1 Nonparametric extensions of the classical linear model

The classical linear model expresses the influence of covariates $X_1, X_2, \ldots, X_p$ on the response variable Y via

$$Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{1}$$

While linearity is a convenient artifact of specifying model 1, the world is full of nonlinear phenomena such as limit cycles and jump resonance. Consequently, nonlinearity must be a modeling consideration to adequately characterize many of the natural underlying mechanisms that generate data. The nonparametric regression model has been widely used in various applications due to its ability to characterize structure in data that linear and other parametric models fail to adequately represent. However, a serious drawback to the general nonparametric model is the ?curse of dimensionality? phenomenon, a term which refers to the fact that the convergence rate of nonparametric smoothing estimators becomes rather slow when the estimation target is a general function of a large number of variables without additional structures. Many efforts have been made to impose structure on the regression function to alleviate this issue, which is broadly described as dimension reduction. Some approaches to restricting the general nonparametric model include: (generalized) additive models [Chen and Tsay, 1993b][Hastie and Tibshirani, 1990], [Hastie and Tibshirani, 1986], Sperlich, Tjostheim & Yang 2002, Stone 1985), partially linear models [Härdle and Liang, 2007], [Zeger and Diggle, 1994], varying coefficient models [Hastie and Tibshirani, 1993], Fan & Zhang, 1999), and their hybrids (Carroll et al., 1997; Fan et al., 1998; Heckman et al., 1998), among others.

An immediate problem of departing from linearity is the need for a class of well-parameterized nonlinear models that are simple yet sufficient in handling most nonlinear phenomena observed in practice. Because there is no unified theory applicable to all nonlinear models, this problem is a difficult one. The main difficulty is that unlike linear models where the functions involved can be treated fairly systematically, the set of all nonlinear models is so broad that systematic treatment is infeasible. The expansiveness of the class of nonlinear models is due to both the innumerable nonlinear functions as well as the different structures within a given class of functions.

Varying coefficient models are a particularly attractive extension of the classical linear model. The appeal of this model is that, by allowing regression coefficients to depend on a smoothing parameter Z, the modeling bias can significantly be reduced while avoiding the"curse of dimensionality". Another advantage of this model is its interpretability, and this model structure arises naturally when one is interested in exploring how regression coefficients change over different groups, such as age. The mean function of the response Y take the form

1.1 Extensions of linear models which are special cases of models in 2,3,3.1

To illustrate the flexibility of the varying coefficient model, we examine some models that may be expressed as special cases, first considering the general nonparametric modeling literature.

2 VC Models with a Univariate Smoothing Variable

$$E\left(Y|\boldsymbol{X}=\boldsymbol{x},\ Z=z\right)=x_1\beta_1\left(z\right)+\cdots+x_p\beta_p\left(z\right) \tag{2}$$

where $\boldsymbol{X}=\left(X_1,X_2,\ldots,X_p\right)^T$ and Z are covariates and $\boldsymbol{\beta}\left(z\right)=\left(\beta_0\left(z\right),\beta_1\left(z\right),\ldots,\beta_p\left(z\right)\right)^T$ are unknown coefficient functions, assumed to be smooth functions of Z. It is worth noting that by taking $X_1\equiv 1$, this model allows for a varying intercept term. This class of models is particularly appealing in longitudinal studies where they allow us to examine the extent to which covariates affect responses over time [Hoover et al., 1998], [Fan and Zhang, 2000].

3 VC Models with a Multivariate Smoothing Variable

The second approach in specifying varying coefficient models is by generalizing model 2 to allow each covariate's coefficient function to depend on different covariates, $\boldsymbol{Z}=\left(Z_1,Z_2,\ldots,Z_p\right)^T$. This leads to modeling the mean response as follows:

$$E\left(Y|\boldsymbol{X}=\boldsymbol{x},\ \boldsymbol{Z}=\boldsymbol{z}\right)=x_1\beta_1\left(z_1\right)+\cdots+x_p\beta_p\left(z_p\right) \tag{3}$$

There are many proposed extensions of model 2 and model 3, including models that allow a covariate to play both the roles of the linear effect covariate (X_j) in addition to the roles of the *smoothing variables* (Z_j). One can see that by letting the $\{\beta_j\}$ be constant for $j=1,\ldots,p$, this reduces to 19 proposed by Hoover, Rice, Wu and Yang. The class of models having the form as specified in 3 is quite extensive; however, imposing an additive structure on the multivariate coefficient functions does not permit explicitly modeling interactions between the smoothing variables.

3.1 Extended VC Models and Functional VC ANOVA models

3.1.1 Smoothing Spline ANOVA models and the extended linear modeling framework of huang, stone

discuss the convergence results presented in [Huang, 1998] and [Huang et al., 1998], pointing out the limitations of the conclusions due to the assumptions of the knots.

[Huang and Stone, 2003] proposed a general framework which further broadened the class of multivariate varying coefficient models defined by the structures for varying coefficient models which have already discussed. In their seminal work, they propose extensions of

previously considered structures for multivariate coefficient functions by leveraging polynomials splines on tensor product spaces. This work was preceeded by [Huang et al., 1998] and [Huang et al., 1998], which simplified and extended the theoretical approach

Discuss the unified framework and corresponding theory presented in [Huang, 2001] and then introduce the tensor product models of [Eilers and Marx, 2003], [Marx and Eilers, 2005]

3.2 Multidimensional Penalized Signal Regression of Eilers, Marx

3.2.1 Univariate B-splines

We first begin by establishing some notation: we choose to divide the function domain, the interval $[\nu_{min}, \nu_{max}]$ into n' intervals of equal length using $n' + 1$ interior knots. B-splines are constructed from polynomial pieces, joined at certain values of x, the knots. Once the knots are given, it is easy to compute the B-splines recursively, for any desired degree of the polynomial; see de Boor (1977, 1978), Cox (1981) or Dierckx (1993). A different track was chosen by O'Sullivan (1986, 1988). He proposed to use a relatively large number of knots. To prevent overfitting, a penalty on the second derivative restricts the flexibility of the fitted curve, similar to the penalty pioneered for smoothing splines by Reinsch (1967) and that has become the standard in much of the spline literature; see, for example, Eubank (1988),Wahba (1990) and Green and Silverman (1994). In this paper we simplify and generalize the approach of O'Sullivan, in such a way that it can be applied in any context where regression on B-splines is useful. Only small modifications of the regression equations are necessary.

A B-spline consists of polynomial pieces, connected in a special way. A very simple example is shown at the left of Figure l(a): one B-spline of degree 1.It consists of two linear pieces; one piece from xl to x2, the other from x2to x3.The knots are xl, x2and x3. To the left of xl and to the right of x3 this B-spline is zero. In the right part of Figure l(a),three more B-splines of degree 1 are shown: each one based on three knots. Of course, we can construct as large a set of B-splines as we like, by introducing more knots. In the left part of Figure l(b), a B-spline of degree 2 is shown. It consists of three quadratic pieces, joined a t two knots. At the joining points not only the ordinates of the polynomial pieces match, but also their first derivatives are equal (but not their second derivatives). The B-spline is based on four adjacent knots: xl, ...,x4. In the right part Figure l(b), three more B-splines of degree 2 are shown.

Note that the B-splines overlap each other. First-degree B-splines overlap with two neighbors, second-degree B-splines with four neighbors and so on. Of course, the leftmost and rightmost splines have less overlap. At a given x, two first-degree (or three second-degree) B-splines are nonzero. These examples illustrate the general properties of a B-spline of degree q:

- it consists of q + 1 polynomial pieces, each of degree q

- the polynomial pieces join at q inner knots
- at the joining points, derivatives up to order q - 1 are continuous
- the B-spline is positive on a domain spanned by q +2 knots; everywhere else it is zero
- except at the boundaries, it overlaps with 2q polynomial pieces of its neighbors
- at a given x, q +1B-splines are nonzero.

Let the domain from xmi, to x, be divided into n' equal intervals by n' +1knots. Each interval will be covered by q +1B-splines of degree q. The total number of knots for construction of the B-splines will be n' +2q +1. The number of B-splines in the regression is n = n' + q . This is easily verified by constructing graphs like those in Figure 1. B-splines are very attractive as base functions for ("nonparametric") univariate regression. A linear combination of (say)third-degree B-splines gives a smooth curve. Once one can compute the B-splines themselves, their application is no more difficult than polynomial regression. De Boor (1978) gave an algorithm to compute B- splines of any degree from B-splines of lower degree. Because a zero-degree B-spline is just a constant on one interval between two knots, it is simple to compute B-splines of any degree. In this paper we use only equidistant knots, but de Boor's algorithm also works for any placement of knots.

Let Bj(x; q) denote the value at x of the jth B- spline of degree q for a given equidistant grid of knots. A fitted curve 5 to data (xi,y,) is the linear combination (x) = C5=lcijBj(x; q). When the de- gree of the B-splines is clear from the context, or immaterial, we use Bj(x) instead of Bj(x; q). The indexing of B-splines needs some care, especially when we are going to use derivatives. The indexing connects a B-spline to a knot; that is, it gives the index of the knot that characterizes the position of the B-spline. Our choice is to take the leftmost knot, the knot at which the B-spline starts to be- come nonzero. In Figure l(a), xl is the positioning knot for the first B-spline. This choice of indexing demands that we introduce q knots to the left of the domain of x. In the formulas that follow for derivatives, the exact bounds of the index in the sums are immaterial, so we have left them out. De Boor (1978) gives a simple formula for derivatives of B-splines:

$$\begin{aligned} h\sum_j \beta_j B_j'(x,q) &= \sum_j \beta_j B_j(x,q-1) - \sum_j \beta_{j+1} B_{j+1}(x,q-1) \\ &= -\sum_j \Delta\beta_{j+1} B_j(x,q-1) \end{aligned} \tag{4}$$

where h is the distance between knots and $\Delta\beta_j = \beta_j - \beta_{j-1}$. By induction, we have that the second derivative may be characterized as follows:

$$h^2 \sum_j \beta_j B_j''(x,q) = \sum_j \Delta^2 \beta_j B_j(x,q-2) \tag{5}$$

where h is the distance between knots and $\Delta^2\beta_j = \Delta\Delta\beta_j = \beta_j - 2\beta_{j-1} + \beta_{j-2}$. This fact is of particular utility when comparing continuous and discrete roughness penalties, which will follow in later discussion.

3.2.2 Tensor Product B-splines

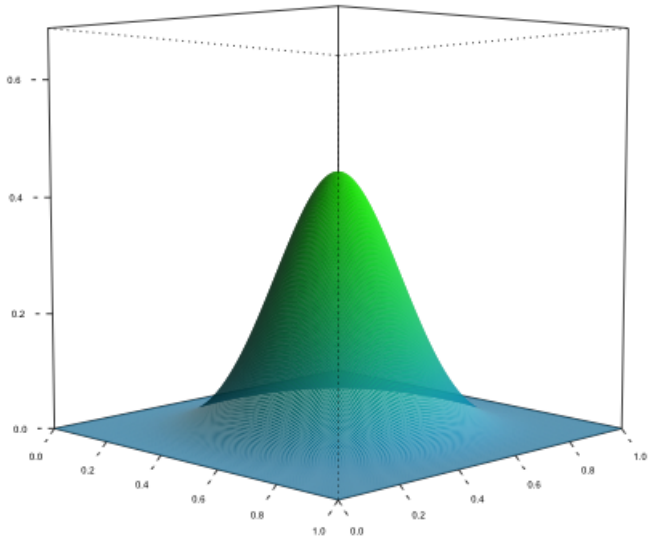

Figure 1: Tensor product of two cubic B-splines

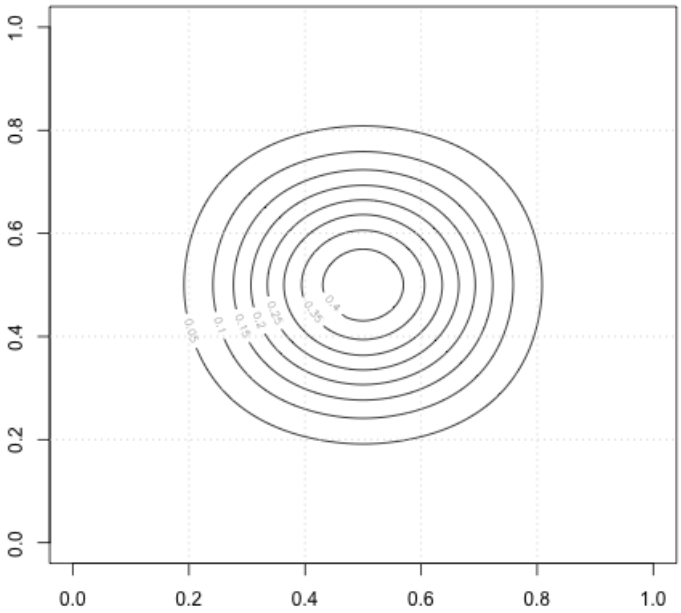

Figure 2: Tensor product of two cubic B-splines

3.2.3 P-Splines Regularization

The choice of knots has been a subject of much research: too many knots lead to overfitting of the data, too few knots lead to underfitting. Some authors have proposed automatic schemes for optimizing the number and the positions of the knots (Friedman and Silverman, 1989; Kooperberg and Stone, 1991,1992). This is a difficult numerical problem and, to our knowledge, no attractive all-purpose scheme exists.

A different track was chosen by O'Sullivan (1986, 1988). He proposed to use a relatively large number of knots. To prevent overfitting, a penalty on the second derivative restricts the flexibility of the fitted curve, similar to the penalty pioneered for smoothing splines by Reinsch (1967) and that has become the standard in much of the spline literature; see, for example, Eubank (1988), Wahba (1990) and Green and Silverman (1994). [Eilers and Marx, 1996] simplify and generalize the approach of O'Sullivan, in such a way that it can be applied in any context where regression on B-splines is useful. Only small modifications of the regression equations are necessary.

The basic idea is not to use the integral of a squared higher derivative of the fitted curve in the penalty, but instead to use a simple difference penalty on the coefficients themselves of adjacent B-splines. We show that both approaches are very similar for second-order differences. In some applications, however, it can be useful to use differences of a smaller or higher order in the penalty. With our approach it is simple to incorporate a penalty

of any order in the (generalized)regression equations. A major problem of any smoothing technique is the choice of the optimal amount of smoothing, in our case the optimal weight of the penalty. We use cross-validation and the Akaike information criterion (AIC). In the latter the effective dimension, that is, the effective number of parameters, of a model plays a crucial role. We follow [Buja et al., 1989] in using the trace of the smoother matrix as the effective dimension. Because we use standard regression techniques, this quantity can be computed easily. We find the trace very useful to compare the effective amount of smoothing for different numbers of knots, different degrees of the B-splines and different orders of penalties.

3.2.4 P-Splines Penalties for Univariate B-Splines

Consider the regression of m data points (x_i, y_i) on a set of n B-splines B_j. The least squares objective function to minimize is

$$\sum_{i=1}^{m} \left\{ y_i - \sum_{j=1}^{n} \beta_j B_j (x_i) \right\}^2 \tag{6}$$

Let the number of knots be relatively large, such that the fitted curve will show more variation than is justified by the data. To make the result less flexible, O'Sullivan (1986, 1988) introduced a penalty on the second derivative of the fitted curve and so formed the objective function

$$\sum_{i=1}^{m} \left\{ y_i - \sum_{j=1}^{n} \beta_j B_j (x_i) \right\}^2 + \lambda \int_{x_{min}}^{x_{max}} \left\{ \sum_{j=1}^{n} \beta_j B_j'' (x) \right\}^2 dx \tag{7}$$

The integral of the square of the second derivative of a fitted function has become common as a smoothness penalty, since the seminal work on smoothing splines by Reinsch (1967). There is nothing special about the second derivative; in fact, lower or higher orders might be used as well. In the context of smoothing splines, the first derivative leads to simple equations, and a piecewise linear fit, while higher derivatives lead to rather complex mathematics, systems of equations with a high bandwidth, and a very smooth fit. [Eilers and Marx, 1996] propose to base the penalty on (higher-order) finite differences of the coefficients of adjacent B-splines:

$$\sum_{i=1}^{m} \left\{ y_i - \sum_{j=1}^{n} \beta_j B_j (x_i) \right\}^2 + \lambda \sum_{j=k+1}^{n} \left(\Delta^k \beta_j \right)^2 \tag{8}$$

This approach reduces the dimensionality of the problem to the number of B-splines, n instead of the number of observations, m , as with smoothing splines. The tuning parameter λ permits continuous control over smoothness of the fit. The difference penalty is a good discrete approximation to the integrated square of the k^{th} derivative, as will be demonstrated below. What is more important: with this penalty moments of the data are conserved and

polynomial regression models occur as limits for large values of λ. See Section 5 for details. We will show below that there is a very strong connection between a penalty on second-order differences of the B-spline coefficients and O'Sullivan's choice of a penalty on the second derivative of the fitted function. However, the difference penalty of [Eilers and Marx, 1996] can be handled mechanically for any order of the differences.

Difference penalties have a long history that goes back at least to Whittaker (1923); recent applications have been described by Green and Yandell (1985) and [Eilers, 1991b], [Eilers, 1991a], [Eilers, 1995]. The difference penalty is easily introduced into the regression equations. That makes it possible to experiment with different orders of the differences. In some cases it is useful to work with even the fourth or higher order. This stems from the fact that for high values of h the fitted curve approaches a parametric (polynomial) model, as will be shown below. [O'Sullivan, 1986] used third-degree B-splines and the following penalty:

$$h^2 P = \lambda \int_{x_{min}}^{x_m ax} \left\{ \sum_j \beta_j B_j'' \left(x, q = 3\right) \right\}^2 dx \tag{9}$$

From the derivative properties of B-splines, it follows that

$$h^2 P = \lambda \int_{x_{min}}^{x_m ax} \sum_j \sum_k \Delta^2 \beta_j \Delta^2 \beta_k B_j \left(x, q = 1\right) B_k \left(x, q = 1\right) dx \tag{10}$$

Most of the cross products of $B_j(x;1)$ and $B_k(x;1)$ vanish as B-splines of degree 1 only overlap when j is $k-1$, k, or $k+1$. Thus, we have that

$$\begin{aligned} h^2 P &= \lambda \int_{x_{min}}^{x_{max}} \left[\left\{ \sum_j \Delta^2 \beta_j B_j \left(x, 1\right) \right\}^2 + 2 \sum_j \Delta^2 \beta_j \Delta^2 \beta_{j-1} B_j \left(x, 1\right) B_{j-1} \left(x, 1\right) \right] dx \\ &= \lambda \left[\sum_j \left(\Delta^2 \beta_j\right)^2 \int_{x_{min}}^{x_{max}} B_j^2 \left(x, 1\right) \; dx + 2 \sum_j \Delta^2 \beta_j \Delta^2 \beta_{j-1} \right] \end{aligned} \tag{11}$$

or

$$\begin{aligned} h^2 P = \lambda \sum_j \left(\Delta^2 \beta_j\right)^2 \int_{x_{min}}^{x_{max}} B_j^2 \left(x, 1\right) dx + 2\lambda \sum_j \Delta^2 \beta_j \Delta^2 \beta_{j-1} \\ + \int_{x_{min}}^{x_{max}} B_j \left(x, 1\right) B_{j-1} \left(x, 1\right) dx \end{aligned} \tag{12}$$

which can be written as

$$h^2 P = \lambda \left\{ c_1 \sum_j \left(\Delta^2 \beta_j\right)^2 + c_2 \sum_j \Delta^2 \beta_j \Delta^2 \beta_{j-1} \right\} \tag{13}$$

where, for given equidistant knots, c_1 and c_2 are constants given by

$$\begin{aligned} c_1 &= \int_{x_{min}}^{x_{max}} B_j^2(x,1)\,dx \\ c_2 &= \int_{x_{min}}^{x_{max}} B_j(x,1)\,B_{j-1}(x,1)\,dx \end{aligned} \tag{14}$$

Thus, we see that O'Sullivan's ridge-like B-spline penalty 9 can be written as a linear combination of Marx and Eilers' difference penalty 8 and the sum of the cross products of neighboring second differences. The second term in 13 leads to a complex objective function when minimizing the penalized likelihood, where seven adjacent spline coefficients occur, as opposed to five if only the first term in 13 is used in the penalty. The added complexity is a consequence of overlapping B-splines, and these complexities grow quickly with higher order differences and B-splines of higher degree, which makes it difficult to construct a procedure for incorporating the penalty in the likelihood equations. Using a difference penalty allows us to sidestep this complexity.

Using the sum of squared errors as the goodness of fit measure, we define $\hat{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)$ to be the minimizer of

$$L(Y,\beta) + \lambda J(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left\{ y_i - \sum_{j=1}^{n} \beta_j B_j(x_i) \right\}^2 + \lambda \sum_{j=k+1}^{n} \left(\Delta^k \beta_j\right)^2$$

In vector notation, this may be written

$$L(Y,\beta) + \lambda J(\beta) = (Y - B\beta)^T (Y - B\beta) + \lambda (D_k\beta)^T (D_k\beta) \tag{15}$$

where where D_k is the matrix representation of the difference operator Ak,and the elements of B are $b_{ij} = B_j(x_i)$. Taking derivatives on both sides of 15 with respect to β gives

$$\begin{aligned} \frac{\partial}{\partial\beta}\left(L(Y,\beta) + \lambda J(\beta)\right) &= \frac{\partial}{\partial\beta}\left(\beta^T B^T B\beta - 2Y^T B^T\beta + \lambda\beta^T D_k^T D_k\beta\right) \\ &= 2B^T B\beta - 2B^T Y + 2\lambda D_k^T D_k\beta \\ &= \left(B^T B + \lambda D_k^T D_k\right)\beta - B^T Y \end{aligned} \tag{16}$$

Setting 16 equal to zero yields the following normal equations:

$$B^T Y = \left(B^T B + \lambda D_k^T D_k\right)\beta \tag{17}$$

When $\lambda = 0$, we have the standard normal equations of linear regression with a B-spline basis. With $k = 0$ we have a special case of ridge regression. When $\lambda > 0$, the penalty only influences the main diagonal and k subdiagonals (on both sides of the main diagonal) of the system of equations. This system has a banded structure because of the limited overlap of the B-splines. It is seldom worth the trouble to exploit this special structure, as the number of equations is equal to the number of splines, which is generally moderate (10-20).

3.3 Nonparametric approaches to modeling nonlinear time series data

Zeger and Diggle (1994) present a partially linear model motivated by the longitudinal data produced by the Multicenter AIDS Cohort Study. The data are of the form $\{(x_{ij}, y_{ij}(t_{ij})) : \quad j = 1, \ldots, m_i;$ where x_{ij} denotes a $p \times 1$ vector of covariates corresponding to $y_{ij}(t_{ij})$, the jth measurement on the i^{th} subject at time t_{ij}. They let

$$Y_{ij}(t) = x_{ij}^T \beta + \mu(t) + W_i(t) + \epsilon_{ij} \tag{18}$$

where $\mu(t)$ is a smooth function of time, and β is a $p \times 1$ vector of regression coefficients. The $\{W_i(t) : \ i = 1, \ldots, n\}$ capture the within-subject dependency structure, defined to be independent replicates of a stationary Gaussian process with mean zero and covariance function $\gamma(v) = \sigma_w^2 \rho(v, \theta)$. The $\{Z_{ij} : \ j = 1, \ldots, m_i \ i = 1, \ldots, n\}$ are mutually independent Normally distributed error terms with mean zero and variance σ_z^2.

Hoover, Rice, Wu and Yang (1998) considered the following model:

$$Y(t) = \boldsymbol{X}^T(t) \boldsymbol{\beta}(t) + \epsilon(t) \tag{19}$$

proposing estimation of the coefficient functions via smoothing splines and local polynomials. $\epsilon(t)$ is defined as in 18 and is assumed to be independent of $\boldsymbol{X}(t)$. Hoover et al (1998) propose the same model, using smoothing splines and kernel smoothing to estimate the components of $\boldsymbol{\beta}(t)$ and develop asymptotic properties of kernel estimators.

For nonlinear time series applications, Chen & Tsay [Chen and Tsay, 1993a] and Xia & Li (1999) develop functional-coefficient autoregressive models. The common research in nonlinear time series analysis has focused on several classes of models, such as the threshold autoregressive (TAR) model of Tong (1983, 1990) and the exponential autoregressive (EXPAR) model of Haggan and Ozaki (1981). In this article we are concerned with empirical modeling of nonlinear time series. In particular we focus on exploring the nonlinear feature of a time series in the process of model building. This is achieved by generalizing directly the linear autoregressive (AR) models and exploiting local characteristics of a given time series. The generalized model is referred to as the functional coefficient autoregressive (FAR) models. Most nonlinear AR models considered in the literature are special cases of the FAR model. It turns out that the FAR models are flexible enough to accommodate most nonlinear features considered in the literature while being simple enough to be treated with relative ease.

4 Model estimation

Zeger and Diggle (1994) carry out estimation of $\mu(t)$ and β as defined in model 18 iteratively via kernel smoothing and generalized least squares. While more flexible than the classical linear model, this still limiting as it does not allow us to explain any dynamic effect of the covariates over time.

In the case of a single common smoothing variable, estimation of 2 via kernel smoothing is quite straightforward. Since the space of the smoothing variable is of only one dimension, smoothing of the p coefficient functions reduces to finding the local least squares fit using a single smoothing bandwidth. This approach, however, may lead to inadequate estimators since the functions $\beta_0(z), \beta_1(z), \ldots, \beta_p(z)$ may need varying degrees of smoothing in the z dimension. To address this,

4.1 Kernel estimation with a single smoothing variable

Suppose we have a random sample of data, consisting of $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, for $i = 1, \ldots, n$. In classical univariate nonparametric regression, we model

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{20}$$

where f is the unknown smooth regression function of interest, and the $\{\epsilon_i\}$ are mutually independent mean-zero errors, with $Var(\epsilon_i) = \sigma_\epsilon^2$. To derive the form of the estimator of the mean function, we consider expressing f in terms of the joint probability distribution of X and Y:

$$\begin{aligned} f(x) = E(Y|X = x) &= \int yp(y|x)\, dy \\ &= \frac{\int yp(y|x)\, dy}{\int p(y|x)\, dy} \end{aligned} \tag{21}$$

Let K denote a kernel function corresponding to a probability density, h denote the smoothing bandwidth, and let

$$K_h(t) = h^{-1} K\left(h^{-1} t\right)$$

The Nadaraya-Watson estimator of the joint density of x and y has form

$$\begin{aligned} \hat{p}(x, y) &= \frac{1}{nh_x h_y} \sum_{i=1}^{n} K_{h_x}\left(\frac{x - x_i}{h_x}\right) K_{h_y}\left(\frac{y - y_i}{h_y}\right) \\ &= \frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - x_i)\, K_{h_y}(y - y_i) \end{aligned} \tag{22}$$

Then, substituting 22 for $p(x, y)$ in the numerator of 21, we can write

$$\int y\hat{p}(x, y)\ dy = \frac{1}{n} \int yK_{h_x}(x - x_i)\, K_{h_y}(y - y_i)$$

Since $\int yK_{h_y}(y - y_i)\, dy = y_i$, we have that

$$\int y\hat{p}(x, y)\ dy = \frac{1}{n} \sum_{i=1}^{n} K_{h_x}(x - x_i)\, y_i \tag{23}$$

Estimating the denominator of 21 in similar fashion, we have

$$\begin{aligned}\int \hat{p}(x,y)\ dy &= \frac{1}{n}\sum_{i=1}^{n} K_{h_x}(x-x_i)\int K_{h_y}(y-y_i)\ dy \\ &= \frac{1}{n}\sum_{i=1}^{n} K_{h_x}(x-x_i) \\ &= \hat{f}_x(x) \end{aligned} \tag{24}$$

Using 23 and 24 as plug-in estimators in 21, then

$$\hat{f}(x) = \sum_{i=1}^{n} W_{h_x}(x,x_i)\, y_i \tag{25}$$

where

$$W_{h_x}(x,x_i) = \frac{K_{h_x}(x-x_i)}{\sum_{i=1}^{n} K_{h_x}(x-x_i)}$$

and $\sum_{i=1}^{n} W_{h_x}(x,x_i) = 1$. One can extend this to the case where the regression function is defined as in 2; the Nadaraya-Watson (NW) estimator of $\boldsymbol{\beta}(z_0) = (\beta_0(z_0,), \beta_1(z_0,), \ldots, \beta_p(z_0,))^T$ minimizes

$$\sum_{i=1}^{n}\left(Y_i - \left(\sum_{j=1}^{p}\alpha_j X_{ij}\right)\right)^2 K_{h_z}(z_0, Z_i)$$

with respect to $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$ for each target point z_0. Let $\mathcal{X}$ denote the $n \times p$ matrix having $i-j^{th}$ element X_{ij}, $\mathcal{W}$ denote the $n \times n$ diagonal matrix with i^{th} diagonal entry $K_{h_z}(z_0, Z_i)$, and let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^T$. Further, let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, then the NW estimator has form

$$\hat{\boldsymbol{\beta}}(z_0) = \left[\mathcal{X}^T\mathcal{W}\mathcal{X}\right]^{-1}\mathcal{X}^T\mathcal{W}\boldsymbol{Y}$$

It is well known that locally weighted averages can exhibit high bias near the boundaries of the smoothing variable domain, due to the asymmetry of the kernel in that region. This bias can also be present on the interior of the domain when the observed values of Z are irregularly sampled, though it is typically less severe in the interior than near the boundaries. To remedy this, one may consider fitting local linear smoothers, which will correct this bias to first order. The local linear smoother minimizes

$$\sum_{i=1}^{n}\left[Y_i - \sum_{j=1}^{p}\left(\alpha_{0j} + \alpha_{1j}(Z_i - z_0)\right)X_{ij}\right]^2 K_{h_z}(z_0, Z_i) \tag{26}$$

with respect to $\boldsymbol{\alpha}_0 = (\alpha_{01}, \ldots, \alpha_{0p})^T$, and $\boldsymbol{\alpha}_1 = (\alpha_{11}, \ldots, \alpha_{1p})^T$. Let $\mathcal{X}$ denote the $n \times 2p$ matrix having $i-j^{th}$ element X_{ij} and $i-(j+p)^{th}$ element $(Z_i - z_0)X_{ij}$ for $1 \le j \le p$, then the minimizer of **??** is given by

$$\hat{\boldsymbol{\beta}}(z_0) = \left[\mathcal{I}_p, \boldsymbol{O}_p\right]\left[\mathcal{X}^T\mathcal{W}\mathcal{X}\right]^{-1}\mathcal{X}^T\mathcal{W}\boldsymbol{Y}$$

where $\mathcal{I}_p$ is the $p \times p$ identity matrix, and $\boldsymbol{O}_p$ is the $p \times p$ zero matrix. Extensions to the case of a single multivariate smoothing variable $\boldsymbol{Z}$, where the mean function is given by

$$E(Y|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z}) = x_1\beta_1(\boldsymbol{z}) + \cdots + x_p\beta_p(\boldsymbol{z})$$

However, while boundary effects associated with the NW estimator are a concern in one dimension, the curse of dimensionality makes these effects much more problematic in two or more dimensions. The fraction of points close to the boundary of the domain approaches one as the dimensionality of the input space grows, and simultaneously maintaining locality (and low bias) as well as sizable number of observations in the neighborhood of the target point, z_0 (low variance) becomes an increasingly tall order.

4.1.1 Kernel bandwidth selection with a single smoothing variable

4.1.2 Asymptotic properties of kernel estimators with a single smoothing variable

4.1.3 Two-step estimation for multiple bandwidths

Model selection as described in 4.1.1 assumes a single smoothing bandwidth h_z as well as a single common kernel function K for every coefficient function β_j. While convenient and straightforward, in practice, the assumption that each coefficient function should receive the same degree of smoothing is likely to be an erroneous one. Fan and Zhang (1999) present an intuitive formulation of their proposed two-stage estimation procedure that allows for each coefficient function to have its own smoothing bandwidth. Assume that $\beta_p(z)$ is smoother than the other $p-1$ coefficient functions, and can be locally approximated by a cubic polynomial:

$$\beta_p(z) \approx b_{0p} + b_{1p}(z - z_0) + b_{2p}(z - z_0)^2 + b_{3p}(z - z_0)^3$$

for any z_0 close to z. Let $\left\{\tilde{b}_{0j}, \tilde{b}_{1j}\right\}$, $j = 1, \ldots, p-1$ and $\tilde{b}_{0p}, \tilde{b}_{1p}, \tilde{b}_{2p}, \tilde{b}_{3p}$ be the minimizers of the weighted sums of squares:

$$\sum_{i=1}^{n}\Big[Y_i - \sum_{j=1}^{p-1}\left\{b_{0j} + b_{1j}(Z_i - z_0)\right\}X_{ij} - \left\{b_{0p} + b_{1p}(z - z_0) + b_{2p}(z - z_0)^2 + b_{3p}(z - z_0)^3\right\}X_{ip}\Big]^2 \times K_{h_1}(Z_i - z_0)$$

If we take $\tilde{\beta}_p^{os}(z_0) = \tilde{b}_{0p}$, then they show that the bias of the the *one-step estimator* is $O(h_0^2)$ and the variance is $O\left((nh_0)^{-1}\right)$. Fan and Zhang (1999) propose a two-step estimation procedure that allows for individual degrees of smoothing of each of the coefficient functions;

Cai (2000) further investigated this two-step approach. In the first step, to estimate $\beta_j(z_0)$, a preliminary estimate, $\tilde{\beta}_j$, is obtained by applying a local cubic smoother to β_j and local linear smoothing to the remaining $p-1$ functions with a single common bandwidth, h_0, for every j. In the second step, a local cubic smoother is again applied to the residuals $Y_i - \sum_{j \neq k} X_{ik}\tilde{\beta}(z_0)$ using function-specific bandwidth to obtain the final estimate of $\beta_j(z_0)$. They present the asymptotic mean-squared error of the estimates obtained by this procedure, and further show that the estimates achieve optimal convergence rates. Cai (2000) demonstrated that even when every coefficient function exhibits the same degree of smoothness, the two-step estimates exhibit the same asymptotic properties as the usual one-step local smoother.

4.2 Kernel estimation with multiple smoothing variables

A proposed extension of model 2 permits each coefficient function to depend on its own smoothing variable:

$$E(Y|\boldsymbol{X} = \boldsymbol{x},\ Z = z) = x_1\beta_1(z_1) + \cdots + x_p\beta_p(z_p)$$

While the expression of the model itself does not make this obvious, estimation of this model is significantly different than the estimation of the model assuming a single common smoothing parameter for every coefficient function. Xue & Yang (2006a) further generalized this model where each coefficient function is replaced by a multivariate function with additive structure:

$$E(Y|\boldsymbol{X} = \boldsymbol{x},\ \boldsymbol{Z} = \boldsymbol{z}) = x_1 \sum_{j=1}^{q} \beta_{1j}(z_1) + \cdots + x_p \sum_{j=1}^{q} \beta_{pj}(z_p) \tag{27}$$

which allows for inclusion of all interaction terms $X_j\beta_{jk}(Z_k)$, $j = 1, \ldots, p$, $k = 1, \ldots, q$. Applying multivariate kernel smoothing locally to each point $\boldsymbol{z} = (z_1, \ldots, z_p)^T$ results in multivariate functions of the entire covariate vector, losing the structure of model 3. To extract proper estimates of the $\{\beta_j\}$, two primary methodologies have been proposed: marginal integration and smooth backfitting. Linton and Nielsen (1995) employ local kernel smoothing to estimate the multivariate coefficient functions $\{\beta_j(\boldsymbol{z})\}$, minimizing

$$n^{-1} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{q} \alpha_j X_{ij} \right)^2 K_{h_1}(z_1, Z_{i1}) \times \cdots \times K_{h_p}(z_p, Z_{ip})$$

for each value of . Integrating the multivariate coefficient functions over the support of the smoothing variables gives marginal estimates of β_j. This approach, however, suffers from the curse of dimensionality, as the attractive statistical properties of the estimators $\hat{\beta}_j$ depend heavily on the consistency of the $\{\alpha_j\}$, which requires $n \times h_1 \times \cdots \times h_p \to \infty$, thus losing the attractive qualities of local methods. The smooth backfitting method initially introduced by Mammen et al. (1999) for additive regression models enjoys both theoretical and numerical advantages over the integration method, and is free of the curse of dimensionality. To estimate $\{\alpha_j\}$, one minimizes the integrated weighted sum of squares

$$\int n^{-1} \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} \alpha_j \left(z_j \right) \right)^2 K_{h_1} \left(z_1, Z_{i1} \right) \times \cdots \times K_{h_p} \left(z_p, Z_{ip} \right) \, d\boldsymbol{z}$$

over the space of function tuples $\mathcal{H} = \{\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p) : \alpha_j (\boldsymbol{z}) = \alpha_j (z_j)\}$, so that the optimization must not be performed for every $\boldsymbol{z}$. For a detailed discussion of these methods, we refer the reader to Linton and Nielsen (1995) and Mammen & Park (2005).

4.3 Basis expansions and penalized likelihood techniques

In classical nonparametric regression problems, $m(x) = E(Y|X)$ is represented by a linear basis expansion in X, so that

$$E(Y|X) = \sum_{j=1}^{M} \beta_m b_m (X)$$

with $M \to \infty$. The general estimation framework may be described as follows: The estimation space, $G = G_n$, is the linear space of bounded functions having finite dimension M_n. For a given loss function $\mathcal{L}$, the estimate of m, $\hat{m}$, is defined to be the element of G_n which minimizes $\mathcal{L}$ and maybe be characterized by the estimates of the basis function coefficients $\beta_1, \ldots, \beta_m$. It is typical that the true mean function does not belong to G_n, and members of G_n are taken to be an approximation to the truth. Typical choices for loss functions include sums of squared errors or negative log likelihood functions. To this end, it is natural to allow the dimension of the estimation space to grow with the sample size. The choice of basis is not a trivial one, and some choices include logarithms, power functions, or wavelets; there is, however, disadvantages to using basis functions with unrestricted support. Piecewise polynomials and splines are families of functions with each member of which having bounded support. This allows for local representations of $m(x)$, while still permitting ease of implementation, as their estimation is carried out through the global optimization of $\mathcal{L}$.

These methods in the classical setting have been explored extensively; Chen (2007) provides an extensive review of the asymptotic behaviour of these estimators. Zhou, Shen, and Wolfe (1998) establish asymptotic normality of univariate regression splines; they present explicit expressions for the asymptotic pointwise bias and variance of the estimator, providing a method of constructing confidence intervals and confidence regions when the knots are asymptotically equally spaced and are distributed according to a continuous density. Their results additionally require that the order of the spline is equal to the order of the derivative of the unknown function to be estimated. Huang et al. (2003) establish asymptotic results for not only the univariate case, but also for tensor product splines and multivariate splines on triangulations.

A general representation of models 2, 3, and 27 may be represented as follows:

$$E(Y|\boldsymbol{X}, \boldsymbol{Z}) = \sum_{i=1}^{q} \boldsymbol{X}_i^T \boldsymbol{\beta}_i (\boldsymbol{Z}_i) \tag{28}$$

where $\boldsymbol{X}_i$ is a $d_i \times 1$ vector, $d_i \geq 1$; $\boldsymbol{X}$ is the collection of all covariates contained in $\{\boldsymbol{X}_i\}$, $i = 1, \ldots, q$. For example, model 27 may be written as above by letting $\boldsymbol{X}_i \equiv \boldsymbol{X} = (X_1, \ldots, X_p)^T$ for every j. The majority of the work in this area has been for the case where $q = 1$. Xue and Yang (2005a) allowed for multivariate coefficient functions, assuming an additive structure by letting

$$\begin{aligned} E(Y|\boldsymbol{X}, \boldsymbol{Z}) &= \sum_{i=1}^{d_1} X_i \beta_i(\boldsymbol{Z}) \\ \beta_i(\boldsymbol{Z}) &= \sum_{j=1}^{d_2} \beta_{ij}(Z_j) \end{aligned} \tag{29}$$

for $i = 1, \ldots, d_1$.

4.4 Smoothing methods with longitudinal data

Models 2, 3, and 27 can be written as follows:

$$Y(t) = \sum_{j=1}^{q} \boldsymbol{X}_j^T \boldsymbol{f}(T) + \epsilon(T) \tag{30}$$

where $\boldsymbol{f} = (f_1, \ldots, f_q)^T$ is the vector of coefficient functions of interest and $\epsilon(t)$ is a mean zero stochastic process. Both the response and covariates are assumed to be observed at subject-specific times, which may be irregularly spaced. Let $\boldsymbol{X}_{ij} = \boldsymbol{X}_i(T_{ij})$ and $Y_{ij} = Y_i(T_{ij})$ denote the observed covariates and responses on subject i at random time points $\{T_{ij}\}$, $j = 1, \ldots, n_i$. Given this structure, model 30 can be written

$$Y_{ij} = \boldsymbol{f}(T_{ij})^T \boldsymbol{X}_{ij} + \epsilon_{ij} \tag{31}$$

where $\epsilon_{ij} = \epsilon(T_{ij})$. The $\{T_{ij}\}$ are assumed to be independent for all i, j; $\boldsymbol{X}_{ij}$ and ϵ_{ij} are assumed to be independent across values of i, but may exhibit within-subject dependency structure. A simple avenue of model estimation for model **??** is to apply local smoothing, where the Nadaraya-Watson estimator minimizes

$$N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(Y_{ij} - \sum_{k=1}^{q} \alpha_k X_{ijk} \right)^2 K_h(t, T_{ij}) \tag{32}$$

with respect to $\alpha = (\alpha_1, \ldots, \alpha_q)^T$, where $N = \sum_{i=1}^{n} n_i$. The specification in 33 places equal weights on all subjects; to assign individual weights to each subject's contribution to the loss function, one may instead minimize

$$n^{-1} \sum_{i=1}^{n} w_i \sum_{j=1}^{n_i} \left(Y_{ij} - \sum_{k=1}^{q} \alpha_k X_{ijk} \right)^2 K_h(t, T_{ij}) \tag{33}$$

where one may specify, for example, $w_i = n_i^{-1}$. Hoover et al. (1998) proposed kernel estimation using local polynomial smoothing, of which the minimization of 33 is a special case. Wu et al present the construction of both point-wise confidence intervals as well as simultaneous confidence regions based on the asymptotic normality of the local kernel smoother.

References

[Anderson, 1973] Anderson, T. (1973). Asymptotically efficient estimation of covariance matrices with linear structure. *The Annals of Statistics*, pages 135–141.

[Banerjee et al., 2008] Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516.

[Bickel and Levina, 2008] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, pages 199–227.

[Buja et al., 1989] Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510.

[Cai and Yuan, 2010] Cai, T. and Yuan, M. (2010). Nonparametric covariance function estimation for functional and longitudinal data. *university of Pennsylvania and Georgia inistitute of technology*.

[Cai, 2002] Cai, Z. (2002). Two-step likelihood estimation procedure for varying-coefficient models. *Journal of Multivariate Analysis*, 82(1):189–209.

[Cai et al., 2000] Cai, Z., Fan, J., and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association*, 95(451):941–956.

[Cai and Tiwari, 2000] Cai, Z. and Tiwari, R. C. (2000). Application of a local linear autoregressive model to bod time series. *Environmetrics*, 11(3):341–350.

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

[Chen and Tsay, 1993b] Chen, R. and Tsay, R. S. (1993b). Nonlinear additive arx models. *Journal of the American Statistical Association*, 88(423):955–967.

[Chen, 2007] Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of econometrics*, 6:5549–5632.

[Chen et al., 2011] Chen, Z., Shi, M., W., G., and Tang, M. (2011). Efficient semiparametric estimation via cholesky decomposition for longitudinal data. *Computational Statistics and Data Analysis*, 55:677–690.

[Cheng and Wei, 2000] Cheng, S. and Wei, L. (2000). Inferences for a semiparametric model with panel data. *Biometrika*, 87(1):89–97.

[Chiang et al., 2001] Chiang, C.-T., Rice, J. A., and Wu, C. O. (2001). Smoothing spline estimation for varying coefficient models with repeatedly measured dependent variables. *Journal of the American Statistical Association*, 96(454):605–619.

[Darroch et al., 1980] Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980). Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics*, pages 522–539.

[Dempster, 1972] Dempster, A. P. (1972). Covariance selection. *Biometrics*, pages 157–175.

[Eilers, 1991a] Eilers, P. (1991a). Nonparametric density estimation with grouped observations. *Statistica neerlandica*, 45(3):255–269.

[Eilers, 1991b] Eilers, P. H. (1991b). Penalized regression in action: Estimating pollution roses from daily averages. *Environmetrics*, 2(1):25–47.

[Eilers, 1995] Eilers, P. H. (1995). Indirect observations, composite link models and penalized likelihood. In *Statistical Modelling*, pages 91–98. Springer.

[Eilers and Marx, 1996] Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102.

[Eilers and Marx, 2003] Eilers, P. H. and Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and intelligent laboratory systems*, 66(2):159–174.

[Fan, 1993] Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, pages 196–216.

[Fan et al., 2007] Fan, J., Huang, T., and Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association*, 102(478):632–641.

[Fan and Wu, 2008] Fan, J. and Wu, Y. (2008). Semiparametric estimation of covariance matrixes for longitudinal data. *Journal of the American Statistical Association*, 103(484).

[Fan and Zhang, 2000] Fan, J. and Zhang, J.-T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):303–322.

[Fan and Zhang, 1999] Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *Annals of Statistics*, pages 1491–1518.

[Friedman et al., 2008] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

[Friedman and Silverman, 1989] Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31(1):3–21.

[Gabriel, 1962] Gabriel, K. (1962). Ante-dependence analysis of an ordered set of variables. *The Annals of Mathematical Statistics*, pages 201–212.

[Gu, 2013] Gu, C. (2013). *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media.

[Härdle and Liang, 2007] Härdle, W. and Liang, H. (2007). Partially linear models. In *Statistical methods for biostatistics and related fields*, pages 87–103. Springer.

[Hastie and Tibshirani, 1986] Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical science*, pages 297–310.

[Hastie and Tibshirani, 1987] Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386.

[Hastie and Tibshirani, 1993] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796.

[Hastie and Tibshirani, 1990] Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43. CRC Press.

[Hoover et al., 1998] Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L.-P. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822.

[Huang, 1998] Huang, J. Z. (1998). Functional anova models for generalized regression. *Journal of multivariate analysis*, 67(1):49–71.

[Huang, 2001] Huang, J. Z. (2001). Concave extended linear modeling: a theoretical synthesis. *Statistica Sinica*, pages 173–197.

[Huang et al., 1998] Huang, J. Z. et al. (1998). Projection estimation in multiple regression with application to functional anova models. *The annals of statistics*, 26(1):242–272.

[Huang and Stone, 2003] Huang, J. Z. and Stone, C. J. (2003). Extended linear modeling with splines. In *Nonlinear Estimation and Classification*, pages 213–233. Springer.

[Huang et al., 2002] Huang, J. Z., Wu, C. O., and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89(1):111–128.

[Huang et al., 2004] Huang, J. Z., Wu, C. O., and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, pages 763–788.

[Kaslow et al., 1987] Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and Rinaldo, C. R. (1987). The multicenter aids cohort study: rationale, organization, and selected characteristics of the participants. *American journal of epidemiology*, 126(2):310–318.

[Lee et al., 2012] Lee, Y. K., Mammen, E., Park, B. U., et al. (2012). Flexible generalized varying coefficient regression models. *The Annals of Statistics*, 40(3):1906–1933.

[Levina et al., 2008] Levina, E., Rothman, A., and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, pages 245–263.

[Lin and Carroll, 2006] Lin, X. and Carroll, R. J. (2006). Semiparametric estimation in general repeated measures problems. *Journal of the Royal Statistical Society. Series B, statistical methodology*, 68:69–88.

[Linton and Nielsen, 1995] Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, pages 93–100.

[Mammen and Park, 2005] Mammen, E. and Park, B. U. (2005). Bandwidth selection for smooth backfitting in additive models. *Annals of Statistics*, pages 1260–1294.

[Marx and Eilers, 2005] Marx, B. D. and Eilers, P. H. (2005). Multidimensional penalized signal regression. *Technometrics*, 47(1):13–22.

[Meinhausen and Buhlmann, 2006] Meinhausen, N. and Buhlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462.

[O'Sullivan, 1986] O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical science*, pages 502–518.

[Peng et al., 2012] Peng, J., Wang, P., Zhou, N., and Zhu, J. (2012). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*.

[Pourahmadi, 1999] Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690.

[Rothman et al., 2008] Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

[Stone et al., 1997] Stone, C. J., Hansen, M. H., Kooperberg, C., Truong, Y. K., et al. (1997). Polynomial splines and their tensor products in extended linear modeling: 1994 wald memorial lecture. *The Annals of Statistics*, 25(4):1371–1470.

[Stone and Huang, 2002] Stone, C. J. and Huang, J. Z. (2002). Free knot splines in concave extended linear modeling. *Journal of Statistical Planning and Inference*, 108(1):219–253.

[Stone and Huang, 2003] Stone, C. J. and Huang, J. Z. (2003). Statistical modeling of diffusion processes with free knot splines. *Journal of statistical planning and inference*, 116(2):451–474.

[Tong et al., 1995] Tong, H., Chan, K., Cox, D., Cutler, C. D., Guégan, D., Jensen, J. L., Johansen, S., Lawrance, A., Lebaron, B., Ozaki, T., et al. (1995). A personal overview of non-linear time series analysis from a chaos perspective [with discussion and rejoinder]. *Scandinavian Journal of Statistics*, pages 399–445.

[Wahba, 1990] Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.

[Wu et al., 1998] Wu, C. O., Chiang, C.-T., and Hoover, D. R. (1998). Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American statistical Association*, 93(444):1388–1402.

[Wu and Pourahmadi, 2003] Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90(4):831–844.

[Yao et al., 2005] Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590.

[Yuan and Lin, 2007] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35.

[Zeger and Diggle, 1994] Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, pages 689–699.

[Zhang and Leng, 2012] Zhang, W. and Leng, C. (2012). A moving average cholesky factor model in covariance modelling for longitudinal data. *Biometrika*, 99(1):141–150.

[Zhou et al., 1998] Zhou, S., Shen, X., Wolfe, D., et al. (1998). Local asymptotics for regression splines and confidence regions. *The annals of statistics*, 26(5):1760–1782.

[Zimmerman and Nunez-Anton, 1997] Zimmerman, D. L. and Nunez-Anton, V. (1997). Structured antedependence models for longitudinal data. In *Modelling Longitudinal and Spatially Correlated Data*, pages 63–76. Springer.