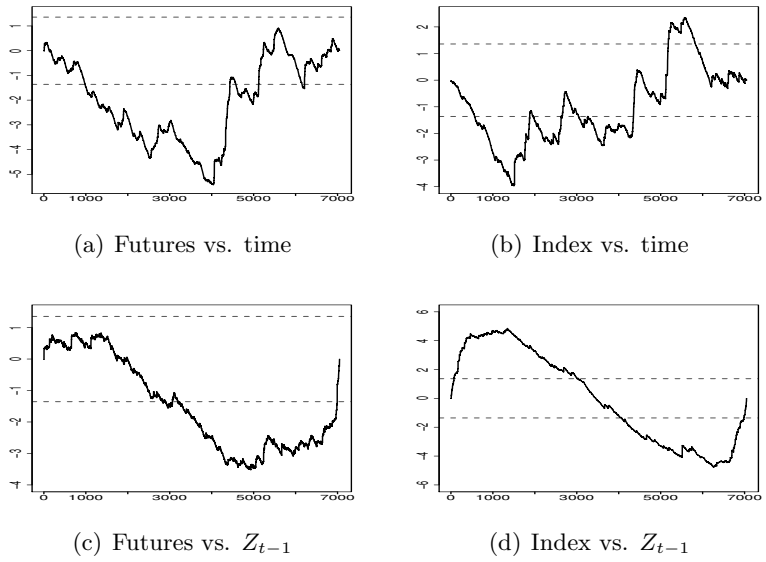
(a) Futures vs. time (b) Index vs. time



(c) Futures vs. Z_{t-1} (d) Index vs. Z_{t-1}

Figure 6.20: Plots of CUSUM processes for May's model (6.14-6.14); (- -) 95% confidence bands .

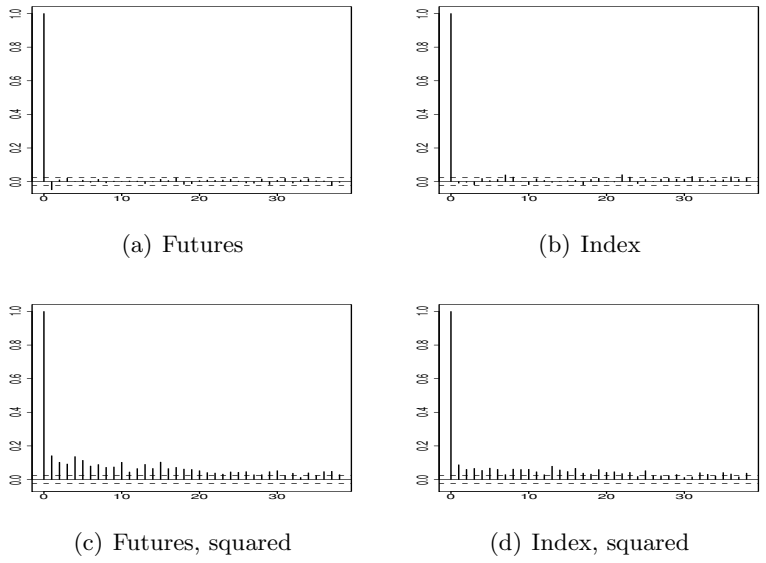
squared errors of the fitted models for May's data, and predictive mean squared errors of one-step-ahead predictions applied to November's data. For the TVECM we present the es-



(a) Futures vs. time (b) Index vs. time

(c) Futures vs. Z_{t-1} (d) Index vs. Z_{t-1}

Figure 6.21: Plots of CUSUM-SQ processes for May's model (6.14-6.14); (- -) 95% confidence bands .



(a) Futures (b) Index

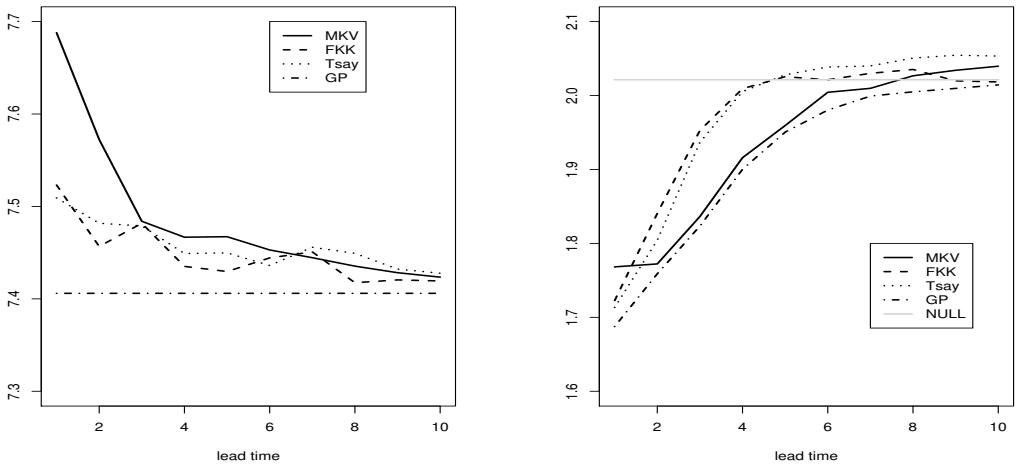
(c) Futures, squared (d) Index, squared

Figure 6.22: Plots of residual and squared residual ACF for May's model (6.14-6.14); (- -) 95% confidence bands .

timation methods of Martens, Kofman and Vorst (MKV), Forbes, Kalb and Kofman (FKK) and Tsay, and we also include a zero-mean null model for comparison. For the Bayesian

model of FKK, we use the maximum a-posteriori estimates of the thresholds and the posterior means of the coefficients to get point predictions. Similarly for our method, we use the posterior means of the functional coefficients to get point predictions. As we can see, MKV and FKK give better in sample but worse out of sample performance. Tsay's model gives the best TVECM results, but its MSPE is still higher than our model's. Moreover, we present in Fig. 6.23 the mean squared predictive error for each model's one- to ten-step-ahead iterative predictions. Overall, our GP-VECM has the lowest predictive error for all lead times. The MKV model improves a lot for the index and for lead times greater than one, but for the futures it still has the highest error. Another interesting characteristic is that the mean futures error for all TVECMs actually decreases with time. This is an indication that these models do not offer any predictive improvements over the null model for the futures returns. This observation is consistent with empirical evidence that futures prices tend to lead index prices, e.g. see Stoll and Whaley [110] or Kawaller, Koch and Koch [66]. This roughly means that information or expectations about market movements are first reflected on the futures prices and later on the index. This lead-lag effect is supported by the argument that futures prices readily reflect new information, whereas the index has many different components that need to change before its level changes. Our model expresses this explicitly by not including any terms for the conditional mean of the futures returns. Thus, it affords an interpretation consistent with the hypothesis that changes in the value of the asset are unpredictable. The other models also point in this direction (compare for example the values of the adjusted R^2 for futures and index returns in Table 6.3) but they do not explicitly account for this.

We believe that the main reason for this overfitting behavior is the way TVECMs are specified and estimated. There is an inherent difficulty in doing these operations in an integrated way and the resulting models tend to be overparametrized. First of all, you cannot



(a) Futures (b) Index

Figure 6.23: Mean square error of iterative predictions versus lead time for November's data, using all three TVECM and our GP-VECM, fitted to May's data.

Table 6.8: Mean squareds errors for May and one-step-ahead mean squared predictive errors for November, using the proposed TVECMs in the literature and our GP-VECM. Standard errors appear in parentheses and all values are rescaled by a factor of 10^4.

		TVECM			
	Null Model	MKV	FKK	Tsay	GP-VECM
MSE for May					
Futures	8.49499	8.38331	8.36555	8.39275	8.49499
	(0.23326)	(0.22717)	(0.22416)	(0.22843)	(0.23326)
Index	2.42336	1.65408	1.64677	1.71767	1.70092
	(0.09539)	(0.05582)	(0.05329)	(0.05805)	(0.05699)
MSPE for November (using May's fit)					
Futures	7.40599	7.68792	7.51579	7.47968	7.40599
	(0.19514)	(0.20191)	(0.19732)	(0.19624)	(0.19514)
Index	2.02139	1.76827	1.72038	1.71339	1.68757
	(0.07346)	(0.05501)	(0.05123)	(0.05199)	(0.05231)

fit autoregressive parameters that are constant through regimes without resorting to profile methods. Moreover, the estimation of the threshold values is cumbersome, relying on grid searches in classical estimation or numerical methods in the Bayesian case, and often

requiring subjective input or ad hoc practices. The autoregressive order and/or variable selection procedure is also not well established. Sometimes the model specification is fixed beforehand, as in FKK and Tsay, whereas other times it is determined in an ad hoc basis, as in MKV where they only keep significant coefficients. Using predictive criteria for this purpose can also be complicated. Exhaustive model search is slow when combined with the grid search for the thresholds, and even greedy methods are tricky to handle because every variable added can change the optimal number of regimes and threshold locations. As a result, we believe that the previously presented TVECMs were overparametrized. For May's data, MKV fit 79 coefficients, only counting the autoregressive coefficients and excluding the threshold and covariance parameters. Both FKK and Tsay fit 108 coefficients for the same data, whereas our model has 26.75 effective degrees of freedom. These numbers are not directly comparable but they are indicative of the complexity of each model. Thus, for descriptive purposes, we believe our model is more parsimonious, since nonlinearity only comes from a couple of terms. One descriptive advantage for which TVECMs are popular in this context is that they give information on so called arbitrage bands. These are the values of the mispricing error above which arbitrage opportunities are profitable. We are skeptical of this interpretation for the thresholds, because all three TVECM give different results. Even though FKK and Tsay use the same number of regimes and autoregressive order, their thresholds are markedly different. The discrepancy comes mainly from the estimation procedure. Tsay uses a predictive criterion, so his regimes have almost the same number of observations in order to reduce predictive variance. Moreover, the assumption that the series changes its behavior in a discrete fashion, when the mispricing crosses a threshold, might not be well founded, given that different TVECM estimation procedures cannot pick up consistent threshold values.

Finally, we point out some practical aspects in model fitting. Before applying a TVECM

both MKV and Tsay perform nonlinearity tests, but our method does not require this step because the feasible set of our selection procedure includes linear models. Moreover, MKV use a test to decide between a three and a five regime model. Our use of a single criterion in model selection avoids the need and dangers of multiple testing. Tsay's procedure acknowledges this fact and the need for integrated estimation, but we believe that certain difficulties still persist, having to do in particular with grid searches. We do not have information on the time requirements to fit the other models, but we believe our method is comparable. It took us approximately 16 hours to fit each month, running R code on a Pentium 4 workstation. The procedure is also highly automated. The only decisions we have to take concern the candidate regressor and argument variables, and they were all taken beforehand. We therefore believe that our GP-VECM is an attractive alternative to the existing TVECM. It tends to give more parsimonious models, it is easy to fit without requiring much subjective or ad hoc input, and it gives better predictive behavior than the other models for the data at hand.

6.3 Nonlinear Stochastic Volatility Model

6.3.1 Introduction

In this section we apply our nonlinear state space (SS) methodology to a stochastic volatility (SV) model. This class of models has recently become very popular in financial econometrics and has many applications especially in derivatives pricing and risk management. The main idea is to allow the variance of a time series to be a stochastic process itself. To be more concrete, suppose we are interested in modeling a financial asset S_t using a SV model. The standard single-asset formulation of the model in continuous time is

$$dS_t = \mu S_t dt + \sigma_t S_t dW_t \tag{6.17}$$

$$df(\sigma_t) = a(\sigma_t)dt + b(\sigma_t)dB_t \tag{6.18}$$

This model combines the usual Black-Scholes dynamics for the log-returns of S_t in (6.17) with a transformed random volatility process $f(\sigma_t)$, following its own stochastic differential equation (6.18). The functions f, a, b can be arbitrary and W_t, B_t are Brownian motions, possibly correlated.

There are three important special cases of the general SV model, depending on the form of (6.18), which we list below

- Hull-White (HW) model: $d\sigma_t^2 = a\sigma_t^2 dt + b\sigma_t^2 dB_t$
- Cox-Ingersoll-Ross (CIR) model: $d\sigma_t^2 = a_1(a_0 - \sigma_t^2)dt + b\sigma_t dB_t$
- log-Ornstein-Uhlenbeck (log-OU) model: $d\log(\sigma_t^2) = a_1(a_0 - \log(\sigma_t^2))dt + bdB_t$

The HW model is essentially a log-normal model for σ_t^2 and can be equivalently expressed as $d\log(\sigma_t^2) = (a + b^2/2)dt + bdB_t = a'dt + bdB_t$, in analogy to the log-OU model. The

model states that the logarithm of the variance process follows a random walk, and Hull and White [61] give an analytical solution to European options when the two processes are uncorrelated. The CIR model uses mean-reverting dynamics with a different scaling for the random increments of the volatility process. It is also known as the Heston model, after Heston [59] who provided a closed form solution for the price of European options, also allowing for non-zero correlation. The log-OU model is an extension of the HW model to mean-reverting dynamics, but no analytic solution is available for option pricing. Finally, all three models have the attractive property that the resulting volatility process $\{\sigma_t\}$ is always positive. For our application the log-volatility models fit our framework best, because we need $b(\sigma_t)$ to be constant or at least state-independent. For this reason, and because it is more general than the HW model, we focus on the log-OU model.

The log-OU model is a continuous time model, but for statistical estimation it must be fit with discrete observations. Likelihood estimation requires the transition probabilities of the continuous stochastic differential equation which are usually intractable, especially in the presence of correlation. Therefore, we work with a discretized version of the continuous model from which we can easily obtain transition probabilities. We adopt the Euler discretization of the log-OU model

$$\frac{\Delta S_t}{S_{t-1}} = \frac{S_t - S_{t-1}}{S_{t-1}} = \mu + \sigma_t \epsilon_t \tag{6.19}$$

$$\Delta \log(\sigma_{t+1}^2) = \log(\sigma_{t+1}^2) - \log(\sigma_t^2) = a_1(a_0 - \log(\sigma_t^2)) + b\xi_t \tag{6.20}$$

We can simplify the notation by letting $R_t = \frac{S_t - S_{t-1}}{S_{t-1}}$ be the return on the asset and $h_t = \log(\sigma_t^2)$ the log-volatility. We also rearrange the long run average of log-volatility a_0 to appear as a scaling factor $\bar{\sigma}$ in the return errors, where $\bar{\sigma}$ can be thought of as the mean volatility level of returns. We do this in order to remove any constant terms in the dynamics

of the latent process, as is usually the case in SS models. The equivalent model becomes

$$R_t = \mu + \sigma_t \bar{\sigma} \epsilon_t \tag{6.21}$$

$$h_{t+1} = a_1 h_t + \sigma_\eta \eta_t \tag{6.22}$$

where $\{\epsilon_t, \eta_t\}$ are possibly correlated standard normal random variables.

This discrete model poses some difficulties in estimation because the process h_t is latent and appears in the variance of the observed returns R_t. Explicit maximum likelihood estimation is not possible, and nonlinear, non-Gaussian filtering schemes such as sequential Monte Carlo (also known as particle filters) have to be employed, e.g. see Doucet et al. [29]. However, Harvey et al. [52] propose a transformation which linearizes the above model. They take the logarithm of the squared excess returns $y_t = \log((R_t - \mu)^2)$, so that the observation equation becomes $y_t = \log(\sigma_t^2) + \log(\bar{\sigma}^2) + \log(\epsilon_t^2)$. The transformation can give infinite values if $R_t - \mu$ is zero, which can happen for example if we fix $\mu = 0$ and some return is zero (this event has positive probability since, in reality, assets take discrete values). To avoid this situation, the authors suggest using an estimate of μ from the data. The transformed error term $\log(\epsilon_t^2)$ is distributed as the logarithm of a chi-squared random variable, in particular it has mean equal to -1.27 and variance equal to $\pi^2/2$, but the authors approximate the error term with a normal distribution with the same moments. Fig. 6.24 presents the density of the log-χ^2 and that of the normal approximation for comparison. Assuming for now that ϵ_t and η_t are uncorrelated, the model after the transformation becomes

$$y_t = \omega + h_t + \sigma_\xi \xi_t \tag{6.23}$$

$$h_{t+1} = a_1 h_t + \sigma_\eta \eta_t \tag{6.24}$$

where $\omega = -1.27 + \log(\bar{\sigma}^2)$, $\sigma_\xi = \sqrt{\pi}/2$, and ξ_t and η_t are i.i.d standard normal. Thus, the

model is recast in the classic linear Gaussian SS form, and the Kalman filtering machinery is readily available for estimation. This is essentially a Quasi Maximum Likelihood (QML) estimation method.

One problem with this approach is that the exact error distribution of $\log(\epsilon_t^2)$ is quite skewed to the left and can accommodate very small observations y_t compared to the approximate Gaussian error ξ_t. Durbin and Koopman [32] proposed a Monte Carlo likelihood correction based on importance sampling for linear SS models where the emission probabilities (i.e. the observation errors) are non Gaussian. Sandmann and Koopman [98] implemented this approach on a SV model similar to ours, but we do not attempt this correction for two reasons. First, the parameter estimates from QML will still be consistent, even though they can suffer from poor small sample properties. Second, the correction cannot provide exact filtering distributions, so in this respect there is no gain. Another issue which we do address, though, is that of correlation between the returns and the volatility process. This is important since empirical research has shown that financial markets react differently depending on the direction of returns, see Bekaert and Wu [6]. In particular, volatility tends to rise in response to big negative returns and fall in response to big positive returns, a phenomenon known as asymmetric volatility. Notice that since the transformation uses the square of the returns, information about the sign, and in extension the correlation, is lost. Harvey and Shephard [53] proposed a SS model conditional on the sign of the return errors $s_t = \text{sign}(\epsilon_t)$ which can recover information on correlation. Letting $\rho = \text{Cor}(\epsilon_t, \eta_t)$ be the correlation of the original return and volatility errors, they show that the conditional linear SS form is

$$y_t = \omega + h_t + \sigma_\xi \xi_t \tag{6.25}$$

$$h_{t+1} = \left(a_1 - \frac{\gamma s_t}{\sigma_\xi^2}\right) h_t + s_t \left(\delta + \frac{\gamma}{\sigma_\xi^2}(y_t - \omega)\right) + \sigma_{\eta'} \eta_t' \tag{6.26}$$

where $\delta = .7979\rho\sigma_\eta$, $\gamma = 1.1061\rho\sigma_\eta$,$\sigma^2_{\eta'} = \sigma^2_\eta - \delta^2 - \gamma^2/\sigma^2_\xi$ and ξ_t, η'_t are i.i.d. standard normal. By estimating the parameters of this model we can estimate the corresponding parameters of model (6.21-6.22), including correlation.

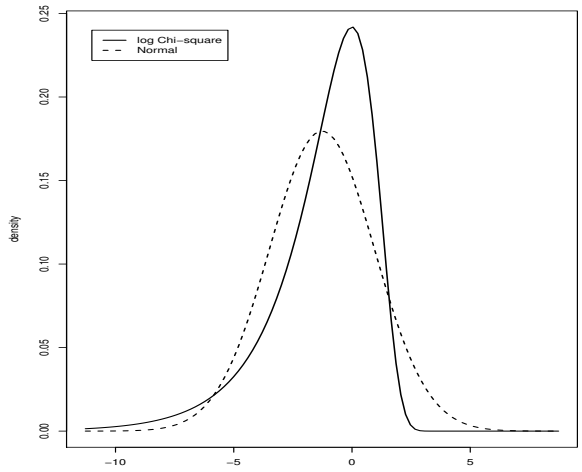


Figure 6.24: Density of log-χ^2 and $\mathcal{N}(-1.27, \pi^2/2)$ random variables.

Our goal is to extend the stochastic volatility model (6.21-6.22) to a nonlinear setting. We do this by adding a nonlinear function of the observation error in the dynamics of the model

$$R_t = \mu + \sigma_t \bar{\sigma} \epsilon_t \tag{6.27}$$

$$h_{t+1} = a_1 h_t + f(\epsilon_t) + \sigma_\eta \eta_t \tag{6.28}$$

The particular choice of argument is based on the remark that the correlation in the errors produces a linear effect of ϵ_{t-1} on h_{t-1}. To verify this notice that $\eta_t = \rho\epsilon_t + \sqrt{1-\rho^2}\eta'_t$, where ϵ_t, η'_t are independent, so the original dynamics in (6.22) can be rewritten as $h_{t+1} = a_1 h_t + \sigma_\eta \rho \epsilon_t + \sigma_\eta \sqrt{1-\rho^2}\eta'_t$. Therefore, it seems natural to extend the linear dependence of h_{t+1} on ϵ_t to a nonlinear one, by introducing the term $f(\epsilon_t)$. In order to fit this model

in our nonparametric NLSS methodology, the argument ϵ_t to the function must be known explicitly by time t, which is not the case for SV models. For practical purposes we use an estimate of ϵ_t, given by dividing the observed excess return $R_t - \mu$ by an estimate of the volatility. Unfortunately, we cannot use the process σ_t, because it is latent and this would break the conditional normality property of the model. For this reason we use an exponentially weighted moving average (EWMA) volatility estimate of the form $\hat{\sigma}_t^2 = \lambda\hat{\sigma}_{t-1}^2 + (1-\lambda)(R_t-\mu)^2$, which only depends on the observations. Thus, the argument variable of f becomes $\hat{\epsilon}_t = \frac{R_t-\mu}{\hat{\sigma}_t}$. The EWMA volatility estimate is a well known benchmark estimate of in-sample volatility, popularized by RiskMetrics [47]. For daily volatilities the value $\lambda = .94$ is recommended, which is what we also use. We can now estimate the function f nonparametrically using our NLSS methodology on the transformed SS model, but we give more details on this in the sequence.

We also look at a competing parametric, non-stochastic volatility model falling under the umbrella of generalized autoregressive conditionally heteroskedastic (GARCH) models. We focus on the first order exponential GARCH (EGARCH) model of Nelson [86], given by

$$R_t = \mu + \sigma_t\epsilon_t \tag{6.29}$$

$$h_{t+1} = a_0 + a_1 h_t + b_1(|\epsilon_t| + g_1\epsilon_t) \tag{6.30}$$

where $\epsilon_t \sim \mathcal{N}(0,1)$. The dynamics of the log-volatility are those of an autoregressive process plus a nonlinear function of the observation error. Note that the process h_t is non stochastic, in the sense that it does not have its own error term. Thus, given its starting value and the observed returns R_t we can exactly reconstruct the volatility process and, by extension, the error ϵ_t. We look at the EGARCH model instead of its precursors, the ARCH and GARCH model, because it ensures the estimated volatility is always positive and it is more closely related to our SS models. It also captures the asymmetry in volatility

by the term $|\epsilon_t| + g_1\epsilon_t$, which allows different behavior depending on the sign of ϵ_t. This can be viewed as a threshold effect on the dynamics of volatility, similar to the threshold model of Glosten, Jagannathan and Runkle [46].

6.3.2 Data and Implementation

We apply our methodology to real data using eight years of daily closing levels on the S&P 500 index, from Jan 1998 to Dec 2005, with a total of 2,012 observations. The data come from the CRSP database of Wharton Research Data Services. From these data we construct the series of returns R_t on the index, which is shown in Fig. 6.25. To simplify model fitting, we use the mean and variance of the returns to estimate μ and the average volatility level $\bar{\sigma}$ in (6.21). We use these values to transform the observations according to Harvey and Shephard [53], and we estimate the LSS model (6.25-6.26). The parameters of the model, namely (a_1, ρ, σ_η), are selected by maximizing the Kalman likelihood, where the initial state distribution in the Kalman filter is $\mathcal{N}(0, \sigma_\eta^2/(1-a_1^2))$, an approximation to the state's stationary distribution.

Next, we turn our attention to our NLSS model, whose dynamics are given by

$$\begin{aligned} y_t &= \omega + h_t + \sigma_\xi \xi_t \\ h_{t+1} &= f(\hat{\epsilon}_t) + \left(a_1 - \frac{\gamma s_t}{\sigma_\xi^2}\right) h_t + s_t \left(\delta + \frac{\gamma}{\sigma_\xi^2}(y_t - \omega)\right) + \sigma_\eta \eta_t \end{aligned}$$

We estimate the function f nonparametrically, using a zero mean GP prior with Gaussian kernel $C(x, x') = \nu^2 \exp\{-(x-x')^2/\ell^2\}$ and a reduced rank approximation. We represent the function as a linear combination $f(x) = \sum_{i=1}^{10} \beta_i C(x, b_i)$ of 10 Gaussian kernels with random coefficients β_i and centered at basis points b_i. For the basis points we use a set of 10 percentiles in the range of $\hat{\epsilon}_t$, in order to avoid the influence of outliers. As we

demonstrated in section 3.5.2, we treat the unknown coefficients as latent variables so that the state equation becomes

$$\begin{bmatrix} h_{t+1} \\ \beta_1 \\ \vdots \\ \beta_{10} \end{bmatrix} = \begin{bmatrix} a_1 - \frac{\gamma s_t}{\sigma_\xi^2} & C(\hat{\epsilon}_t, b_1) & \dots & C(\hat{\epsilon}_t, b_{10}) \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} h_t \\ \beta_1 \\ \vdots \\ \beta_{10} \end{bmatrix} + \begin{bmatrix} s_t \left(\delta + \frac{\gamma}{\sigma_\xi^2}(y_t - \omega) \right) \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} \sigma_\eta \eta_t \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

From this point it is straightforward to apply the Kalman filter. The initial distribution of the state vector is multivariate normal and incorporates the prior on f. For the first coordinate, the log-volatility, we use the same $\mathcal{N}(0, \sigma_\eta^2/(1 - a_1^2))$ as for the LSS model. The remaining coordinates, the coefficients, independently follow a zero mean multivariate normal with variance given by inverse matrix of the Gaussian kernel evaluated at the basis points. The complete set of the model's parameters is $(a_1, \rho, \sigma_\eta, \nu, \ell)$, the last two coming from the covariance kernel. Once again they are selected by maximizing the Kalman likelihood, with initial values adapted from the SS model. Finally, we also estimate the parameters (a_0, a_1, b_1, g_1) of the EGARCH model (6.29-6.30) using conditional maximum likelihood.

The resulting parameter estimates for all three models are given in Table 6.9. Notice that the autoregressive parameter a_1 is very close to 1 for all models, which suggests that volatility has long memory. There also seems to be a high correlation between returns and volatility in the SV models, where the negative sign is in accordance with asymmetric

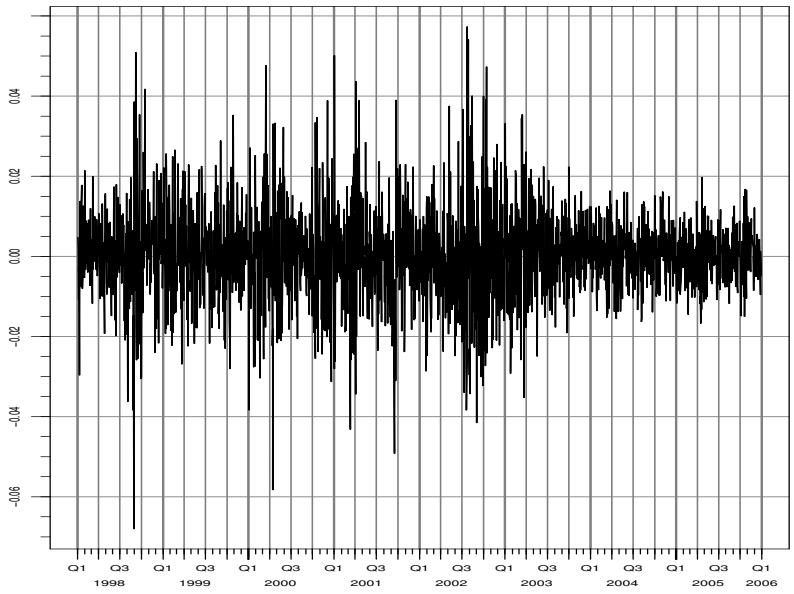


Figure 6.25: Plot of daily returns on the S&P 5000 index, Jan 1998 to Dec 2005.

volatility. The posterior distribution of f is presented in Fig. 6.26, and the estimate shows a pronounced bump around zero while becoming negative on both sides. For comparison purposes, Fig. 6.27 presents the effect of the return error ϵ_t on log-volatility h_{t+1} for all three models. As we pointed out before, this is linear for the LSS model, with slope $\rho\sigma_\eta$. For our NLSS methodology the curve represents the posterior mean of the function f plus the linear effect from correlation. The main difference of the NLSS model is that it produces an increase in volatility for values of ϵ_t around zero. This nonlinear effect does not offer an intuitive physical interpretation, but we rather believe it serves as an adjustment to the dynamics of the process, stemming from the logarithmic transformation. When ϵ_t is close to zero the logarithm of the squared return y_t becomes very negative and pushes the log-volatility h_t down. Hence, a plausible explanation for the form of f is that it tries to counteract this effect. We also calculate the fitted volatility process from each model. For the two SS models, we use the Kalman smoother to get the conditional distribution of $h_t|y_{1:T}$

which is normal, and we define the estimated volatility as the mean of the corresponding log-normal distribution. A plot of the fitted volatility process from 2002 to 2006 for the three models is given in Fig. 6.28, together with the EWMA estimate. It seems that the fits of the models are pretty close, and they also follow the EWMA in sample estimate of volatility. We do not provide diagnostics because our SS models use the transformed data, which are not faithful to the original SV model. The methodology is rather focused on estimating the SV model parameters using QML. Therefore, it is more appropriate to differentiate the models based on how well they describe the evolution of the process, using these estimates.

Table 6.9: Parameter estimates for the three volatility models fitted to S&P 500 returns.

Model					
LSS		NLSS		EGARCH	
a_1	0.9911475	a_1	0.9863075	a_1	0.98797838
ρ	-0.8325423	ρ	-0.8377840	a_0	-0.18476149
σ_η	0.1236039	σ_η	0.2053147	b_1	0.09303694
		ν	0.1158954	g_1	-1.15261145
		ℓ	0.5545815		
$\mu = 0.000196804$, $\bar{\sigma} = 0.01197680$					

6.3.3 Option Pricing

We want to compare our nonlinear model to the competing ones in a practical setting. Volatility models have diverse applications in finance, but we are particularly interested in option pricing. This setting is appropriate because it requires a model for the evolution of volatility. First, we give a brief overview of option pricing under SV models, for a detailed exposition see Fouque et al. [38]. A basic result in mathematical finance states that the arbitrage-free price of an option is given by its expected discounted payoff under a special measure $\mathbb{Q}$, e.g. see Björk [8]. The measure $\mathbb{Q}$ must be equivalent to the observed measure $\mathbb{P}$, but they need not coincide. Under this measure all discounted traded assets are martingales,

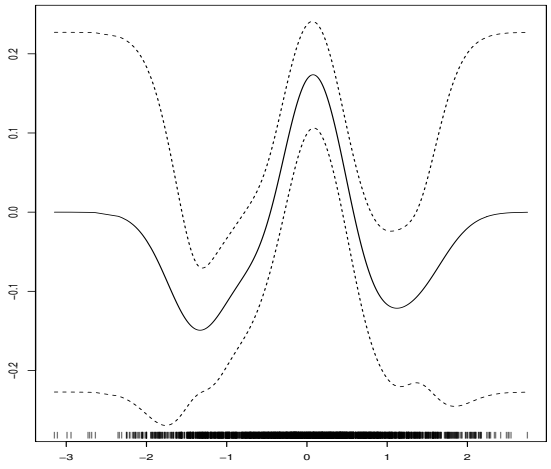


Figure 6.26: Plot of posterior function of f in (6.28).

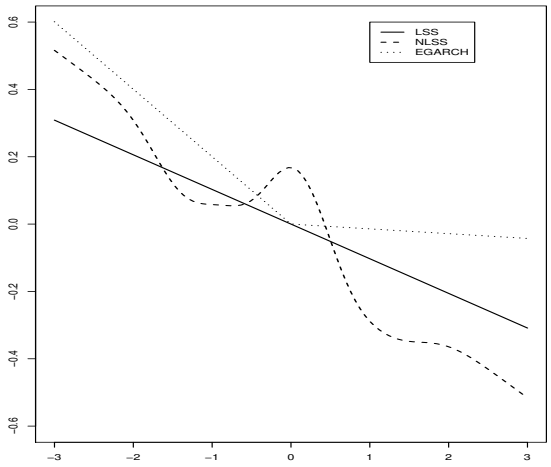


Figure 6.27: Plot of the estimated effect of return error ϵ_t on log-volatility h_{t+1} for all three models.

meaning that the expected rate of returns on traded assets equals the risk free rate r, i.e. $\mathrm{E}_{\mathbb{Q}}[S_T] = S_0 \exp\{rT\}$ (in case the asset pays dividends at a rate d, then r is replaced by $r - d$). For stochastic volatility models there is a complication because the option price of an asset also depends on the volatility process which is not traded. As a consequence,

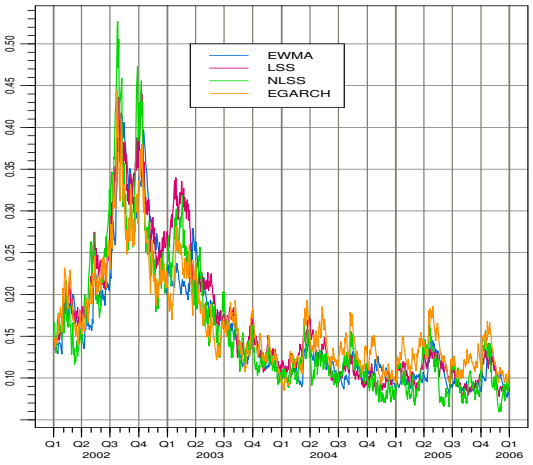


Figure 6.28: Plot of the estimated volatility from all three models and the EWMA volatility, Jan 2002 to Dec 2005.

there is no way to hedge perfectly against random fluctuations in volatility. In this case the market is said to be incomplete because there exist many alternative measures $\mathbb{Q}$ which give consistent option prices, in the sense of absence of arbitrage. In our log-OU model, a valid candidate for $\mathbb{Q}$ is any measure with dynamics

$$\begin{aligned} dS_t &= rS_t dt + \sigma_t S_t d\tilde{W}_t \\ d\log(\sigma_t^2) &= \left[a_1\left(a_0 - \log(\sigma_t^2)\right) + b\left(\rho\lambda_t^S + \sqrt{1-\rho^2}\lambda_t^\sigma\right)\right]dt + bd\tilde{B}_t \end{aligned}$$

where $\tilde{W}_t, \tilde{B}_t$ are Brownian motions with the same correlation ρ as under $\mathbb{P}$, $\lambda_t^S = (\mu - r)/\sigma_t$ and λ_t^σ is any adapted, suitably regular process (in particular it can depend on S_t, σ_t). The terms λ_t^S and λ_t^σ define the change of drift in $\tilde{W}_t$ and $\tilde{B}_t$, respectively, and are called the market and volatility risk premia, respectively. The particular measure that provides the observed option prices is determined by the investor's risk preferences toward volatility, as reflected in λ_t^σ. In practice, it is common to restrict attention to a simple class of measures and use calibration to select the one that matches more closely the observed prices. Usually,

the measure $\mathbb{Q}$ is assumed to belong to the same family of models as the observed measure $\mathbb{P}$, so in our log-OU example we would have

$$d\log(\sigma_t^2) \;=\; \left[a_1'(a_0' - \log(\sigma_t^2)) + b\rho\lambda_t^S\right] dt + bd\tilde{B}_t$$

under $\mathbb{Q}$, and we would have to select the parameters a_1', a_0' by calibration.

Turning back to our discretized log-OU setting, the previous theory can be translated as follows. For pricing options in the linear SV model we use the $\mathbb{Q}$ measure dynamics

$$S_t \;=\; S_{t-1}\exp\left\{r - \sigma_t^2\bar{\sigma}^2/2 + \sigma_t\bar{\sigma}\epsilon_t\right\} \tag{6.31}$$

$$h_{t+1} \;=\; a_0' + \left(a_1' - \frac{\gamma s_t}{\sigma_\xi^2}\right) h_t + s_t\left(\delta + \frac{\gamma}{\sigma_\xi^2}(y_t - \omega)\right) + \sigma_\eta\eta_t \tag{6.32}$$

where $R_t = \log(S_t/S_{t-1})$, $y_t = \log((R_t - r)^2)$, $s_t = \text{sign}(\epsilon_t - \lambda_t^S)$, $\lambda_t^S = (\mu - r)/(\bar{\sigma}\sigma_t)$, $\{\epsilon_t, \eta_t\}$ are standard normal with correlation ρ, and r is the daily risk free rate. Similarly for our nonlinear SV model, the log-volatility dynamics become

$$h_{t+1} \;=\; a_0' + f(\epsilon_t - \lambda_t^S) + \left(a_1' - \frac{\gamma s_t}{\sigma_\xi^2}\right) h_t + s_t\left(\delta + \frac{\gamma}{\sigma_\xi^2}(y_t - \omega)\right) + \sigma_\eta\eta_t \tag{6.33}$$

Notice that besides the values of a_0', a_1', all other parameters in these models are determined by the $\mathbb{P}$ measure dynamics and can be estimated by observations. For completeness, we also give the risk neutral dynamics for the EGARCH model, as proposed by Duan [30]. Since volatility for this model is non-stochastic, there is a unique measure $\mathbb{Q}$ for which log-volatility follows

$$h_{t+1} \;=\; a_0 + a_1 h_t + b_1[|\epsilon_t - \lambda_t^S| + g_1(\epsilon_t - \lambda_t^S)] \tag{6.34}$$

where $\lambda_t^S = (\mu - r)/\sigma_t$. Note that in the EGARCH model, all parameters of interest are defined in terms of the observed measure. For European options, whose value depends only on the price of the asset at the expiration time T, we need to calculate expectations of the form $\mathrm{E}_{\mathbb{Q}}[f(S_T)|S_0, \sigma_0]$. The distribution of the asset prices in all three models specified above cannot be given explicitly, so expectations under $\mathbb{Q}$ are estimated by Monte Carlo simulation.

Our options data consist of a year's worth of daily European option closing prices on the S&P 500 index, from Jan to Dec 2006. The data come from the OptionMetrics database of Wharton Research Data Services. For each day there are different contract specifications offered, depending on expiry date and strike price. We restrict attention to the most liquid contracts, so we only keep options whose strike price is within 10% of the current index price. In order for the effects of different models to be distinguishable we need a sufficient time horizon, so we look at contracts whose expiry date is at least one and two months into the future. Longer expiration dates of up to one year are available, but we exclude these because they are less liquid and their prices might depend on other considerations outside our framework, for example interest rate risk. For each contract we have a highest bid and lowest ask price, from which we form a single option price by taking their midpoint. Furthermore, we remove options whose bid price is zero to avoid bias (because the ask price is positive even for options with no value), as well as prices that violate no arbitrage conditions. The clean data consist of 13,489 prices divided into four groups according to option type (call or put) and expiration date (at least one or two months). On top of that, we extract the required continuously compounded dividend and risk free rates from the data base, the latter being computed from zero coupon bonds with approximately the same maturity date.

We use the fitted volatility models to estimate the observed option prices. At first, we make

the simplifying assumption that the volatility risk premium is zero, so that $a'_0 = a_0, a'_1 = a_1$ and $\mathbb{Q}$ is uniquely defined by the fitted models. Theoretically, this describes the situation when investors are neutral toward volatility risk, i.e. their risk preferences are independent of volatility changes. In order to estimate the integral $\mathrm{E}_{\mathbb{Q}}[f(S_T)|S_0, \sigma_0]$ by Monte Carlo simulation, we generate daily paths of the index and volatility processes as described above. For the initial volatility σ_0 we use the EWMA estimate of volatility of that day for all three models, because we want to compare results based only on their dynamics. Moreover, we found that the EWMA estimates as starting values give better results for all three models than do their intrinsic estimates. Finally, we use antithetic and control variable techniques to improve the accuracy of the Monte Carlo method. The control variable we use is the price of the option under constant volatility equal to σ_0, which is given explicitly by the Black-Scholes formula. The same technique for SV models is also advocated by Hull and White [61]. We calculate the mean absolute error of the estimated prices for each of the three models, using 50,000 simulated paths, and for the Black-Scholes model for constant σ_0 volatility. Table 6.10 presents the results by option type and expiry date. Overall, all models seem do better for closest expiry dates and for put contracts. The Black-Scholes errors are substantially higher, suggesting the constant volatility assumption is not empirically supported. For varying volatility, the EGARCH and LSS models give very similar results, whereas our nonlinear model seems to provide an improvement.

In order to better understand the differences between the models we look at the final distribution of the asset S_T under $\mathbb{Q}$, since it uniquely defines the option prices. Fig. 6.29 presents density estimates for the distribution of the 2-month index price together with the theoretical log-normal density of the same variable under the Black-Scholes model. For this plot, the values of S_0, σ_0 are set to those of the first day of options data. All three non-constant volatility models exhibit an obvious negative skew compared to the log-normal

Table 6.10: Mean absolute error of estimated option prices, Jan-Dec 2006, without volatility risk premium; standard errors in parentheses.

	Call		Put		
Model	1 month	2 months	1 month	2 months	Total
B-S	0.31216	0.50622	0.28258	0.49303	0.39785
	(0.00386)	(0.00598)	(0.00333)	(0.00574)	(0.00258)
EGARCH	0.19995	0.22491	0.17688	0.20691	0.20181
	(0.00257)	(0.00295)	(0.002090)	(0.00270)	(0.00130)
LSS	0.19999	0.22879	0.18003	0.21169	0.20481
	(0.00258)	(0.00307)	(0.00208)	(0.00279)	(0.00133)
NLSS	0.15602	0.202001	0.13736	0.18479	0.16972
	(0.00220)	(0.00304)	(0.00188)	(0.00290)	(0.00129)
# obs.	3,219	3,336	3,543	3,391	13,489

density, even though all distributions have the same mean. This is a well documented property of empirical option pricing measures and relates to the existence of the volatility smile, e.g. see Dennis and Mayhew [27]. It has to do with the fact that, under the measure $\mathbb{Q}$, prices have the potential of a bigger downfall relative to the Black-Scholes model. We can see from the plot that the NLSS model distribution has a fatter left tail, and this causes the difference in the estimated option prices compared to the other models.

We also introduce risk premia in the SV models and calibrate the risk neutral measure $\mathbb{Q}$ to observed option prices. We divide our option data in a training and test set of six months each. We select the parameters a_0' and a_1' by minimizing the mean absolute option pricing error for the first six months of data, and then estimate prices for the following six months using these values. For the LSS model we got $a_0' = -4.2565 \times 10^{-5}$, $a_1' = .98976$ and for the NLSS model we got $a_0' = -.01425$, $a_1' = .97827$. Table 6.11 presents the mean absolute error of the estimated prices for the test period, where we also include the results for the SV models without volatility risk premia (i.e. $\lambda^\sigma = 0$) for comparison. Overall, the error is higher for the test period as compared to the whole year of data, but our previous remarks still hold. The introduction of the risk premium offers an improvement for the LSS model,

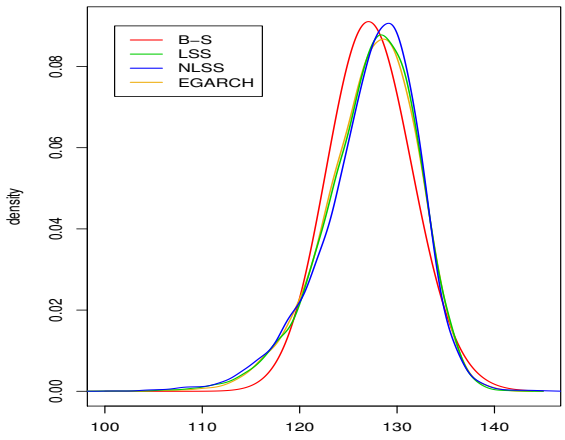


Figure 6.29: Plot of density estimates of the 2-month index level for the three volatility models under the pricing measure $\mathbb{Q}$, together with the theoretical log-normal density of the Black-Scholes model.

but not for our NLSS model. Nevertheless, our model still provides lower pricing errors.

We now make some general comments. We have presented a simple way in which our GP methodology can be practically applied in a SS setting. We extended the SS approach for SV modeling of Harvey et al. [52] by introducing a nonlinear term in the dynamics. This terms seems to compensate for certain effects of the required data transformation, and led to better option pricing results in our example. Moreover, the nonlinearity can be treated easily, requiring a small modification beyond Kalman filtering for linear models. Another advantage of our method, owing to the relation with the LSS model, is that it can be readily extended to a multivariate setting. For example, we can model the volatilities of many assets using just a few volatility factors. On the negative side, we need a transformation to recast the SV model into an approximate SS model, and we are essentially doing QML estimation. As a result, we need big samples to ensure that our estimates are well behaved. For this reason we used eight years of daily data in our application, which is a relatively big time

span for such a model to be the consistent. In practice, we would opt for more frequent data if we had access to them.

Table 6.11: Mean absolute error of estimated option prices, Jul-Dec 2006, with and without volatility risk premia; standard errors in parentheses.

	Call		Put		
Model	1 month	2 months	1 month	2 months	Total
BS	0.37027	0.63214	0.31443	0.59249	0.48447
	(0.00588)	(0.00860)	(0.00523)	(0.00858)	(0.00406)
EGARCH	0.25205	0.30365	0.2070431	0.27170	0.25976
	(0.00380)	(0.00400)	(0.00310)	(0.00367)	(0.00188)
LSS ($\lambda^\sigma = 0$)	0.25426	0.30262	0.21133	0.27322	0.26150
	(0.00379)	(0.00428)	(0.00310)	(0.00388)	(0.00174)
LSS ($\lambda^\sigma \neq 0$)	0.24346	0.27464	0.20247	0.24890	0.24301
	(0.00353)	(0.00378)	(0.00271)	(0.00324)	(0.00169)
NLSS ($\lambda^\sigma = 0$)	0.18940	0.19695	0.15687	0.18142	0.18126
	(0.00337)	(0.00393)	(0.00269)	(0.00360)	(0.00174)
NLSS ($\lambda^\sigma \neq 0$)	0.20272	0.22838	0.16718	0.20845	0.20224
	(0.00330)	(0.00349)	(0.00258)	(0.00313)	(0.00160)
# obs.	1,547	1,805	1,691	1,830	6,873

Chapter 7

Summary and Future Work

7.1 Summary and Contributions

In this final chapter we provide a summary of the thesis and highlight our contributions. Our main goal was the development of a nonparametric estimation methodology based on GPs for analyzing time series. We adopted the FAR model of Chen and Tsay [22] for describing the series dynamics and used GP regression for estimating the coefficient functions, in analogy to the framework of O'Hagan [88] for independent data. Our methodology has significant departures from O'Hagan, though, and our contributions lie in addressing both practical and theoretical issues pertinent to the nature of time series. We proposed an empirical Bayes procedure for specifying the GP prior that allows different smoothness for different functions and leads to parsimonious models, and we described the resulting estimation and prediction mechanics. In particular, our method can accommodate combinations of constant and varying coefficients, and is geared toward providing models with stable dynamics. We also compared it to available nonlinear and nonparametric methods. We pointed out its advantages over threshold models and local estimation with regard to modeling flexibility, and over splines with regard to extrapolation.

In terms of applicability, we developed an approximate inference scheme that overcomes the restrictions of our method on the sample sizes it can handle. Specifically, we extended the reduced rank approximation technique from GP regression to our FAR setting. Our proposed approximation scheme uses a simple representation for the functions and introduces a small perturbation in the prior covariance kernels. It reduces our method's computational cost from cubic to linear in the data, and we provided evidence of its suitability for estimating smooth coefficient functions. Based on this approximation, we extended our method to multivariate and SS models. For the latter, we indicated how they can be treated conveniently using the Kalman filter.

We also explored the theoretical properties of our method. We used the connection between GP and regularized regression to prove the consistency of our functional coefficient estimates from a frequentist perspective. To this end, we used identifiability and ergodicity conditions, assuming the smoothness of each function is known. We made a distinction between Markovian and more general time series regression models, adapting the conditions accordingly. Moreover, we provided an asymptotic result for approximate inference which established the convergence of the estimates to appropriate projections of the true functions.

Furthermore, we proposed an integrated way of performing statistical inference based on our methodology. We developed a greedy model selection algorithm that is capable of capturing a broad range of model specifications, and that avoids overfitting by allowing constant coefficients. In addition, we suggested a suite of procedures for evaluating a model's fit, which were collected from the literature. These were especially suited for our method, and they included statistical and graphical procedures based on residuals and simulation.

Finally, we presented three applications of our methodology to real data sets. The first was on a simple univariate series from natural sciences, whereas the second concerned a large-sample bivariate series from financial econometrics. The last application involved our state-space methodology for analyzing a stochastic volatility model, and also used financial data. In all three cases we provided results from alternative approaches and we demonstrated the advantages of our method. For the first and second applications, in particular, our method gave more parsimonious models with at least as good predictive behavior as that of competing ones. For the last application, our method permitted more flexibility in the volatility dynamics and provided an improvement for option pricing relative to the alternatives.

7.2 Future Work

We conclude by presenting directions for future research which we believe have special interest and are likely to extend the applicability of our method. Initially, we would like to try alternative estimation approaches for our model, and at a first level fully Bayesian estimation. The likelihood of our model given the hyperparameters is Gaussian, therefore conditionally conjugate priors can be used for the mean of the coefficients and the error variance. However, we still require appropriate priors for the smoothing parameters. Moreover, the posterior is analytically intractable, so we would have to resort to MCMC methods. One possible extension of the MCMC methodology would be the incorporation of reversible jump schemes between models with constant and varying coefficients. Another extension, which is particular to the approximation framework, would be to model number and location of basis as free parameters, similar to the Bayesian adaptive regression splines of DiMatteo et al. [28]. At a second level, we would like to perform estimation for our model using error distributions other than Gaussian. Dropping the normality assumption, our model can be viewed as a nonlinear, non-Gaussian state-space model where the function

evaluations are the unobserved states. Both the likelihood of the model and the posterior distribution of the coefficients become analytically intractable, but we can apply particle filtering for approximating them numerically, where again reduced rank approximations can speed up the process. Details of the filtering algorithm are given in Kitagawa [69] and methods for maximum likelihood estimation in this setting are given in Johansen et al. [65]. Although MCMC and particle filters are significantly more involved and computationally intensive than our empirical Bayes method, it would be instructive to compare the different estimation approaches, especially in the latter of non-Gaussian errors and for small data sets.

Another important area of investigation is that of alternative specifications for our model. First, we would like to experiment with different covariance kernels, besides the squared exponential. Our choice of kernel so far was motivated by convenience and the need to fit relatively smooth coefficient functions. An interesting possibility would be the use of nonstationary kernels, i.e. kernels with variable smoothness at different areas of the input space. There are at least two cases where this could be useful: one is for controlling the shape of the functions at regions where we have less information, like boundaries or areas where the regressors are close to zero, and the other is for estimating coefficients with abrupt changes, inspired by the wide applicability of TAR models. For the latter case, there are nonstationary examples like the neural network kernel (see section 4.2.3 of Rasmussen and Williams [96]) which permit the estimated function to have level shifts. In addition to covariance kernels, we would also like to experiment with different prior mean functions. In this respect we have only used constant functions, but in some cases it might be preferable to allow more flexibility. For example, a sigmoid prior would be useful for controlling the bias of the functional coefficient estimates separately outside the two ends of the observed range. Lastly, we would like to examine adaptive modeling, where the coefficient arguments are unknown linear combinations of a set of variables, with weights to be estimated from

the data. Such models appear in section 8.4 of Fan and Yao [36].

Moreover, we would like to elaborate on the nonlinear state space methodology that we have presented. There are still some issues which we have not addressed, such as measuring the complexity of a fitted model and developing a structured model selection procedure. We also want to test our method in a multivariate setting. A candidate application we have identified in econometrics, and which is suited for our method concerns, factor models for the term structure of interest rates. An example of this approach for linear SS models using Kalman filtering is given in Duan and Simonato [31], and we want to build nonlinear extensions of this. For financial data especially, we also want to extend our model for dealing with ARCH errors. As long as the distribution of the error is normal, our model can handle parametric forms of conditional heteroskedasticity. While preserving conditional normality, this feature can be incorporated in the filtering and likelihood computations and estimation can be addressed in an integrated way.

Finally, we would like to further explore the theoretical properties of our model, and there are at least three directions of research which we believe are worthwhile. The first concerns establishing convergence rates for the estimators, to complement our consistency result. These are useful for comparing nonparametric methods within smoothness classes of functions for the true coefficients. They are also an important step toward the second direction which relates to adaptive estimation. For our consistency result we assumed that the smoothness of each function is known beforehand, but this is an unrealistic assumption. Therefore,we would like to be able to evaluate or develop procedures for selecting the relevant smoothing parameters. Moreover, the fixed RKHS assumption for the true functions can be restrictive relative to general smoothness classes, so it would be interesting to investigate procedures where the smoothness decreases with the amount of data. Lastly,

we would also like to provide a consistency result for the model selection procedure we presented, possibly including the constancy test.

Appendix A

Reproducing Kernel Hilbert Spaces

We provide a brief introduction to reproducing kernel Hilbert spaces (RKHS) together with an overview of the basic results we use in Chapter 4. A more detailed exposition can be found in Berlinet and Thomas-Agnan [7] or Rasmussen and Williams [96]. We begin with the definition of a RKHS; let $\mathcal{K}$ be a Hilbert space of real functions with domain $\mathcal{X}$ and inner product $\langle \cdot, \cdot \rangle_{\mathcal{K}}$.

Definition A.1 (RKHS)**.** *The Hilbert space $\mathcal{K}$ is called a RKHS if all point evaluations are bounded linear functionals.*

An immediate consequence of Definition A.1, through the Riesz representation theorem, is that for every $x \in \mathcal{X}$ there is an element $k_x \in \mathcal{K}$ such that

$$f(x) = \langle f, k_x \rangle_{\mathcal{K}}, \quad \forall f \in \mathcal{K}.$$

We note that RKHS are *smooth* spaces of functions, in the sense that norm convergence implies pointwise convergence. To see this, let $\{f_n\}_{n \geq 1}$ be a sequence that converges to f

in $\mathcal{K}$, under the usual norm $\|f\|_{\mathcal{K}} = \sqrt{\langle f, f\rangle_{\mathcal{K}}}$. For any $x \in \mathcal{X}$, we have

$$\begin{aligned} |f_n(x) - f(x)| &= |\langle f_n - f, k_x\rangle_{\mathcal{K}}| \\ &\leq M\|f_n - f\|_{\mathcal{K}}, \quad M > 0 \end{aligned}$$

It is obvious from this that the usual L_2 space with Lebesgue measure is not a RKHS. We also give the definition of a reproducing kernel for a general Hilbert space.

Definition A.2 (Reproducing Kernel)**.** *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a reproducing kernel of the Hilbert space $\mathcal{K}$ if and only if*

i. For any $x \in \mathcal{X}$, $k(\cdot, x) \in \mathcal{K}$

ii. For any $x \in \mathcal{X}$ and $f \in \mathcal{K}$, $\langle f, k(\cdot, x)\rangle_{\mathcal{K}} = f(x)$

The second condition is called the *reproducing property*. An equivalent definition of a RKHS is that of a Hilbert space that possesses a reproducing kernel. For a given RKHS $\mathcal{K}$ we can find a reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by defining $k(x, x') = \langle k_x, k_{x'}\rangle_{\mathcal{K}}$ for all $x, x' \in \mathcal{X}$; it is easy to verify that k satisfies the conditions of Definition A.2. One important property of reproducing kernels is that they are positive definite functions, where such a functions are defined below.

Definition A.3 (Positive Definite Function)**.** *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite function if for all $n \geq 1$, $(a_1, \ldots, a_n) \in \mathbb{R}^n$ and $(x_1, \ldots, x_n) \in \mathcal{X}^n$*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0$$

The following theorem establishes a correspondence between RKHS and positive definite kernels.

Theorem A.4 (Moore-Aronszajn)**.** *For every positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ there exists a unique RKHS of real functions on $\mathcal{X}$, and vice versa.*

This allows us to define a RKHS by means of a positive definite kernel. Given a kernel k and its corresponding RKHS $\mathcal{K}$, we know that $\mathcal{K}$ contains the space of functions $\mathcal{K}_L = \{f : f(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i), \quad \forall n \geq 1,\ (a_1, \ldots, a_n) \in \mathbb{R}^n,\ (x_1, \ldots, x_n) \in \mathcal{X}^n\}$. In fact, $\mathcal{K}$ can be constructed as the Cauchy complement of $\mathcal{K}_L$ and for functions in $\mathcal{K}_L$ their norm is given by $\|f\|_{\mathcal{K}}^2 = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j k(x_i, x_j)$. A list of different kernels and their corresponding RKHS is given in [7], including standard examples from nonparametric regression like Sobolev spaces. Next, we look at the eigen decomposition of a reproducing kernel. First we define the eigenfunctions and eigenvalues of a kernel.

Definition A.5 (Kernel Eigenfunction)**.** *Let k be a positive definite kernel on a compact space $\mathcal{X}$ and μ be a strictly positive measure on $\mathcal{X}$. A function ϕ is said to be an eigenfunction of k, with associated eigenvalue λ, if*

$$\int_x k(x, x')\phi(x) d\mu(x) \quad = \lambda\phi(x')$$

In general, there are an infinite number of eigenfunctions and they can be chosen such that they are orthonormal w.r.t. μ. Mercer's theorem allows us to express the kernel k in terms of eigenfunctions and eigenvalues.

Theorem A.6 (Mercer)**.** *Let $\mathcal{X} \subset \mathbb{R}^m$ be compact, μ be a strictly positive Borel measure on $\mathcal{X}$ and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous positive definite kernel in $L_\infty(\mathcal{X}^2, \mu^2)$. Then there exist eigenfunctions $\phi_i \in L_2(\mathcal{X}, \mu)$ and associated eigenvalues $\lambda_i > 0$ such that*

i. $\int_{\mathcal{X}} \phi_i(x)\phi_j(x) d\mu(x) = \delta_{ij}$*, where δ_{ij} is the Kronecker delta*

ii. $\sum_{i=1}^\infty \lambda_i < \infty$

iii. $k(x, x') = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(x')$, *where the convergence of the series is absolute and uniform in* $\mathcal{X}^2$ *(*μ^2 *almost everywhere).*

These eigenfunctions form a basis for the RKHS and we can express any function $f \in \mathcal{K}$ as

$$f(\cdot) = \sum_{j=1}^{\infty} f_j \phi_i(\cdot)$$

where $\{f_j\}_{j=1}^{\infty}$ are given by $f_j = \int f(x)\phi_j(x)d\mu(x)$ and $\|f\|_{\mathcal{K}}^2 = \sum_j f_j^2/\lambda_j$.

Bibliography

[1] H. Akaike. "A New Look at the Statistical Model Identification". *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] D. W. K. Andrews. "Non-Strong Mixing Autoregressive Processes". *Journal of Applied Probability*, 21(4):930–934, 1984.

[3] C. F. Ansley, R. Kohn, and C.-M. Wong. "Nonparametric Spline Regression with Prior Information". *Biometrika*, 80(1):75–80, 1993.

[4] N. S. Balke and T. B. Fomby. "Threshold Cointegration". *International Economic Review*, 38(3):627–645, 1997.

[5] A. R. Barron. "The Strong Ergodic Theorem for Densities: Generalized Shannon-McMillan-Breiman Theorem". *Annals of Probability*, 13(4):1292–1303, 1985.

[6] G. Bekaert and G. Wu. "Asymmetric Volatility and Risk in Equity Markets". *The Review of Financial Studies*, 13(1):1–42, 2000.

[7] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

[8] T. Björk. *Arbitrage Theory in Continuous Time*. Oxford University Press, 1998.

[9] D. Bosq. *Lecture Notes in Statistics: Nonparametric Statistics for Stochastic Processes*. Springer-Verlag, New York, 1998.

[10] P. Bougerol and N. Picard. "Strict Stationarity of Generalized Autoregressive Processes". *The Annals of Probability*, 20(4):1714–1730, 1992.

[11] L. Breiman and J. H. Friedman. "Estimating Optimal Transformations for Multiple Regression and Correlation". *Journal of the American Statistical Association*, 80(391):580–598, 1985.

[12] R. J. Brenner and K. F. Kroner. "Arbitrage, Cointegration, and Testing the Unbiasedness Hypothesis in Financial Markets". *The Journal of Financial and Quantitative Analysis*, 30(1):23–42, 1995.

[13] D. R. Brillinger. "An Introduction to Polyspectra". *The Annals of Mathematical Statistics*, 36(5):1351–1374, 1965.

[14] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer-Verlag, New-York, 2nd edition, 1991.

[15] R. L. Brown, J. Durbin, and J. M. Evans. "Techniques for Testing the Constancy of Regression Relationships over Time". *Journal of the Royal Statistical Society, Ser. B*, 37(2):149–192, 1975.

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

[17] M. J. Campbell and A. M. Walker. "A Survey of Statistical Work on the Mackenzie River Series of Annual Canadian Lynx Trappings for the Years 1821-1934 and a New Analysis". *Journal of the Royal Statistical Society, Ser. A*, 140(4):411–431, 1977.

[18] K. S. Chan. "Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregressive Model". *The Annals of Statistics*, 21(1):520–533, 1993.

[19] K. S. Chan, J. D. Petruccelli, H. Tong, and S. W. Woolford. "A Multiple-Threshold AR(1) Model". *Journal of Applied Probability*, 22(2):267–279, 1985.

[20] R. Chen. *Two Classes of Non-linear Time Series*. PhD thesis, Carnegie Mellon University, Dept. of Statistics, 1990. Unpublished PhD Thesis.

[21] R. Chen and L.-M. Liu. "Functional Coefficient Autoregressive Models: Estimation and Tests of Hypotheses". *Journal of Time Series Analysis*, 22(2):151–173, 2001.

[22] R. Chen and R. S. Tsay. "Functional-Coefficient Autoregressive Models". *Journal of the American Statistical Association*, 88(421):298–308, 1993.

[23] R. Chen and R. S. Tsay. "Nonlinear Additive ARX Models". *Journal of the American Statistical Association*, 88(423):955–967, 1993.

[24] T. Choi. *Posterior Consistency in Nonparametric Regression Problems under Gaussian Process Priors*. PhD thesis, Carnegie Mellon University, Dept. of Statistics, 2005.

[25] N. A. C. Cressie. *Statistics for Spatial Data*. Wiley, 1993.

[26] Y. A. Davydov. "Convergence of Distributions Generated by Stationary Stochastic Processes". *Theory of Probability and its Applications*, 13(4):691–696, 1968.

[27] P. Dennis and S. Mayhew. "Risk-Neutral Skewness: Evidence from Stock Options". *The Journal of Financial and Quantitative Analysis*, 37(3):471–493, 2002.

[28] I. DiMatteo, C. R. Genovese, and R. E. Kass. "Bayesian Curve-Fitting with Free-Knot Splines". *Biometrika*, 88(4):1055–1071, 2001.

[29] A. Doucet, N. D. Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2001.

[30] J.-C. Duan. "The GARCH Option Pricing Model". *Mathematical Finance*, 5(1):13–32, 1995.

[31] J.-C. Duan and J.-G. Simonato. "Estimating and Testing Exponential-Affine Term Structure Models by Kalman Filter". *Review of Quantitative Finance and Accounting September*, 13(2):111–35, 1999.

[32] J. Durbin and S. J. Koopman. "Monte Carlo Maximum Likelihood Estimation for Non-Gaussian State Space Models". *Biometrika*, 84(3):669–684, 1997.

[33] B. Efron and C. Morris. "Stein's Estimation Rule and Its Competitors–An Empirical Bayes Approach". *Journal of the American Statistical Association*, 68(341):117–130, 1973.

[34] R. F. Engle. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation". *Econometrica*, 50(4):987–1007, 1982.

[35] R. F. Engle and C. W. J. Granger. "Co-Integration and Error Correction: Representation, Estimation, and Testing". *Econometrica*, 55(2):251–276, 1987.

[36] J. Fan and Q. Yao. *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag, New-York, 2003.

[37] C. S. Forbes, G. R. J. Kalb, and P. Kofman. "Bayesian Arbitrage Threshold Analysis". *Journal of Business & Economic Statistics*, 17(3):364–372, 1999.

[38] J.-P. Fouque, G. Papanicolaou, and R. Sircar. *Derivatives in Financial Markets with Stochastic Volatility*. Cambridge University Press, 2000.

[39] S. Frühwirth-Schnatter. "Recursive Residuals and Model Diagnostics for Normal and Non-normal State Space Models". *Journal of Environmental and Ecological Statistics*, 3(4), 1996.

[40] R. Gerlach, C. Carter, and R. Kohn. "Diagnostics for Time Series Analysis". *Journal of Time Series Analysis*, 20(3):309–330, 1999.

[41] J. Geweke and N. Terui. "Bayesian Threshold Autoregressive Models for Nonlinear Time Series". *Journal of Time Series Analysis*, 14:441–454, 1993.

[42] Z. Ghahramani and S. Roweis. "Learning Nonlinear Dynamical Systems Using an EM Algorithm". In M. S. Kearns, S. A. Solla, and D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*, pages 431–437. The MIT Press, 1999.

[43] S. Ghosal and A. van der Vaart. "Convergence Rates of Posterior Distributions for Non-IID Data". *The Annals of Statistics*, 35(1):192–223, 2007.

[44] M. Gibbs and D. MacKay. "Efficient implementation of Gaussian processes". Unpublished manuscript. Cavendish Laboratory, Cambridge, UK (available at `http://citeseer.ist.psu.edu/gibbs97efficient.html`) , 1997.

[45] A. Girard, C. E. Rasmussen, J. Quiñonero-Candela, and R. Murray-Smith. "Gaussian Process with Uncertain Inputs - Application to Multiple-Step-Ahead Time Series Forecasting". In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 529–536. The MIT Press, 2003.

[46] G. R. Glosten, R. Jagannathan, and D. E. Runkle. "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks". *The Journal of Finance*, 48(5):1779–1801, 1993.

[47] R. Group. *RiskMetrics - Technical Document*. J.P. Morgan/Reuters, 1996. (available at `http://www.riskmetrics.com/publications/techdocs/rmcovv.html`).

[48] G. K. Grunwald, R. J. Hyndman, L. Tedesco, and R. L. Tweedie. "Non-Gaussian Conditional Linear AR(1) Models". *Australian and New Zealand Journal of Statistics*, 42(4):479–495, 2000.

[49] V. Haggan and T. Ozaki. " Modelling Nonlinear Random Vibrations Using an Amplitude-Dependent Autoregressive Time Series Model". *Biometrika*, 68(1):189–196, 1981.

[50] J. D. Hamilton. "Analysis of Time Series Subject to Changes in Regime". *Journal of Econometrics*, 45(1-2):39–70, 1990.

[51] W. Härdle, H. Lütkepohl, and R. Chen. "A Review of Nonparametric Time Series Analysis". *International Statistical Review*, 65(1):49–72, 1997.

[52] A. Harvey, E. Ruiz, and N. Shephard. "Multivariate Stochastic Variance Models". *The Review of Economic Studies*, 61(2):247–264, 1994.

[53] A. Harvey and N. Shephard. "An Asymmetric Stochastic Volatility Model for Asset Returns". *Journal of Business and Economic Statistics*, 14(4):429–434, 1996.

[54] A. C. Harvey. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge University Press, 1991.

[55] T. Hastie and R. Tibshirani. "Generalized Additive Models". *Statistical Science*, 1(3):297–310, 1986.

[56] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Monographs on Statistics and Applied Probability. Chapman and Hall, 1990.

[57] T. Hastie and R. Tibshirani. "Varying-Coefficient Models". *Journal of The Royal Statistical Society, Ser. B*, 55(4):757–796, 1993.

[58] T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York, 2001.

[59] S. L. Heston. "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options". *The Review of Financial Studies*, 6(2):327–343, 1993.

[60] J. Z. Huang and H. Shen. "Functional Coefficient Regression Models for Non-Linear Time Series: A Polynomial Spline Approach". *Scandinavian Journal of Statistics*, 31:515–534, 2004.

[61] J. Hull and A. White. "The Pricing of Options on Assets with Stochastic Volatilities". *The Journal of Finance*, 42(2):281–300, 1987.

[62] C. Inclán and G. C. Tiao. "Use of Cumulative Sums of Squares for Retrospective Detection of Changes in Variance". *Journal of the American Statistical Association*, 89(427):913–923, 1994.

[63] C. M. Jarque and A. K. Bera. "A Test for Normality of Observations and Regression Residuals". *International Statistical Review*, 55(2):163–172, 1987.

[64] F. Jianqing and W. Zhang. "Statistical Estimation in Varying Coefficient Models". *The Annals of Statistics*, 27(5):1491–1518, 1999.

[65] A. M. Johansen, A. Doucet, and M. Davy. "Particle Methods for Maximum Likelihood Estimation in Latent Variable Models". *Statistics and Computing*, 18(1):47–57, 2008.

[66] I. G. Kawaller, P. D. Koch, and T. W. Koch. "The Temporal Price Relationship Between S&P 500 Futures and the S&P 500 Index". *The Journal of Finance*, 42(5):1309–1329, 1987.

[67] G. S. Kimeldorf and G. Wahba. "A Correspondence Between Bayesian Estimation on Stochastic Processes and Smoothing by Splines". *The Annals of Mathematical Statistics*, 41(2):495–502, 1970.

[68] G. S. Kimeldorf and G. Wahba. "Some Results on Tchebycheffian Spline Functions". *Journal of Mathematical Analysis and Applications*, 33:82–95, 1971.

[69] G. Kitagawa. "Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models". *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

[70] G. Koop, M. H. Pesaran, and S. M. Potter. "Impulse response analysis in nonlinear multivariate models". *Journal of Econometrics*, 74(1):119–147, 1996.

[71] N. Lawrence, M. Seeger, and R. Herbrich. "Fast Sparse Gaussian Process Methods: The Informative Vector Machine". In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 609–616. The MIT Press, 2003.

[72] T. C. Lin and M. Pourahmadi. "Nonparametric and Non-linear Models and Data Mining in Time Series: a Case Study on the Canadian Lynx Data". *Journal of the Royal Statistical Society, Ser. C*, 47(2):187–201, 1998.

[73] Y. Lin and L. D. Brown. "Statistical Properties of the Method of Regularization with Periodic Gaussian Reproducing Kernel". *The Annals of Statistics*, 32(4):1723–1743, 2004.

[74] D. V. Lindley and A. F. M. Smith. "Bayes Estimates for the Linear Model". *Journal of the Royal Statistical Society, Ser. B*, 34(1):1–41, 1972.

[75] O. Linton and J. P. Nielsen. "A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration". *Biometrica*, 82(1):93–100, 1995.

[76] R. S. Lipster and A. N. Shiryaev. *Statistics of Random Processes II: Applications*. Springer-Verlag, 2001.

[77] G. M. Ljung and G. E. P. Box. "On a Measure of Lack of Fit in Time Series Models". *Biometrika*, 65(2):297–303, 1978.

[78] H. Lütkepohl. *Introduction to Multiple Time Series Analysis*. Springer-Verlag, 1991.

[79] M. Martens, P. Kofman, and T. C. F. Vorst. "A Threshold Error-Correction Model for Intraday Futures and Index Returns". *Journal of Applied Econometrics*, 13(3):245–263, 1998.

[80] E. Masry and J. Fan. "Local Polynomial Estimation of Regression Functions for Mixing Processes". *Scandinavian Journal of Statistics*, 24(2):165–179, 1997.

[81] E. Masry and D. Tjøstheim. "Additive Nonlinear ARX Time Series and Projection Estimates". *Econometric Theory*, 13(2):214–252, 1997.

[82] A. I. McLeod and W. K. Li. "Diagnostic Checking ARMA Time Series Models Using Squared-Residual Autocorrelations". *Journal of Time Series Analysis*, 4:269–273, 1983.

[83] S. P. Meyn and R. L. Tweedie. "Stability of Markovian Processes I: Criteria for Discrete-Time Chains". *Advances in Applied Probability*, 24(3):542–574, 1992.

[84] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, London, 1993.

[85] P. A. P. Moran. "The Statistical Analysis of the Canadian Lynx Cycle. I. Structure and Prediction". *Australian Journal of Zoology*, 1:163–173, 1953.

[86] D. B. Nelson. "Conditional Heteroskedasticity in Asset Returns: A New Approach". *Econometrica*, 59(2):347–370, 1991.

[87] D. F. Nicholls and B. G. Quinn. *Random Coefficient Autoregressive Models: An Introduction*. Springer-Verlag, New York, 1982.

[88] A. O'Hagan. "Curve Fitting and Optimal Design for Prediction" (with discussion). *Journal of The Royal Statistical Society, Ser. B*, 40(1):1–42, 1978.

[89] S. D. Oman. "A Different Empirical Bayes Interpretation of Ridge and Stein Estimators". *Journal of the Royal Statistical Society, Ser. B*, 46(3):544–557, 1984.

[90] OptionMetrics. *OptionMetrics Data Manual*, 2006. (available at `http://wrds.wharton.upenn.edu/ds/optionm/manuals/IvyDBReference.pdf`).

[91] C. J. Paciorek and M. J. Schervish. "Nonstationary Covariance Functions for Gaussian Process Regression". In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 609–616. The MIT Press, 2004.

[92] J. D. Petruccelli. "A Comparison of tests for SETAR-type Nonlinearity in Time Series". *Journal of Forecasting*, 9:25–36, 1990.

[93] J. D. Petruccelli and N. Davies. "A Portmanteau Test for Self-Exciting Threshold Autoregressive-Type Nonlinearity in Time Series". *Biometrika*, 73(3):687–694, 1986.

[94] J. Quiñonero Candela and C. E. Rasmussen. "A Unifying View of Sparse Approximate Gaussian Process Regression". *Journal of Machine Learning Research*, 6:1939–1959, 2005.

[95] T. S. Rao. "On the Theory of Bilinear Time Series Models". *Journal of the Royal Statistical Society, Ser. B*, 43(2):244–255, 1981.

[96] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge MA, 2006.

[97] M. Rosenblatt. "Remarks on a Multivariate Transformation". *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

[98] G. Sandmann and S. J. Koopman. "Estimation of Stochastic Volatility Models via Monte Carlo Maximum Likelihood". *Journal of Econometrics*, 87:271–301, 1998.

[99] B. Schölkopf, R. Herbrich, and A. J. Smola. "A Generalized Representer Theorem". In *Proceedings of the 14th Annual Conference on Computational Learning Theory, 2001*, volume 2111 of *Lecture Notes in Computer Science*, pages 416–426. Springer Berlin / Heidelberg, 2001.

[100] G. Schwarz. "Estimating the Dimension of a Model". *The Annals of Statistics*, 6(2):461–464, 1978.

[101] M. Seeger. *Bayesian Gaussian Process Models: PAC-Bayesian Generalization Error Bounds and Sparse Approximations*. PhD thesis, University of Edinburgh, School of Informatics, 2003. (available at `http://www.kyb.tuebingen.mpg.de/bs/people/seeger/papers/thesis.pdf`).

[102] M. Seeger, C. K. I. Williams, and N. Lawrence. "Fast Forward Felection to Speed Up Sparse Gaussian Process Regression". In C. Bishop and B. J. Frey, editors, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Society for Artificial Intelligence and Statistics, 2003.

[103] M. Seeger, C. K. I. Williams, and N. D. Lawrence. "Fast Forward Selection to Speed Up Sparse Gaussian Process Regression". In C. Bishop and B. J. Frey, editors, *Proceedings of the 9^{th} International Workshop on Artificial Intelligence and Statistics*. Society for Artificial Inteligence and Statistics, 2003.

[104] S. S. Shapiro and M. B. Wilk. "An Analysis of Variance Test for Normality (Complete Samples)". *Biometrika*, 52(3):591–611, 1965.

[105] X. Shen and L. Wasserman. "Rates of Convergence of Posterior Distributions". *The Annals of Statistics*, 29(3):687–714, 2001.

[106] R. Shumway and D. Stoffer. "An Approach to Time Series Smoothing and Forecasting Using the EM Algorithm". *Journal of Time Series Analysis*, 3:253–264, 1982.

[107] B. W. Silverman. "Some Aspects of the Spline Smoothing Approach to Non-Parametric Regression Curve Fitting" (with discussion). *Journal of the Royal Statistical Society, Ser. B*, 47:1–52, 1985.

[108] J. Q. Smith. "Diagnostic Checks of non-Standard Time Series Models". *Journal of Forecasting*, 4(3):283–291, 1985.

[109] M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, New York, 1999.

[110] H. R. Stoll and R. E. Whaley. "The Dynamics of Stock Index and Stock Index Futures Returns". *The Journal of Financial and Quantitative Analysis*, 25(4):441–468, 1990.

[111] T. Teräsvirta. "Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models". *Journal of the American Statistical Association*, 89(425):208–218, 1994.

[112] D. Tjøstheim. "Non-Linear Time Series and Markov Chains". *Advances in Applied Probability*, 22(3):587–611, 1990.

[113] H. Tong. "Some Comments on the Canadian Lynx Data". *Journal of the Royal Statistical Society, Ser. A*, 1977.

[114] H. Tong. *Non-linear Time Series: A Dynamical Systems Approach*. Oxford University Press, Oxford, 1990.

[115] H. Tong and K. S. Lim. "Threshold Autoregression, Limit Cycles and Cyclical Data". *Journal of the Royal Statistical Society, Ser. B*, 42(3):245–292, 1980.

[116] R. S. Tsay. "Testing and Modeling Threshold Autoregressive Processes". *Journal of the American Statistical Association*, 84(405):231–240, 1989.

[117] R. S. Tsay. "Testing and Modeling Multivariate Threshold Models". *Journal of the American Statistical Association*, 93(443):1188–1202, 1998.

[118] G. Wahba. "Bayesian "Confidence Intervals" for the Cross-Validated Smoothing Spline". *Journal of the Royal Statistical Society, Ser. B*, 45(1):133–150, 1983.

[119] J. M. Wang, D. J. Fleet, and A. Hertzmann. "Gaussian Process Dynamical Models". In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 1441–1448. The MIT Press, 2006.

[120] L. Wasserman. *All of Nonparametric Statistics*. Springer-Verlag, 2006.

[121] M. West and J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York, 2nd edition, 1997.

[122] C. K. I. Williams and M. Seeger. "Using the Nyström Method to Speed Up Kernel Machines". *Advances in Neurel Information Processing Systems*, 13:682–688, 2001.

[123] A. Zellner. *An Introduction to Bayesian Inference in Econometrics*. John Wiley & Sons, New York, 1971.