# Gaussian Processes for Functional-Coefficient Autoregressive Models

Ph.D Thesis Proposal

Sotirios Damouras
Department of Statistics, Carnegie Mellon University

February 13, 2007

**Abstract**

This work is concerned with nonlinear time series models and, in particular, with nonparametric models for the dynamics of the mean of the time series. We build on the functional-coefficient autoregressive (FAR) model of Chen and Tsay (1993) which is a generalization of the autoregressive (AR) model where the coefficients are varying and are given by functions of the lagged values of the series. We adopt a Bayesian approach for nonparametric functional estimation, modelling the coefficient functions as Gaussain Processes (GPs). We investigate practical implementation issues for our model and describe efficient ways to conduct estimation. We illustrate our proposed method by a study of a simulated and a real data example and we discuss how it can improve over existing estimation techniques for FAR models. Finally, we propose directions for future work which will be aimed, mainly, towards obtaining theoretical results for the procedure, deriving numerical approximation methods for large data sets and assessing the fit of the model.

## 1 Introduction

Linear time series models have played a dominant role in the field of time series analysis. Since its formal development by Box and Jenkins, the autoregressive moving average (ARMA) model has been the most widely used and extensively studied model for time series; a more or less full account of its theoretical and practical aspects is given in [4]. Real time series, however, can exhibit an entire spectrum of nonlinear behavior that is not adequately described by ARMA models. This has spurred interest in the field of nonlinear time series and during the last couple of decades there has appeared a plethora of new models and techniques. Most of these nonlinear models were parametric in nature, but recently, and with the help of increased computational power, attention has shifted to nonparametric models.

The class of nonlinear time series models we focus on are intended to describe the dynamics of the mean; we do not consider models for the variance such as the autoregressive conditional heteroskedastic (ARCH) model and its variants. In the following section, we give a brief description of the relevant nonlinear time series models; a more detailed review of these is given in Härdle, Lütkepohl and Chen [13] and in Fan and Yao [9] (ch. 4 and 8). All of the models we will be looking at fall under the general class of the nonlinear autoregressive (NLAR) model, which is given by

$$x_t = f(x_{t-1}, \ldots, x_{t-p}) + \epsilon_t, \tag{1}$$

where $\{\epsilon_t\}$ is a white noise sequence, independent of $\{x_t\}$. The function $f(\cdot)$ represents the conditional mean of $x_t$ given the past values $(x_{t-1}, \ldots, x_{t-p})$ of the series. In the parametric setting, we assume the function $f(\cdot)$ belongs to a specific parametric class and is characterized by a fixed finite number of parameters. Choosing a particular parametric form usually requires some knowledge of the behavior of the data and can lead to modelling biases if the class of functions is too narrow. Neural networks are broad parametric classes with universal function approximation properties, but they tend to be overparametrized and their estimation is problematic. For this reason, nonparametric estimation techniques have been proposed, which allow $f$ to belong to some flexible class of functions. The most popular of these techniques for time series are kernel regression and local polynomial regression (see Stone [25] for a description of the procedures).

Despite its generality and flexibility, the model in (1) has the important disadvantage that it suffers from the "curse of dimensionality". The term is used to describe the difficulties and compromises in accuracy that arrise from nonparametric estimation in high dimensions and is a well known phenomenon (see Hastie, Tibshirani and Friedman [16]). This is particularly relevant in our case if the autoregressive order $p$ of the model is high, since the function $f$ maps $\mathbb{R}^p \to \mathbb{R}$. In the context of time series there is an added complication; due to the dynamic nature of the data it is often the case that the vector $[x_{t-1}, \ldots, x_{t-p}]$ (the argument to the function $f$) does not cover the entire domain of $f$, a property which is important for estimation. Dynamical systems can have trajectories such that $[x_{t-1}, \ldots, x_{t-p}]$ moves around a lower dimensional manifold of $\mathbb{R}^p$, as in systems with limit cycles for example.

Due to the above issues, it is reasonable to restrict model (1) to a more parsimonious class. A useful and popular approach in the nonparametric regression setting is to assume an additive form for $f$. This leads to the generalized additive model (GAM) of Hastie and Tibshirani [14]. In the context of time series, it becomes the nonlinear additive autoregressive (NLAAR) model, given by

$$x_t = f_1(x_{t-1}) + \ldots + f_p(x_{t-p}) + \epsilon_t. \tag{2}$$

Estimation for additive models in regression is conducted through backfitting algorithms [14], but for time series data which are serially correlated, the convergence of the algorithms can be slow

and problematic, as reported in Chen and Tsay [8]. In order to rectify this, Linton and Nielsen [20] suggested an estimation procedure for the functions based on marginal integration. Even so, additive models can still be susceptible to non-identifibility issues.

A model which is more favored in time series and which is also the focus of the current work, is the functional-coefficient autoregressive (FAR) model of Chen and Tsay [7]. It is the time series analogue of the varying-coefficient regression model of Hastie and Tibshirani [15] and is given by

$$x_t = f_1(U_t^{(1)})x_{t-1} + \ldots + f_p(U_t^{(p)})x_{t-p} + \epsilon_t. \tag{3}$$

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, so they can and will depend on lagged values of $\{x_t\}$. Thus, the FAR model (3) does not strictly nest the NLAAR model (2). In fact, it is usually the case that all functional coefficients share the same argument. One 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 values $x_{t-i}$. The FAR class also has as parametric subcases the threshold autoregressive (TAR) model of Tong [26] and the exponential autoregressive (EXPAR) model of Haggan and Ozaki [11], which we review in the following section.

Two popular nonparametric estimation approaches for the FAR model have been suggested in the literature and they are the arranged local regression (ALR) procedure of Chen and Tsay [7] and the local linear regression (LLR) procedure of Cai, Fan and Yao [5]. Both of them belong to the broader family of kernel smoothing techniques. This is also the cause of a shortcoming for these techniques, that they cannot be used unless all functional coefficient arguments are the same, i.e. $U_t^{(1)} = \ldots = U_t^{(p)}$ in (3). The reason for this is that estimation is performed locally based on some weighting scheme, where locallity in the data is defined w.r.t. the arguments of the functions $f_i$. Therefore, all functional coefficients must share the same arguments. As a consequence, all functional coefficient estimates also share the same smoothing parameter, or bandwidth, and this can compromise the accuracy of estimation. In particular, the function estimates will tend to have the same smoothness. More recently, another nonparametric approach was suggested by Huang and Shen [17] which relies on spline methods and has the potential to overcome the previous problems. The authors use a scheme which relies on the number and positions of the knots to achieve smoothing, instead of using regularization, and this can have have some undesirable implications. We discuss these issues in more detail in the next section.

In section 3 we present our proposed modelling and estimation approach to the FAR model. We adopt the same dynamics as in (3) but we let the functional coefficients be random and follow Gaussian Processes (GPs). We describe the resulting estimation and prediction procedures and discuss the prior specification of the model. In section 4 we present preliminary results from the

application of our method to a simulated and a real data set. We also compare the results to those of competing models and we demonstrate how our method can overcome some of their limitations. Finally, in section 5 we propose directions for our future efforts that will complement this work.

## 2 Review of Relevant models

We start the exposition with the conceptually most simple model, the TAR model. Tong [26] gives an extensive coverage of the model properties and estimation procedures, with many examples. The model assumes serial autocorrelation in the data, much like an AR process, but where the coefficients of autoregression are non-constant and change levels according to different regimes. For example, consider a TAR model of the first order with two regimes. Let $\alpha_1^{(0)}, \alpha_1^{(1)}$ be the two levels of the autoregressive coefficient, and $I_t \in \{0, 1\}$ be an $\mathcal{F}_{t-1}$ measurable indicator random variable, unspecified for now. The dynamics of $x_t$ are given by:

$$x_t = \begin{cases} \alpha_1^{(0)} x_{t-1} + \epsilon_t, & \text{if } I_t = 0 \\ \alpha_1^{(1)} x_{t-1} + \epsilon_t, & \text{if } I_t = 1 \end{cases} \tag{4}$$

It is conceptually easy to extend the model to include higher order lagged values of $x_t$ and more regimes. The variable $I_t$ controls the regime and there exist many different flavors of the TAR model, depending on how it is specified. The most common is to have $I_t$ be a function of some lagged value of the process itself, i.e. $I_t = f(x_{t-d})$, where $d \geq 1$ is called the delay parameter. This model is also known as a self-exciting threshold autoregressive model (SETAR); other possibilities are to allow $I_t$ to depend on exogenous variables, or to model $I_t$ as an independent latent Markov process as in [12]. For the remainder, we will use the term TAR to refer to the SETAR model. The regimes in the TAR model are specified according to the region where the lagged value $x_{t-d}$ lies. For the model given in (4), for example, we have to specify a set $R \subset \mathbb{R}$ such that $I_t = \mathbb{I}(x_{t-d} \in R)$, where $\mathbb{I}(\cdot)$ is the indicator function. We can rewrite the model in (4) as

$$x_t = \begin{cases} \alpha_1^{(0)} x_{t-1} + \epsilon_t, & \text{if } x_{t-d} \in R \\ \alpha_1^{(1)} x_{t-1} + \epsilon_t, & \text{if } x_{t-d} \in R^c \end{cases}$$

We can extend the model by allowing higher order autoregression, more regimes and more variables for defining the regimes. For practical purposes, higher order autoregression can be treated easily. But if we increase the number of regimes or, especially, the number of variables that define the regime, then estimation becomes very difficult. Notice that we decide the regime in terms of partitions of $\mathbb{R}^q$, where $q$ is the dimension of the variable based on which we define the regime. Estimation for TAR models is based on minimizing aggregate sums of squares of linear regreessions accross different regimes, but as pointed out by Tong (see [26], page 303) the objective function is

not smooth with respect to the parameters that define the regimes, i.e. the regime boundaries. This poses problems in estimation when we have many regimes or complicated partition boundaries for high dimensional spaces. These limitations have restricted practical applications of TAR models mainly to situations where there are a few, say $k$, regimes, and $I_t \in \{1, \ldots, k\}$ is decided by a single lagged variable $x_{t-d}$ in the following way

$$I_t \;=\; \begin{cases} 1, & \text{if } x_{t-d} < r_1 \\ i, & \text{if } r_{i-1} \le x_{t-d} < r_i, \;\; i = 2, \ldots, k-1 \\ k, & \text{if } r_{k-1} \le x_{t-d} \end{cases}, \tag{5}$$

where $r_1 < \ldots < r_{k-1}$.

The FAR model (3) is a generalization of the TAR model, and it allows the autoregressive coefficients to be given by general functions $f_i$ with arguments $U_t^{(i)}$, which are $\mathcal{F}_{t-1}$ measurable. As before, we can think of $U_t$ as a collection of lagged values of the process. The TAR model can be thought of as a piecewise linear approximation to the FAR model, if the regime is allowed to depend on all of the $U_t^{(i)}$'s. Another parametric model which falls under the FAR class is the exponential autoregressive (EXPAR) model of Haggan and Ozaki [11], which is given by

$$x_t = \sum_{i=1}^{p} \left(\alpha_i + (\beta_i + \gamma_i x_{t-d}) \exp\{-\theta_i x_{t-d}^2\}\right) x_{t-i} + \epsilon_t. \tag{6}$$

The TAR model, despite its simplicity, is capable of capturing a broad range of nonlinear behavior and is, in fact, very popular. The EXPAR model was developed to describe amplitude dependent frequency phenomena which are typical in random vibration models and its use is not as wide.

We can allow more general, nonparametric functional forms for the parameters in (3). A first nonparametric estimation approach was proposed by Chen and Tsay [7], where the authors use the ALR procedure. The main idea dates back to Tsay [27], and it involves arranging the data points $\{x_t, [x_{t-1}, \ldots, x_{t-p}]\}$ with respect to $U_t$, which is now the common argument to all functions. In order to estimate the functional coefficients at a point $U$, this method uses a window around $U$ and performs a linear regression over the arranged data points with $U_t$ falling within that window

$$x_t = a_1 x_{t-1} + \ldots + a_p x_{t-p}; \;\; \forall t \text{ such that } U_t \in [U-h, U+h]. \tag{7}$$

The estimates of the functional coefficients at $U$ are then given by the fitted regression parameters: $\hat{f}_i(U) = \hat{a}_i$. 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 $U_t$. 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 [7] do not actually work with the resulting estimated functions, but, instead, they advocate using them to infer a parametric functional form for the coefficients and then use all the data to estimate those new parameters by least squares. They 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 [5]. They estimate the coefficient functions by employing the LLR procedure. The locallity is again induced by the common function argument variable $U_t$, but the procedure is different since a kernel is used to weight the neighboring observations instead of a window. Moreover, the authors use a linear function to approximate the coefficient functions localy around $U_0$ in the following fashion

$$f_i(U) \approx a_i + b_i(U - U_0).$$

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 maximize the weighted sum of squares

$$\sum_t \left( x_t - \sum_{i=1}^{p} \big(a_i + b_i(U_t - U)\big) x_{t-i} \right)^2 K_h(U_t - U), \tag{8}$$

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 prove, under certain conditions, asymptotic normality of the estimates at a given point and they suggest a procedure for selecting $h$ using multifold cross-validation based on the rate of convergence.

We mention here the two limitations of these approaches which we alluded to before and which become apparent by now. First, all of the coefficient functions need to share the same argument $U_t$ and second, all of the estimated functions share the same smoothing parameter (either the window or the the kernel bandwidth). In practice, these can lead to important modelling and estimation disadvantages. For the LLR procedure in the varying coefficient setting, this problem was aknowledged by Fan and Zhang [18], 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 afforementioned problem, but they show that the thus obtained functional coefficient estimates achieve optimal rates of convergence. Their results do not apply to time series, however, so this method was not suggested for the FAR model.

For the varying coefficient model in regression, other estimation approaches have been proposed which overcome these limitations. Hastie and Tibshirani [15] allow the coefficient functions to have

different arguments and estimate the model by minimizing least squares plus a smoothness penalty on the functions $f_i$. The penalty parameter is different for each function, therefore allowing variable smoothness. This approach leads to a cubic spline specification for each function $f_i$, whith knots at the locations where each argument is observed. They show that estimating the spline coefficients is computationally expensive in this setting, so they propose estimating each function one at a time, recursively, using a backfitting algorithm. It is possible to use a similar procedure in the time series setting for the FAR model by using penalized conditional least squares. This allows as to carry out estimation similarly to a regression problem, however, the performance and the properties of the estimators are completely different between the two settings. The errors used in least squares estimation are not independent and, as we already mentioned, backfitting procedures can be unreliable. A different method was proposed by Huang and Shen [17] which still uses splines and 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 model for the functional coefficients is

$$f_i(U) = \sum_{j=1}^{k_i} \alpha_{i,j} B_{i,j}(U), \quad i = 1, \ldots, p, \tag{9}$$

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 esimated by minimizing the conditional sum of squares

$$\sum_t \left( x_t - \sum_{i=1}^{p} \left( \sum_{j=1}^{k_i} \alpha_{i,j} B_{i,j}(U_t^{(i)}) \right) x_{t-i} \right)^2 . \tag{10}$$

Note that without regularization, the solution is given by a simple linear regression and this can cause problems if the system is ill-conditioned. There remains the question about the positions and the number of the knots. The authors suggest using equally spaced knots and AIC to decide their number, motivated by the performance of their method in simulation experiments. The authors argue that for most applications the number of knots needed will be small, usually less than five for each function, and as a result their estimation procedure will be fast. Note that the use of AIC also bypasses the need for cross-validation since no regularization is used. We should, however, point out that the choice of equally spaced knots is ad-hoc and that time series data are unevenly distributed more often than not. Throughout their paper, the authors did not use different arguments for the functional coefficients and they sometimes used 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.

## 3 Proposed Model

In our setting, the functions $f_i$ are actually random and are modelled by GPs. We assume the same dynamics as in (3) and we put a prior GP over the functions

$$\begin{aligned} x_t &= f_1(U_t^{(1)})x_{t-1} + \ldots + f_p(U_t^{(p)})x_{t-p} + \epsilon_t \\ \epsilon_t &\sim \mathcal{N}(0, \sigma^2) \\ f_i &\sim \mathcal{GP}(\mu_i, C_i), \;\; i = 1, \ldots, p, \end{aligned} \tag{11}$$

where $\mu_i(\cdot)$, $C_i(\cdot, \cdot)$ are the mean and covariance functions of $f_i$, and the functions $f_i$ are a-priori independent. In the above equation we assumed the errors $\epsilon_t$ are i.i.d. normal in order to use the conjugacy property of normals.

This approach has also been suggested in a regression setting by O'Hagan [22] as early as 1978. This was actually one of the first applications of GPs for nonparametric estimation, a field which has been expanding rapidly in recent years in the statistics and machine learning communities (the recent book by Rasmussen and Williams [23] gives a very nice overview of GPs used for regression and classification). The author looked at a varying-coefficient regression model where the functional coefficients are modelled by GPs. He provides posterior and predictive distributions and looks at optimal experimental designs for the model from a decision theoretic standpoint. With regard to modelling, the curve fitting approach proposed in [22] is very similar in spirit to our approach. The author, however, did not investigate closely the prior specification of the model and particularly the choice of covariance structure which controls the smoothing. He use a Kronecker product form for the covariance matrix which requires the same arguments and the same degree of smoothing for all functions. Moreover, he does not discuss practical implementation issues which we also address.

Assume we have a fixed mean and covariance function for our prior specification and a sample of size $T$ from model (11). Also, let $f_{i,t} = f_i(U_t^{(i)})$ and define the following

$$\boldsymbol{f}^\top = [f_{1,p+1}, \ldots, f_{1,T}, f_{2,p+1}, \ldots, f_{2,T}, \ldots, f_{p,p+1}, \ldots, f_{p,T}]$$

The vector $\boldsymbol{f}$ is the random vector of the functional coefficients appearing in the model, ordered first by function and then by time. We can show that the posterior distribution of the functional coefficients is also multivariate normal. A-priori, each function is independent but a-posteriori all functions are correlated, with the covariance function depending on the observations. There is an alternative perspective of this model which bypasses explicit consideration of the functional coefficients. We can look at the conditional means of the observations which are given by

$$z_t = f_{1,t}x_{t-1} + \ldots + f_{p,t}x_{t-p}$$

The variable $z_t$ (conditional on $x_{t-1}, \ldots, x_{t-p}$) follows a normal distribution as a linear combination of normals. This allows us to work with the model directly in terms of the conditional mean of the observations. This way, we reduce the dimension of the problem significantly, in fact we make it independent of the order of the model $p$. Let $n = T - p$ be the actual number of observations used, after we condition on the first $p$. Working with $\boldsymbol{f}$, we have to consider $pn$ normal random variables, whereas working with the conditional means $z_t$ we only have to consider $n$ normal random variables. For likelihood calculations and predictions from the model we just need the conditional means. Working with the functional coefficient vector $\boldsymbol{f}$ is only justified when we specifically want to get the posterior distribution of the functions. This point is important because calculations for multivariate normal random vectors are computationaly expensive; they require inversion of large covariance matrices. Estimation of the model based on the functional coefficients scales as $\mathcal{O}(p^2n^3)$, whereas estimation based on the conditional means scales as $\mathcal{O}(n^3)$ and this difference can be substantial for high autoregressive order models.

We now discuss the prior specification of the model. We still have to decide on the mean and covariance functions $\mu_i(\cdot)$ and $C_i(\cdot, \cdot)$ respectively. For $\mu_i(\cdot)$ we propose a simple constant function, since it is not very important for estimation. The covariance function, however, is very important as it controls the properties of the functions and especially their smoothness. There are a number of alternatives in the literature depending on the behavior one wants to model. Our setting allows each function to have a different form, apart from their having different parameters. The most popular choice is the squared exponential, which we also adopt in the examples. Assuming we have decided on the particular form, we still have to select the parameters which control the functions. To this end, one approach is to treat them as hyperparameters, put a prior distribution on them and run an MCMC simulation to approximate their posterior. Alternatively, we can use an empirical Bayes procedure and maximize the marginal (conditional) likelihood of the data w.r.t. these parameters. All likelihood calculations can be performed using the conditional means, as we noted before. The gradient of the marginal likelihood is also available for the empirical Bayes approach with little extra effort. The empirical Bayes procedure scales as $\mathcal{O}(n^3 + n_\theta n^2)$ for each step of the numerical maximization scheme compared to $\mathcal{O}(n_\theta n^3)$ for each iteration of MCMC, where $n_\theta$ is the total number of parameters. We used the empirical Bayes approach in the examples because of its speed.

We now make some comments on our proposed procedure and try to put it in context relative to other methods. All of the nonparametric methods we consider for model (3), namely kernel smoothing, local linear regression, smoothing splines and our method, have also been proposed in the regression setting. The relevance with time series and the autoregressive setting is that after we condition on the initial $p$ observations we can carry out the analysis as in the regression setting, by using conditional likelihood or conditional least squares. The major difference, though, is that the errors used in least squares are serially correlated and this complicates matters significantly. In particular, it invalidates simple leave-one-out cross-validation for the choice of smoothing parameters. This is an important advantage for using likelihood methods for time series. First, we should note

that the use of the conditional likelihood is justified by the Shannon-McMillan-Breiman theorem for densities (see Barron [1]), which holds under certain ergodicity conditions. This means that it is reasonable to use the same likelihood procedures as in regression for our model. For procedures which depend on leave-one-out cross-validation for the choice of smoothing parameters in regression, we need to use a different scheme that takes into account serial dependence. The majority of the literature suggests multifold cross-validation schemes, which leave out consecutive sequences of data on which the accuracy of predictions is measured. The justification of the method usually requires stronger assumptions on the series, but more importantly, the implementation involves ad-hoc choices for the sizes and positions of the sequences as well as the type of predictions (one- versus multi-step-ahead). Another important disadvantage, which is specific to the FAR setting, is that we need to choose many smoothing parameters for models which allow variable bandwidths for each function. This implies a grid search over $\mathbb{R}^p$, which can get computationally expensive even for moderate $p$. The empirical Bayes approach, on the other hand, uses a gradient-based numerical maximization that is much more convenient and fast. We now look at the different methods separately. As we pointed out before, both ALR and LLR are restricted to the case where all functions have the same argument in (3). Moreover, they only use a single bandwidth parameter for estimation, which can lead to compromised accuracy. Spline methods have the potential to overcome these disadvantages. Splines using a maximal set of knots and regularization scales as $\mathcal{O}(p^3n^3)$, even worse than our method. Using a smaller number of knots for each function, in order to control the smoothing, can speed up the method significantly ( scales as $\mathcal{O}\big((\sum_{i=1}^p k_i)^3\big)$, where $k_i$ is the number of spline bases for function $f_i$). Both spline methods, though, can run into diffuculties from the need to select multiple smoothing parameters, as we said before. Moreover, both methods become much slower if we use multidimensional splines, i.e. if the arguments to the functions $f_i$ are not scalar. This is a case where our methods is essentially unaffected. Another feature of our method, which is a byproduct of the Bayesian approach, is that it gives exact measures of uncertainty for the estimated functions. However, the main limitation of GP regression for the FAR models, but also in general, is that it becomes practically infeasible when the number of observations is high. So, to summarize, we believe that our method is an attractive alternative to existing techniques for nonparametric estimation in FAR models, which can be easily applied to data sets of up a couple of thousand obsrvations.

# 4 Preliminary Results

## 4.1 Simulated data

We provide a simulated time series example that is especially tailored to demonstrate the advantages the proposed model. We consider a FAR model of order two, where the functions have very different smoothness. The first is a wiggly sinusoidal with decaying amplitude and the second is a smooth logistic. The two coefficient functions $f_1$, $f_2$ are given as

$$f_1(U) = \left(\cos(\pi U)\exp\{-U^2/10\} - 1\right)/2, \quad f_2(U) = \frac{\exp\{U/2\}/2}{1+\exp\{U/2\}}.$$

We simulate 400 observations from the following time series

$$x_t = f_1(x_{t-1})x_{t-1} + f_2(x_{t-1})x_{t-1} + \epsilon_t; \quad \epsilon_t \sim \mathcal{N}(0, 0.3^2) \tag{12}$$

The plot of the simulated time series is shown in Fig. 1. We estimate the functional coefficients by the LLR method of Cai, Fan and Yao, the spline method of Huang ang Shen and our proposed GP methodology. The smoothing parameter selection is done according to the suggested procedure for each method. For LLR, the optimal bandwidth is 0.675; for the spline method, 5 and 1 equally spaced knots were selected for $f_1$ and $f_2$ respectively. We used a squared exponential covariance function for our GP processes with smoothing parameters 0.35 and 84.81 selected by maximizing the marginal likelihood. Note that both splines and our method accounted for the different smoothness. A plot of the estimated functional coefficients, together with the true functions, is given in Fig. 2.

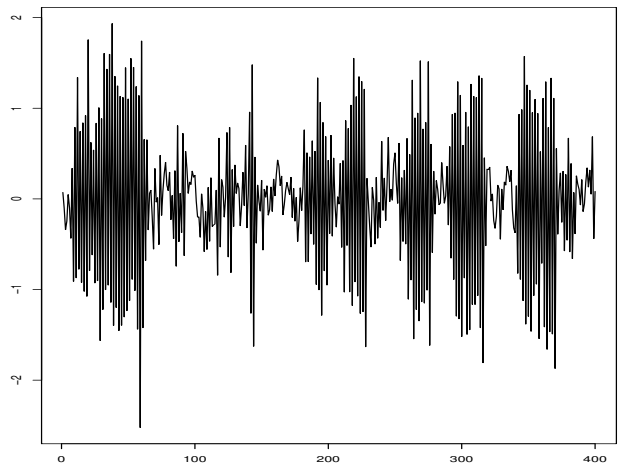
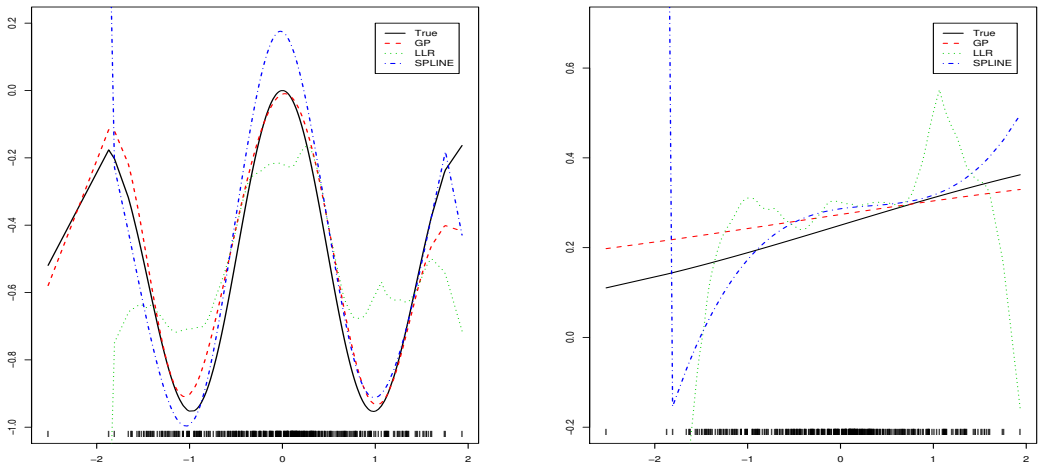



Figure 1: Plot of simulated data from model (12).

As we can see from the plots of the estimates, the LLR method fails to capture the difference in smoothnes of the functional coefficients. This is not due to the particular selected bandwidth, but it is an inherent limitation of the estimation procedure. The spline method accounts for variable smoothnes, as does our GP method, and both give estimates close to the true functions. The striking result from these plots is the very bad behavior at the edges of the observed range of the




(a) (b)

Figure 2: Plots of true and estimated functional coefficients for (a) $f_1$ and (b) $f_2$ using GP, LLR and spline methods for the simulated data.

functions for the LLR and spline methods. There are only few and sparse data at the edges and both methods rely on regression which might not be well behaved at these points. For LLR, we use a weighted local regression, but the bandwidth is too small to include enough data points to get good results. For the spline method, the problem is basically due to the lack of regularization. The method tries to control smoothnes via the number of equally spaced knots, but the arguments to the functions do not spread regularly throughout the range and this is typical behavior for time series. As a result, the spline basis coefficients at the edges can behave very badly. The GP method, on the other hand, estimates the true function very well and does not suffer from problematic behavior at the edges. That is because our model will shrink the functional coefficients outside the observed range towards their prior mean, which is a constant function. Thus, the GP model will tend to give estimates outside the observed range that approximate an AR model with coefficients equal to the function's prior means. This is a case where the prior bias is welcome, especially if we want to make multi-step-ahead predictions or simulations from the FAR model. In these situations the range of the functions and the position at which we need to predict are dynamic and not known beforehand.

### 4.2 Canadian Lynx Data

In this section, we apply our model to real data. We look at the famous Canadian lynx time series, which is the annual record of the number of Canadian lynx trapped in the MacKenzie river district from 1821 to 1934. We will work with the base 10 logarithm of the data in order to stabilize the variance; the plot of the transformed series is shown in Fig. 3. These data have traditionally served as an example for the need of nonlinear and nonparametric time series models. As suggested in the literature, we will use a second order FAR model, where the functional coefficients both depend on the lag-two variable, i.e. $U_t = x_{t-2}$. The model is

$$x_t \quad = \quad f_1(x_{t-2})x_{t-1} + f_2(x_{t-2})x_{t-2} + \epsilon_t$$

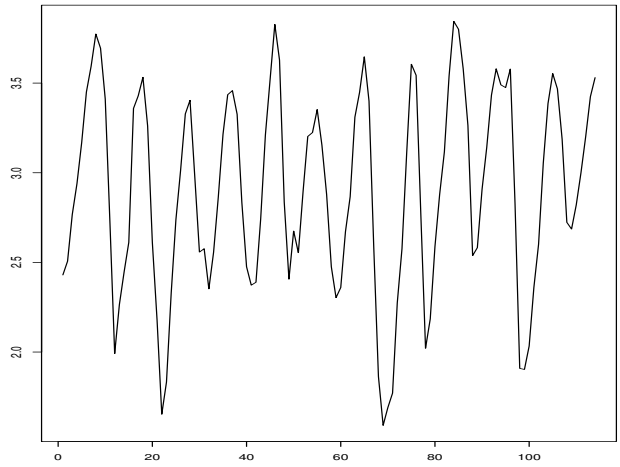
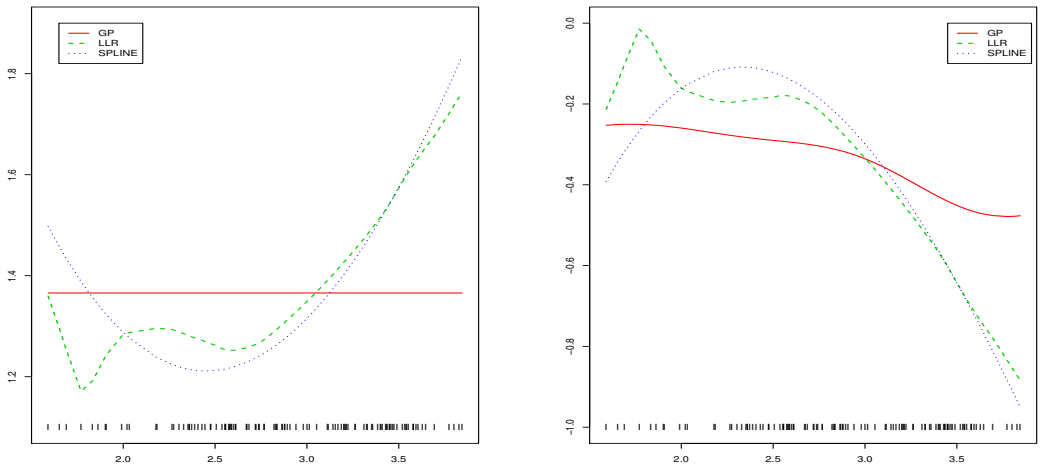



Figure 3: Plot of the logarithm of the Canadian lynx time series.

We apply our GP, the LLR and the spline method and we also use a parametric TAR model. The resulting posterior functional coefficients for the nonparametric methods are presented in Fig. 4. The least squares estimated TAR model is

$$x_t \quad = \quad \begin{cases} 0.59 + 1.25x_{t-1} - 0.42x_{t-2} + \epsilon_t, & \text{if } x_{t-2} \leq 3.25 \\ 2.23 + 1.52x_{t-1} - 1.24x_{t-2} + \epsilon_t, & \text{if } x_{t-2} > 3.25 \end{cases}$$

As we can see, the fitted GP model suggests that $f_1$ is constant, which makes the coefficient of $x_{t-1}$ independent of $x_{t-2}$, whereas none of the other models supports this conclusion. Moreover, we got




(a) (b)

Figure 4: Plots of estimated functional coefficients for (a) $f_1$ and (b) $f_2$ for the Canadian lynx data.

this result without having to change the specification of our model. 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 achieved easily, but we would still have to change the specification of the model. The way we control smoothing, by the number of knots, does not in general allow us to fit a constant function. For all models, however, $f_2$ becomes more negative as $x_{t-2}$ increases. This is in accordance with population dynamics principles which apply to the lynx data. 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 fit of the models; the fitted values from all four models are plotted in Fig 5 and they are practically indistinguishable. We also look at the predictive performance of these models. We refit the models to the first 102 data 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 all models, the parameters are chosen based on the first 102 data. The one-step-ahead predictions are shown in Fig. 6 and the multi-step-ahead predictions are shown in Fig. 7. The one-step-ahead predictions are very close for all models but this is not surprising, since they all have similar

fitted values, which can be thought of as within sample one-step-ahead predictions. However, our modelling procedure seems to give improved multi-step-ahead predictions, which follow the true process more closely. So, the proposed GP estimation method gives a more parsimonious model with at least as good fit to the data as the other models and better predictive performance.

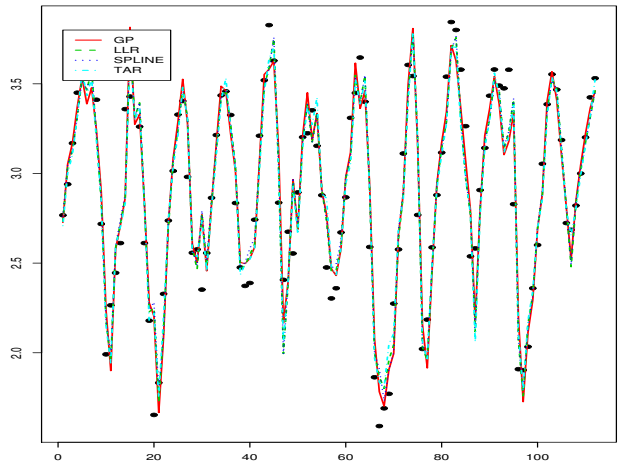



Figure 5: Fitted values from all four models applied to the Canadian lynx data, dots represent true values.

# 5 Future Work

In this section, we list three areas of interest that will be the focus of our future efforts in hopes of shaping the current work into a thesis. These are described in more detail in the following paragraphs; at the end we also suggest other possible directions of investigation, if time permits.

1. Consistency of the estimation procedure.
2. Approximation schemes for large data sets.
3. Evaluating the fit of the model.

## 5.1 Consistency

We want to explore the theoretical properties of our approach and, in particular, to provide a consistency result for our estimator. This will at least make our method comparable to the competing methods we have presented, for which there exist results of this kind. For the ALR method,

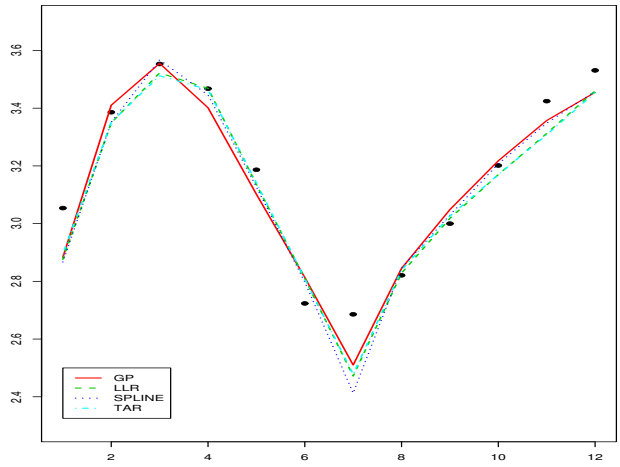



Figure 6: One-step-ahead predictions from all three models applied to the Canadian lynx data, dots represent true values.

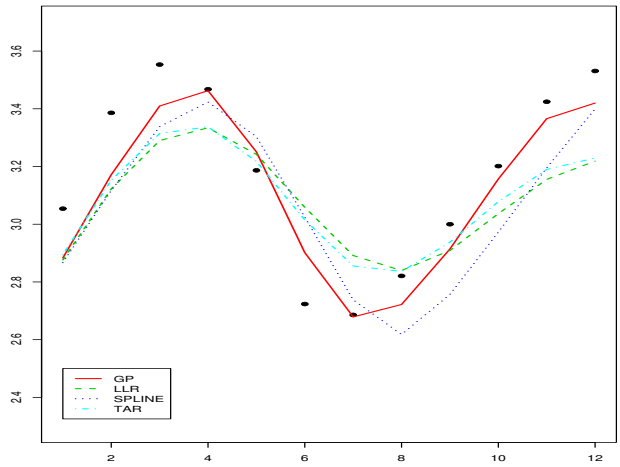



Figure 7: Multi-step-ahead predictions from all three models applied to the Canadian lynx data, dots represent true values.

Chen [6] proves mean square consistency of the functional coefficients evaluated at the observed data points. For the LLR method, Cai, Fan and Yao [5] derive the asymptotic distribution of the

functional coefficients evaluated at a fixed point, together with a bias-variance decomposition and rates of convergence. Finally, for the spline method, Huang and Chen [17] prove $L^2$ consistency of the estimated functions over a truncated range. The type of consistency result we are aiming at has to do with the posterior mean of the functional coefficients and not with their entire posterior distribution. For this end, we will view our estimators as solutions to a penalized least squares problem, using the idea of reproducing kernel Hilbert spaces (RKHS); see Berlinet and Thomas-Agnan [2] for a description of RKHS. Specifically, each posterior mean function in our model can be written (under zero prior mean) as a linear combination of its covariance functions centered at the observed arguments, in the following way

$$\mathbb{E}[f_i(\cdot)|x_{p+1:T}] = \sum_t \alpha_{i,t} C_i(\cdot, U_t^{(i)}); \quad i = 1, \ldots, p \tag{13}$$

Thus, the posterior mean function belongs to its corresponding RKHS $\mathcal{C}_i$, defined by the covariance function (kernel) $C_i(\cdot,\cdot)$. The estimator in (13) is the solution to the penalized least squares problem

$$\min_{f_i \in \mathcal{C}_i,\, i=1,\ldots,p} \left\{ \frac{1}{\sigma^2} \sum_t (x_t - f_1(U_t^{(1)})x_{t-1} - \ldots - f_p(U_t^{(p)})x_{t-p})^2 + \sum_i \|f_i\|_{\mathcal{C}_i}^2 \right\} \tag{14}$$

The setting is similar to that of smoothing splines, with the difference that the smoothness penalty is based on the RKHS norm $\|\cdot\|_{\mathcal{C}_i}$, instead of the $L^2$ norm. Using this formulation, we will try to prove consistency of our estimators under appropriate assumptions. Roughly, we will need to assume the true functions belong to the corresponding RKHS and the process satisfies conditions such that the sum in (14) converges and the functional coefficients are identifiable. Asymptotically, we expect the smoothness penalty to die off and the squared errors to be minimized by our estimators converging to the true functions.

### 5.2 Approximation Methods

As we mentioned earlier, GP regression is computationally expensive, requiring $\mathcal{O}(n^3)$ operations where $n$ is the number of observations. This significantly limits the applicability of the exact method to relatively small data sets. It is, therefore, important to develop numerical approximation schemes for inference when the number of observations is high. Different approximation schemes have been proposed in the literature in order to cope with this problem. Gibbs and McKay [10] use iterative solutions to the linear system of equations which scale as $\mathcal{O}(mn^2)$, where $m$ is the number of iterations. Although this offers an improvement, the scale factor of $n^2$ can still be prohibiting. A faster way of doing inference is based on a reduced rank approximation of the prior covariance matrix of the functions, as described by Rasmussen and Williams [23] (ch. 8). The idea is to approximate the covariance matrix $\boldsymbol{C}$ as

$$\boldsymbol{C} = \boldsymbol{W}\boldsymbol{V}\boldsymbol{W}^{\top} \tag{15}$$

where $\boldsymbol{W}$ is $n \times m$, $\boldsymbol{V}$ is full-rank $m \times m$ and $m < n$. The cost of carrying out the computations using the approximation in (15) scales as $\mathcal{O}(m^2 n)$, where $m$ is the reduced rank of $\boldsymbol{C}$. This is a significant improvement and allows us to work with large numbers of observations. There are two common ways in the literature to approximate the covariance matrix. The first is the Nyström method (see Williams and Seeger [28]), which relies on an approximate eigenvalue decomposition of the covariance kernel and 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 than $n$ number of knots. Both methods are reported to give similar estimates for well behaved data sets, but the SR method has the added advantage that it can be formulated as a consistent model. This avoids undesirable consequences of the approximation, such as negative variances, and provides a proper likelihood for use in parameter selection. In the case of the FAR model we expect the quality of the approximation to be good, given that the coefficient functions have few arguments and tend to be quite smooth for most applications (so that their covariance matrices are practically lower rank than $n$). From preliminary experiments, the SR method seems to behave quite well but we want to further develop the approximation methods for parameter selection and estimation and make extensive tests of their performance.

### 5.3 Measures of Fit

Finally, we need a set of tools for assessing and comparing model fits when we implement our estimation procedure. We refer to these collectively as measures of fit, but we will use them to address different types of questions. First, we need a procedure for order selection and comparison of different model specifications. To this end, we can use information criteria such as AIC or BIC; our procedure provides us with a likelihood which we can penalize for model complexity. We note that our resulting fitted values are in the form of a nonlinear smoother (since the smoother matrix is data dependent). Thus, the standard effective degrees of freedom (EDF) measure from nonparametric regression is no longer recommended and we will use the alternative EDF definition suggested by Lin and Pourahmadi [19]. Another procedure which is useful for comparing nonparametric versus parametric fits for the same model specification is described in [5]. The authors propose a residual sum of squares ratio statistic and approximate its distribution by nonparametric bootstrap. Moreover, we will investigate goodness of fit tests employing the idea of universal residuals as illustrated in Smith [24] and Brockwell [3]. These residuals are obtained by a conditional distribution transformation and are better suited for goodness of fit tests in non-normal and nonlinear time series. We will also look into other possible methods for evaluating fit and try to develop diagnostic checks for identifying weaknesses in the model. We will illustrate these by applying our method to the formal statistical analysis of financial data. A particular problem we have in mind is modeling the dynamics of the relationship between futures and spot prices of indeces, for which nonlinear time

series models have been applied; see Martens, Kofman and Vorst [21].

Depending on time limitations, we would also like to look at other applications that we have in mind. The first is testing for nonlinearity in time series. Most available tests for nonlinearity depend on comparisons of fitness measures between models. Our method gives uncertainty measures directly on the coefficients, so we could devise a procedure for testing if the functional coefficients are constant, i.e. the model is linear. We would also like to try our method for modeling conditional variances, similar to the ARCH model, by assuming that squared errors follow a log-normal distribution.

## References


[1] A. R. Barron. "The Strong Ergodic Theorem for Densities: Generalized Shannon-McMillan-Breiman Theorem". *Annals of Probability*, 13(4):1292–1303, 1985.

[2] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

[3] A. E. Brockwell. "Universal Residuals: A Multivariate Transformation". Technical report, Dept. of Statistics, Carnegie Mellon University, 2007.

[4] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer-Verlag, New-York, 1991.

[5] 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.

[6] R. Chen. *Two Classes of Non-linear Time Series*. PhD thesis, Carnegie Mellon University, Dept. of Statistics, 1990.

[7] R. Chen and R. S. Tsay. "Functional-Coefficient Autoregressive Models". *Journal of the American Statistical Association*, 88(421):298–308, 1993.

[8] R. Chen and R. S. Tsay. "Nonlinear Additive ARX Models". *Journal of the American Statistical Association*, 88(423):955–967, 1993.

[9] J. Fan and Q. Yao. *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer-Verlag, New-York, 2003.

[10] M. Gibbs and D. MacKay. "Efficient implementation of Gaussian processes". Unpublished manuscript. Clavendish Laboratory, Cambridge, UK. `http://citeseer.ist.psu.edu/gibbs97efficient.html`.

[11] V. Haggan and T. Ozaki. "Modelling Nonlinear Random Vibrations Using an Amplitude-Dependent Autoregressive Time Series Model". *Biometrika*, 68(1):189–196, 1981.

[12] J. D. Hamilton. "Analysis of Time Series Subject to Changes in Regime". *Journal of Econometrics*, 45(1-2):39–70, 1990.

[13] W. Härdle, H. Lütkepohl, and R. Chen. "A Review of Nonparametric Time Series Analysis". *International Statistical Review*, 65(1):49–72, 1997.

[14] T. Hastie and R. Tibshirani. "Generalized Additive Models". *Statistical Science*, 1(3):297–310, 1986.

[15] T. Hastie and R. Tibshirani. "Varying-Coefficient Models". *Journal of The Royal Statistical Society: Series B*, 55(4):757–796, 1993.

[16] T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York, 2001.

[17] 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.

[18] F. Jianqing and W. Zhang. "Statistical Estimation in Varying Coefficient Models". *The Annals of Statistics*, 27(5):1491–1518, 1999.

[19] 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: Series C*, 47(2):187–201, 1998.

[20] O. Linton and J. P. Nielsen. "A Kernel Method of Estimating Structured Nonparametric Regression Based on Marginal Integration". *Biometrica*, 82(1):93–100, 1995.

[21] 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.

[22] A. O'Hagan. "Curve Fitting and Optimal Design for Prediction (with discussion)". *Journal of The Royal Statistical Society: Series B*, 40(1):1–42, 1978.

[23] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge MA, 2006.

[24] J. Q. Smith. "Diagnostics Checks of Non-Standard Time Series Models". *Journal of Forecasting*, 4:283–291, 1985.

[25] C. J. Stone. "Consistent Nonparametric Regression". *The Annals of Statistics*, 5(4):595–620, 1977.

[26] H. Tong. *Non-linear Time Series: A Dynamical Systems Approach*. Oxford University Press, Oxford, 1990.

[27] R. S. Tsay. "Testing and Modeling Threshold Autoregressive Processes". *Journal of the American Statistical Association*, 84(405):231–240, 1989.

[28] 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.