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 $\left(X_j\right)$ in addition to the roles of the *smoothing variables* $\left(Z_j\right)$. One can see that by letting the $\{\beta_j\}$ be constant for $j=1,\ldots,p$, this reduces to 34 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

Using an approach similar to that described in Marx and Eilers (1999) and Eilers and Marx (2003), [Marx and Eilers, 2005] provide an extremely practical solution for functional linear models as presented in Ramsay and Silverman (1997, Chapters 10 and 11). We use the entire two-dimensional signal as regressors for model building, generalizing 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.

O'Sullivan proposed model fitting using 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, 1999], [Craven and Wahba, 1978], [Wahba, 1990] and Green and Silverman (1994). To regularize, they impose constraints which take into account the spatial structure of the regressors, while ensuring smoothness in the coefficient surface. We take two steps towards smoothness: (a) We purposely overfit the coefficient surface (not the signal) using two-dimensional tensor product B-splines, making the surface more flexible than needed. (b) We penalize estimation of the surface using difference penalties on each of the rows and columns of the tensor product B-spline coefficients.

The first step provides an initial reduction in parameter estimation through smoothness, as we will see that the tensor product B-splines are driven by relatively few parameters. The overfitting in this step is in the spirit of P-splines (Eilers and Marx, 1996) and is done to circumvent knot selection schemes. The second step ensures further smoothness, regularizing yet allowing general surfaces; The two tuning parameters associated with the row and column penalties allow for continuous control over the surface.

We term our approach presented in this article as Multidimensional Penalized Signal Regression (MPSR) and some of its gains include: (a) The entire signal can be used as regressors. (b) The number of highly spatially correlated regressors can far exceed the number of observations. (c) The parameterization (and the effective dimension) of the surface is dramatically reduced; the system of equations is manageable. (d) The candidate surface can be very general (non-additive), yet heavy penalization will yield polynomial surfaces. (e) Since the approach is grounded in standard (penalized) regression, delete-one diagnostics (e.g. cross-validation) are accessible. (f) The approach is easily transplanted to the generalized linear model (e.g. binary response) framework. (g) Since the two-dimensional signals and single estimated coefficient surface (and twice standard error surfaces) have a common indexing plane, potentially important regions can be visually identified.

P-splines simplify the work of O?Sullivan (1986). He noticed that if we model a function as a sum of B-splines, the familiar measure of roughness, the integrated squared second derivative, can be expressed as a quadratic function of the coefficients. P-splines go one step further: they use equally-spaced B-splines and discard the derivative completely. Roughness is expressed as the sum of squares of differences of coefficients. Differences are extremely easy to compute and generalization to higher orders is straightforward.

Before discussion of the proposed smoothing method, we first review the essential properties of B-spline basis functions. Figure 5 displays the essential building block to the proposed method: a bicubic basis function.

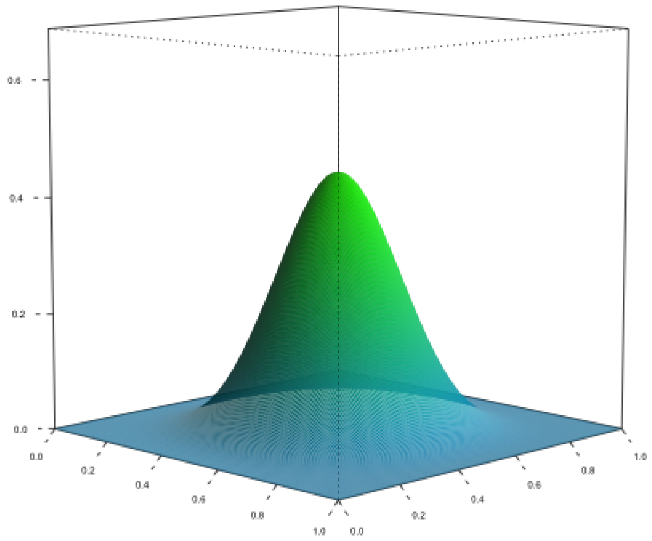

Figure 1: Tensor product of two cubic B-splines

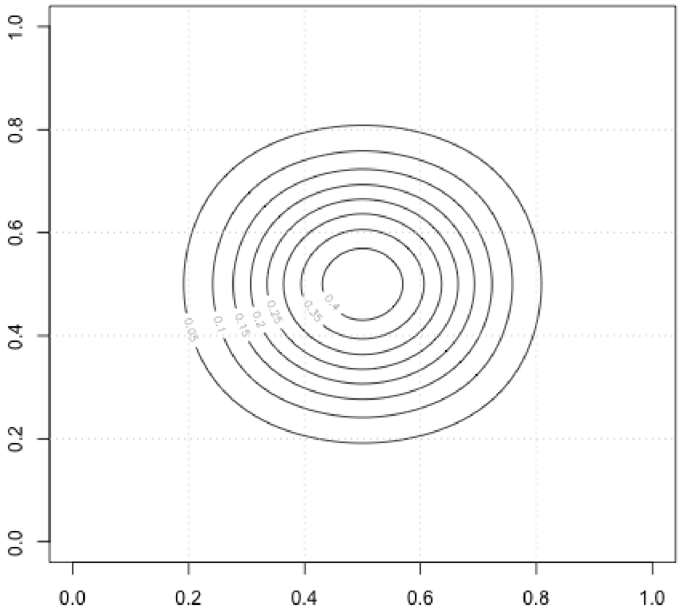

Figure 2: Tensor product of two cubic B-splines

Temporarily focusing on a single axis, we first present some specific details of B-splines in the univariate case.

3.2.1 Univariate B-splines

B-splines are constructed from polynomial pieces, joined at certain values of the domain t, called 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).

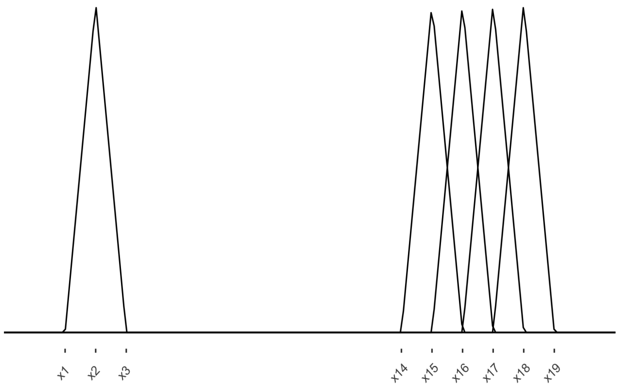

Figure 3: B-splines of degree $q = 1$

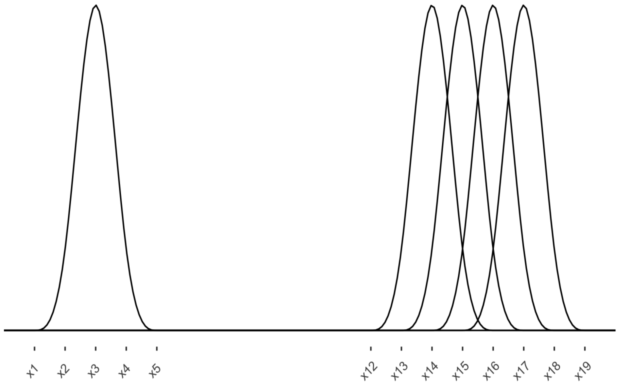

Figure 4: B-splines of degree $q = 3$

A B-spline consists of polynomial pieces connected in a particular way; a very simple example is shown in of Figure 3: five linear B-splines. One the left is an isolated B-spline of degree 1; it consists of two linear pieces: one piece from t_1 to t_2, the other from t_2 to t_3. The knots are t_1, t_2, and t_3. To the left of t_1 and to the right of t_3, this B-spline takes value zero. To the right on the same axis, four more B-splines of degree 1 are shown: each one based on three knots. More B-splines are added to the collection with the introduction of additional knots.

Figure 4 displays five B-splines of degree 3. It consists of four cubic pieces, joined at three inner knots. At the joining points of these cubic polynomials, not only the ordinates of the polynomial pieces match, but their first derivatives are also equal. (However, their second derivatives are not necessarily equal). The B-spline is based on five adjacent knots: $t_1, \ldots, t_5$. To the right, four additional cubic B-splines are shown.

First-degree B-splines overlap with two neighbors, second-degree B-splines with four neighbors, and so on, with the leftmost and rightmost splines have being exceptions to this rule. At a given t, two first-degree (or four cubic) 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 t, $q+1$ B-splines are nonzero.

We choose to divide the domain t_{min} to t_{max} into n' equal intervals, using $n'+1$ interior knots. Taking each boundary into consideration, a complete basis needs $n'+2q+1$ total knots, where q is the degree of the B-spline. This is easily verified by constructing graphs like those in Figure **??**. B-splines are very attractive as basis 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.

Denote the knots as: $\phi_1, \ldots, \phi_{n'+2q+1}$, and the collection of knots simply by ϕ. The total number of B-splines on the axis is $n = n' + q$. For indexing purposes it is convenient to associate each B-spline, $B_j\left(t\right)$ with exactly one of the $j = 1, \ldots, n$ knots. We denote a full basis using matrix B, which has dimension $m \times n$. These basic properties comprise the essential background necessary for our pursuits, but for comprehensive treatment, we refer the reader to de Boor (1978) and Dierckx (1993).

De Boor (1978) presented 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. For the sake of simplicity in presentation, we only discuss the construction and properties given equidistant knots, but de Boor's algorithm also works for any placement of knots.

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 becomes nonzero. Let $B_j\left(t;q\right)$ denote the value at t of the j^{th} B-spline of degree q for a given equidistant grid of knots. A fitted curve $\hat{y}$ to data $\left(t_i, y_i\right)$ is given by the linear combination $\hat{y}\left(t\right) = \sum_{j=1}^{n} \beta_j B_j\left(t;q\right)$. When the degree of the B-splines is clear from the context, or immaterial, we use $B_j\left(t\right)$ instead of $B_j\left(t;q\right)$.

De Boor (1978) gives a simple formula for derivatives of B-splines:

$$\begin{aligned} h \sum_j \beta_j B_j' (t,q) &= \sum_j \beta_j B_j (t, q-1) - \sum_j \beta_{j+1} B_{j+1} (t, q-1) \\ &= -\sum_j \Delta\beta_{j+1} B_j (t, q-1) \end{aligned} \tag{4}$$

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:

$$h^2 \sum_j \beta_j B_j'' (t,q) = \sum_j \Delta^2 \beta_j B_j (t, q-2) \tag{5}$$

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 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.

3.2.3 P-Spline Difference Penalties for Univariate B-Splines

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

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

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

$$\sum_{i=1}^{m} \left\{ y_i - \sum_{j=1}^{n} \beta_j B_j\left(t_i\right) \right\}^2 + \lambda \int_{t_{min}}^{t_{max}} \left\{ \sum_{j=1}^{n} \beta_j B_j''\left(t\right) \right\}^2 \, dt \tag{7}$$

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:

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

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:

$$h^2 P = \lambda \int_{t_{min}}^{t_max} \left\{ \sum_j \beta_j B_j'' \left(t, q = 3\right) \right\}^2 \, dt \tag{9}$$

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

$$h^2 P = \lambda \int_{t_{min}}^{t_max} \sum_j \sum_k \Delta^2 \beta_j \Delta^2 \beta_k B_j \left(t, q = 1\right) B_k \left(t, q = 1\right) dt \tag{10}$$

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

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

or

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

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

$$\begin{aligned} c_1 &= \int_{t_{min}}^{t_{max}} B_j^2 \left(t, 1\right) dt \\ c_2 &= \int_{t_{min}}^{t_{max}} B_j \left(t, 1\right) B_{j-1} \left(t, 1\right) dt \end{aligned} \tag{14}$$

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

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

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 Δ_k,and the elements of B are $b_{ij} = B_j(t_i)$. Taking derivatives on both sides of 15 with respect to β gives

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

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).

3.2.4 Properties of P-Splines

A fruitful way of looking at P-splines is to give the coefficients a central position as a skeleton, with the B-splines merely putting “the flesh on the bones.” This is illustrated in Figure 4. A smoother sequence of coefficients leads to a smoother curve. The number of splines and coefficients is immaterial, as long as the latter are smooth. The role of the penalty is to make such happen.

The number of B-splines can be (much) larger than the number of observations. The penalty makes the fitting procedure well-conditioned. This should be taken literally: even a thousand splines will fit ten observations without problems. Such is the power of the penalty. Figure 5 illustrates this for simulated data. There are 10 data points and 40 (+3) cubic B-splines. Unfortunately, this property of P-splines (and other types of penalized splines) is not generally appreciated. But one simply cannot have too many B-splines. A wise choice is to use 100 of them, unless computational constraints (in large models) come into sight. We will return to this example in Section 4, after introducing the effective model dimension, and further address this issue of many splines in Appendix B.

P-splines have a number of useful properties, partially inherited from B-splines. We give a short overview, with somewhat informal proofs. In the first place: P-splines show no boundary effects, as many types of kernel smoothers do. By this, we mean the spreading of a fitted curve or density outside of the (physical) domain of the data, generally accompanied by bending toward zero. In Section 8 this aspect is considered in some detail, in the context of density smoothing. P-splines can fit polynomial data exactly. Let data (t_i, y_i) be given. If the y_i are a polynomial in t of degree k, then B-splines of degree k or higher will exactly fit the data (de Boor, 1977).The same is true for P-splines, if the order of the penalty is $k + 1$ or higher, whatever the value of λ. To see that this is true, take the case of a first-order penalty and the fit to data y that are constant (a polynomial of degree 0). Because Cy=ldjBj(t) = c, we have that C)=,Ci B)(ti)= 0, for all t. Then it follows from the relationship between differences and derivatives in (1)that all $\Delta\beta_j$ are zero, and thus that C)=2Aaj = 0. Consequently, the penalty has no effect and the fit is the same as for unpenalized B-splines. This reasoning can easily be extended by induction to data with a linear relationship between t and y, and a second order difference penalty. P-splines can conserve moments of the data. For a linear model with P-splines of degree k +1 and a penalty of order k +1,or higher, it holds that

for all values of A, where 5i= C)=lbUfj are the fit- ted values. For GLM's with canonical links it holds that

This property is especially useful in the context of density smoothing: the mean and variance of the estimated density will be equal to mean and variance of the data, for any amount of smoothing. This is an advantage compared to kernel smoothers: these inflate the variance increasingly with stronger smoothing. The limit of a P-splines fit with strong smoothing is a polynomial. For large values of h and a penalty of order k, the fitted series will approach a polynomial of degree $k-1$, if the degree of the B-splines is equal to, or higher than, k. Once again, the relationships between derivatives of a B-spline fit and differences of coefficients, as in (1) and (2), are the key. Take the example of a second-order difference penalty: when h is large, Cy=3(A2aj) 2 has to be very near zero. Thus each of the second differences has to be near zero, and thus the second derivative of the fit has to be near zero everywhere. In view of these very useful results, it seems that B-splines and difference penalties are the ideal marriage.

It is important to focus on the linearized smoothing problem that is solved at each iteration, because we will make use of properties of the smoothing matrix. From (16) follows for the hat matrix H:

The trace of H approaches k as λ increases. A proof goes as follows: let...

3.2.5 Model Selection and Tuning

Now that we can easily influence the smoothness of a fitted curve with A, we need some way to choose an "optimal" value for it. We propose to use the Akaike information criterion (AIC). The basic idea.of AIC is to correct the log- likelihood of a fitted model for the effective number of parameters. An extensive discussion and applications can be found in Sakamoto, Ishiguro and Kitagawa (1986). Instead of the log-likelihood, the deviance is easier to use. The definition of AIC is equivalent to

where dim(a, A) is the (effective) dimension of the vector of parameters, a , and dev(y; a , A) is the deviance. Computation of the deviance is straightforward, but how shall we determine the effective dimension of our P-spline fit? We find a solution in Hastie and Tibshirani (1990). They discuss the effective dimensions of linear smoother and propose to use the trace of the smoother matrix as an approximation. In our case that means dim(a) = tr(H). Note that $tr\left(H\right) = n$ when $\lambda = 0$, as in (nonsingular) standard linear regression.

As $tr\left(AB\right) = tr\left(BA\right)$ (for conformable matrices), it is computationally advantageous to use

The latter expression involves only n-by-n matrices, whereas H is an $m \times m$ matrix. An estimate of the variance is needed, and one approach is to take the variance of the residuals from the $\hat{y}_i$ that are computed when $\lambda = 0$, say, $\hat{\sigma}_0^2$:

$$\text{AIC}\left(\lambda\right) = \sum_{i=1}^{m} \frac{\left(y_i - \hat{\mu}_i\right)^2}{\hat{\sigma}_0^2} + 2tr\left(H\right) - 2m\log\left(\hat{\sigma}_0\right) - m\log 2\pi \tag{18}$$

This choice for the variance is rather arbitrary, as it depends on the number of knots. Alternatives can be based on (generalized) cross-validation. For ordinary cross-validation we compute

$$\text{CV}\left(\lambda\right) = \sum_{i=1}^{m} \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}}\right)^2 \tag{19}$$

where the h_{ii} are the diagonal elements of the hat matrix, H. For generalized cross-validation (Wahba, 1990), we compute

$$\text{GCV}\left(\lambda\right) = \sum_{i=1}^{m} \frac{\left(y_i - \hat{y}_i\right)^2}{\left(m - \sum_{j=1}^{m} h_{jj}\right)^2} \tag{20}$$

The difference between both quantities is generally small. The best λ is the value that minimizes $\text{CV}\left(\lambda\right)$ or $\text{GCV}\left(\lambda\right)$. The variance of the residuals at the optimal λ is a natural choice to use as an estimate of $\hat{\sigma}_0^2$ for the computation of $\text{AIC}\left(\lambda\right)$. It is practical to work with modified versions of $\text{CV}\left(\lambda\right)$ and $\text{GCV}\left(\lambda\right)$, with values that can be interpreted as estimates of the cross-validation standard deviation:

$$\begin{aligned} \text{C}\bar{\text{V}}\left(\lambda\right) &= \sqrt{m^{-1}\text{CV}\left(\lambda\right)} \\ \text{GC}\bar{\text{V}}\left(\lambda\right) &= \sqrt{m\text{GCV}\left(\lambda\right)} \end{aligned} \tag{21}$$

The two terms in $\mathrm{AIC}(\lambda)$ represent the deviance and the trace of the smoother matrix, respectively. The latter term, say $T(\lambda) = tr(H(\lambda))$, is of interest on its own, because it can be interpreted as the effective dimension of the fitted curve.

$T(\lambda)$ is useful to compare fits for different numbers of knots and orders of penalties, whereas λ can vary over a large range of values and has no clear intuitive appeal. We will show in an example below that a plot of AIC against T is a useful diagnostic tool. In the case of P-splines, the maximum value that $T(\lambda)$ can attain is equal to the number of B-splines (when $\lambda = 0$). The actual maximum depends on the number and the distributions of the data points. The minimum value of $T(\lambda)$ occurs when λ goes to infinity and is equal to the order of the difference penalty. This agrees with the fact that for high values of λ, the fit of P-splines approaches a polynomial of degree $k-1$.

3.3 Tensor Product B-splines

Tensor product B-splines exist in the $t \times \tilde{t}$ plane. For our presentation, n $(\tilde{n})$ equally-spaced indexing knots are placed on the t $(\tilde{t})$ axis to yield a regularly-spaced grid, carving out the plane into sub-rectangles. The $r^{th} - s^{th}$ single tensor product $B_r(t)$ $(B_s(\tilde{t}))$, as presented in Figure 5 and Figure 6, is positive in the rectangular region defined by the knots $R = [\phi_r, \phi_{r+q+2}] \times \left[\tilde{\phi}_s, \tilde{\phi}_{s+\tilde{q}+2}\right]$ or on a support of spanned by $(q+2) \times (\tilde{q}+2)$ knots. Similar to univariate B-splines, we index each tensor product by one of the $n \times \tilde{n}$ knot pairs, where

$$\begin{aligned} B_r(\tilde{t})\, B_s(\tilde{t}) &> 0 \text{ for all } t,\ \tilde{t} \quad \in R \\ &= 0 \text{ for all } t,\ \tilde{t} \quad \notin R \end{aligned}$$

for $r = 1, \ldots, n$, $s = 1, \ldots, \tilde{n}$. Figure 7 displays nine tensor product B-splines, which represents only a portion of a full basis. A graphic of a complete basis would be of little illustrative use, as the“hills” overlap quite a lot, making each individual basis function difficult to isolate from the rest. Associated with each “hill” in Figure 7 , there is an unknown coefficient.

The complete tensor product B-spline basis thus has an unknown coefficient matrix, denoted by $\Gamma_{n\times\tilde{n}} = [\gamma_{rs}]$. For given knot grid, a very flexible surface can be approximated, e.g. at the digitized coordinates. For $j = 1, \ldots, p$ and $k = 1, \ldots, \tilde{p}$,

$$\alpha(t_j, \tilde{t}_k) = \sum_{r=1}^{n} \sum_{s=1}^{\tilde{n}} B_r(t_j)\, B_s(\tilde{t}_k)\, \gamma_{rs} \tag{22}$$

The surface is defined by relatively few parameters $(n\tilde{n})$, where changing Γ implies changes the surface.

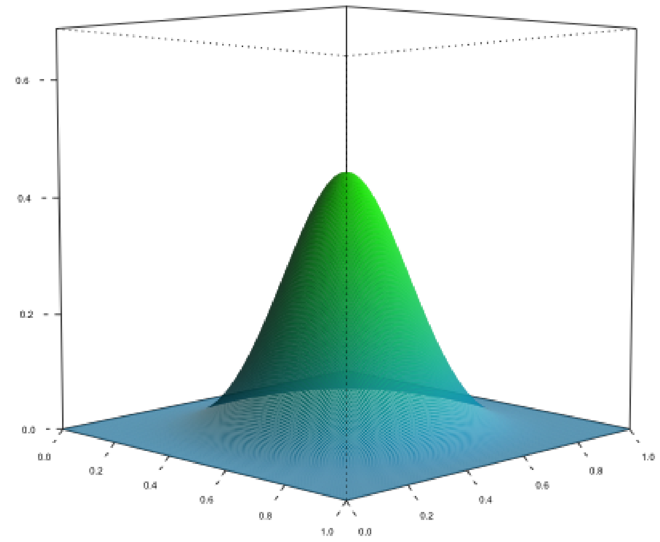

Figure 5: Tensor product of two cubic B-splines

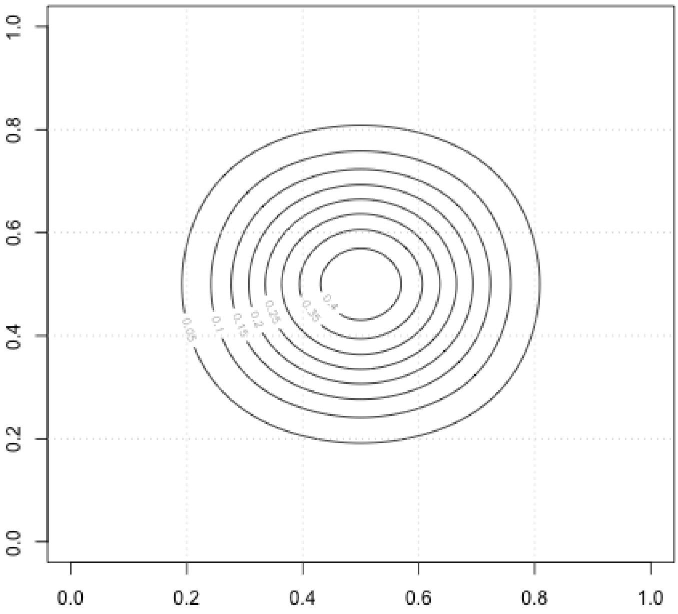

Figure 6: Tensor product of two cubic B-splines

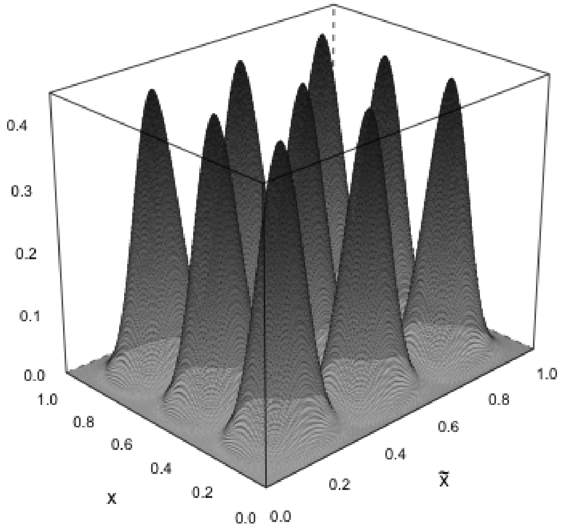

Figure 7: Landscape of cubic B-splines, a portion of a full bivariate basis

3.3.1 Computation

It is computationally efficient to express the surface in "unfolded" notation. Denote the support coordinate matrix $C = (v \otimes \mathbf{1}_{\tilde{p}}, \mathbf{1}_p \otimes \tilde{v})$ of dimension $p\tilde{p} \times 2$. Let B_1 and B_2 denote the matrices of dimensions $p\tilde{p} \times n$ and $p\tilde{p} \times \tilde{n}$ with entries corresponding to the univariate B-spline bases evaluated at each entry of the first and second column of C, respectively. The unfolded expression at the support coordinates then has the standard multiple regression form

$$\text{vec}\left\{\alpha\left(t, \tilde{t}\right)\right\} = T^{*}\gamma \tag{23}$$

where $\gamma = \text{vec}\left(\Gamma\right)$. Define T^{*} to be the matrix that is the result of taking the element-wise product of the matrices $(B_1 \otimes \mathbf{1}_n')$ and $(\mathbf{1}_{\tilde{n}}' \otimes B_2)$, written

$$T^{*} = (B_1 \otimes \mathbf{1}_n') \odot (\mathbf{1}_{\tilde{n}}' \otimes B_2) \tag{24}$$

where $\otimes$ denotes the Kronecker product operator.

3.4 Penalized Two-Dimensional Coefficient Surfaces

Let $X_i = [x_{ijk}]$ denote the i^{th} matrix of regression covariates, having dimension $p \times \tilde{p}$. Given X_i, the signal regressor support matrix C, and the corresponding coefficient surface $\alpha\left(t, \tilde{t}\right)$, we express the mean function

$$\mu_i = \sum_{j=1}^{p} \sum_{k=1}^{\tilde{p}} x_{ijk} \alpha \left(t_j, \tilde{t}_k \right) \tag{25}$$

where $i = 1, \ldots, m$, $j = 1, \ldots, p$, and $k = 1, \ldots, \tilde{p}$. Substituting 22 into 25, we can expression the mean regression function in terms of the B-spline basis expansion:

$$\mu_i = \sum_{j=1}^{p} \sum_{k=1}^{\tilde{p}} x_{ijk} \sum_{r=1}^{n} \sum_{s=1}^{\tilde{n}} B_r \left(t_j \right) B_s \left(\tilde{t}_k \right) \gamma_{rs} = \boldsymbol{x}' T^* \gamma \tag{26}$$

where $\boldsymbol{x}_i' = \text{vec}\left(X_i \right)$. A straightforward goodness of fit measure is the squared norm of the residual vector:

$$\mathcal{Q}\left(\gamma \right) = |y - \boldsymbol{X} T^* \gamma|^2 = |y - M\gamma|^2 \tag{27}$$

where we define $M = \boldsymbol{X} T^*$, and $\boldsymbol{X}$ is the $m \times p\tilde{p}$ matrix of vectorized regressors, $\boldsymbol{x}_i$, $i = 1, \ldots, m$. While expressing the coefficient surface in terms of the tensor product B-splines reduces the model dimension, there are still $n\tilde{n}$ unknown parameters to be estimated from the data. Even moderately complex surfaces may require increasing the number of knots on the grid to allow enough flexibility for adequate estimation. There may also be regions without measurements, leading to a sparse $\boldsymbol{X}$, containing many zeros. Imposing structure through a regularization term will alleviate these potential issues.

In the spirit of P-splines as discussed in section 3.2.3, discrete roughness or difference penalties are imposed on Γ. Although we implement penalization on the vector form of coefficients ?, the motivation and mechanics of penalization is perhaps best seen through the matrix of coefficients ?. In fact a separate difference penalty is assigned to each of its rows and each of its columns. The penalties have structure to effectively break the linkage in the penalty from row to row or from column to column. The objective function is now modified, using penalties, to minimize

$$\mathcal{Q}_\lambda \left(\gamma \right) = \text{goodness of fit measure} + \text{row penalty} + \text{column penalty} \tag{28}$$

We choose to measure the fit of the coefficient surface using the residual sums of squares. This specification permits ease of computation, particularly when the penalty function can be expressed as a quadratic function of the unknown parameters. The objective function is indexed by $\lambda = (\lambda_1, \lambda_2)$, which are penalty parameters which control the complexity of the fitted surface. To facilitate smoothness across both dimensions of the coefficient surface, the penalty term is comprised of two parts: one which places a difference penalty on the rows of Γ, and the second imposes the difference penalty on the columns of Γ. Specifically, we define $\hat{\gamma}$ to be the minimizer of

$$
\begin{aligned}
\mathcal{Q}_{\lambda}(\gamma) &= \sum_{i=1}^{m} (y_i - \mu_i)^2 + \lambda_1 \sum_{r=1}^{n} \gamma'_{r\bullet} D'_k D_k \gamma_{r\bullet} + \lambda_2 \sum_{s=1}^{\tilde{n}} \gamma'_{\bullet s} D'_{\tilde{k}} D_{\tilde{k}} \gamma_{\bullet s} \\
&= \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{p} \sum_{k=1}^{\tilde{p}} x_{ijk} \alpha\left(t_j, \tilde{t}_k\right) \right)^2 + \lambda_1 \sum_{r=1}^{n} \gamma'_{r\bullet} D'_k D_k \gamma_{r\bullet} + \lambda_2 \sum_{s=1}^{\tilde{n}} \gamma'_{\bullet s} D'_{\tilde{k}} D_{\tilde{k}} \gamma_{\bullet s} \\
&= |y - \boldsymbol{X} T^{*} \gamma|^2 + \lambda_1 \sum_{r=1}^{n} \gamma'_{r\bullet} D'_k D_k \gamma_{r\bullet} + \lambda_2 \sum_{s=1}^{\tilde{n}} \gamma'_{\bullet s} D'_{\tilde{k}} D_{\tilde{k}} \gamma_{\bullet s} \\
&= |y - M\gamma|^2 + \lambda_1 |P_1 \gamma|^2 + \lambda_2 |P_2 \gamma|^2
\end{aligned}
\tag{29}
$$

where $\gamma_{r\bullet}$ and $\gamma_{\bullet s}$ denote the r^{th} row and s^{th} column of Γ, respectively. Kronecker products and matrix notation allow for compact representation of each of the components of the penalty:

$$
\begin{aligned}
P_1 &= (D'_k D_k) \otimes I_{\tilde{n}} \\
P_2 &= I_n \otimes \left(D'_{\tilde{k}} D_{\tilde{k}}\right)
\end{aligned}
\tag{30}
$$

where I_d denotes the $d \times d$ identity matrix; k and $\tilde{k}$ denote the orders of the difference penalties imposed on the column and row dimensions, respectively. P_1 and P_2 have corresponding fixed dimensions $\left[\tilde{n}\,(n-k)\,\right] \times n\tilde{n}$ and $\left[n\left(\tilde{n}-\tilde{k}\right)\,\right] \times n\tilde{n}$. In principle, the order of the differences in the penalty term can be considered as additional hyper-parameters, but in practice are typically fixed by the user, as one would typically fix the order of the differencing as one fixes the order of the squared derivative, as in the specification of the ridge regression penalty, for example. Further discussion of this topic will be presented in sections to follow.

29 relies on two non-negative penalty parameters λ_1 and λ_2. These parameters allow for what can practically be viewed as continuous control of the smoothness of the surface in each dimension: the row dimension and the column dimension. This gives added flexibility, allowing for different degrees of smoothness across rows and columns. Figure 6 displays a possible scenario resulting from strong row (top panel) and strong column (bottom panel) penalization using a second order penalty on each row and column with large λ_1 and λ_2. Notice that the limiting behavior for each row and column is linear, but reversals of slopes are possible from one row (or column) to the next.

An example of a first and second order penalty matrix D for small $n = \tilde{n} = 3$ (for one row or column of Γ) looks like

$$
D_1 = \begin{bmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \quad D_2 = \begin{bmatrix} 1 & -2 & 1 \end{bmatrix}
$$

Each row and column of Γ has its own banded differencing matrix, D, so that the projection onto the penalized column space has structure

$$P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \otimes D_1 = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

and the projection operator corresponding to the row penalty is of the form

$$P_2 = D_1 \otimes \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ & & & & & & & & \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix}$$

One should also note that this formulation allows for seamless accommodation of different orders of differencing for rows and columns.

Taking derivatives of 29 with respect to γ, we have

$$\frac{\partial}{\partial \gamma} \mathcal{Q}_\lambda (\gamma) \quad = \quad \left[M'M + \lambda_1 P_1' P_1 + \lambda_2 P_2' P_2 \right] \gamma - M'y$$

and setting equal to zero, yields the explicit P-spline estimator

$$ga\hat{m}ma = \left(M'M + \lambda_1 P_1' P_1 + \lambda_2 P_2' P_2\right)^{-1} M'y \tag{31}$$

The predicted response values are then given by $\hat{y} = M\hat{\gamma}$, and the effective hat matrix is

$$H = M \left(M'M + \lambda_1 P_1' P_1 + \lambda_2 P_2' P_2\right)^{-1} M' \tag{32}$$

An attractive feature of this estimator is that the dimension of the system of normal equations remains fixed at $n\tilde{n}$ even while the resolution of the observation grid, $p\tilde{p}$, increases. The solution can be easily modified to include an intercept term, α_0 and its corresponding penalty. Additionally, one may include an overall ridge penalty:

$$\lambda_0 |\gamma|^2, \quad \lambda_0 > 0$$

To accommodate an intercept term, one simply augments the modified design matrix with a vector of ones, letting $\breve{M} = [\mathbf{1}|M]$. Letting $\breve{P}_i = [\mathbf{0}|]$ for $i = 1, 2$, and $\breve{I} = \operatorname{diag}(\mathbf{0}, I_{n\tilde{n}})$, the modified estimator is given by

$$(\hat{\alpha}_0, \hat{\gamma}')' = \left(\breve{M}'\breve{M} + \lambda_1 \breve{P}_1' \breve{P}_1 + \lambda_2 \breve{P}_2' \breve{P}_2 + \lambda_0 \breve{I}\right)^{-1} \breve{M}' y$$

where the addition of the zero vectors to the penalty projection matrices and the identity matrix ensure that the intercept term is not penalized. Advantages of this approach is that we sidestep the choice of the number and placement of the spline knots, which is a difficult optimization problem in itself. The penalty terms introduce little additional computational complexity, since the smoothing parameters have no impact on $\tilde{M}'M$ and $\tilde{M}'y$ and thus do not need to be recomputed when these parameters are varied. For fixed k, as λ_i, $i = 1, 2$ increases γ becomes smoother in its corresponding dimension, approaching a polynomial of degree $k - 1$. As the hat matrix is easily obtained, leave-one-out cross validation error estimates are also easy to obtain.

3.5 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

$$Y_{ij}(t) = x_{ij}^T \beta + \mu(t) + W_i(t) + \epsilon_{ij} \tag{33}$$

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:

$$Y(t) = \boldsymbol{X}^T(t)\, \boldsymbol{\beta}(t) + \epsilon(t) \tag{34}$$

proposing estimation of the coefficient functions via smoothing splines and local polynomials. $\epsilon(t)$ is defined as in 33 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 33 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

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{35}$$

where f is the unknown smooth regression function of interest, and the $\{\epsilon_i\}$ are mutually independent mean-zero errors, with $Var(\epsilon_i) = \sigma^2_\epsilon$. 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:

$$\begin{aligned} f(x) = E(Y|X = x) &= \int yp(y|x)\,dy \\ &= \frac{\int yp(y|x)\,dy}{\int p(y|x)\,dy} \end{aligned} \tag{36}$$

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

$$\begin{aligned}\hat{p}\left(x,y\right) &= \frac{1}{nh_xh_y}\sum_{i=1}^{n}K_{h_x}\left(\frac{x-x_i}{h_x}\right)K_{h_y}\left(\frac{y-y_i}{h_y}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}K_{h_x}\left(x-x_i\right)K_{h_y}\left(y-y_i\right) \qquad (37)\end{aligned}$$

Then, substituting 37 for $p\left(x,y\right)$ in the numerator of 36, we can write

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

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

$$\int y\hat{p}\left(x,y\right)\ dy = \frac{1}{n}\sum_{i=1}^{n}K_{h_x}\left(x-x_i\right)y_i \qquad (38)$$

Estimating the denominator of 36 in similar fashion, we have

$$\begin{aligned}\int \hat{p}\left(x,y\right)\ dy &= \frac{1}{n}\sum_{i=1}^{n}K_{h_x}\left(x-x_i\right)\int K_{h_y}\left(y-y_i\right)\ dy \\ &= \frac{1}{n}\sum_{i=1}^{n}K_{h_x}\left(x-x_i\right) \\ &= \hat{f}_x\left(x\right) \qquad (39)\end{aligned}$$

Using 38 and 39 as plug-in estimators in 36, then

$$\hat{f}\left(x\right) = \sum_{i=1}^{n}W_{h_x}\left(x,x_i\right)y_i \qquad (40)$$

where

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

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

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

with respect to $\boldsymbol{\alpha} = \left(\alpha_1,\ldots,\alpha_p\right)^T$ 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

$$\sum_{i=1}^{n} \left[Y_i - \sum_{j=1}^{p} \left(\alpha_{0j} + \alpha_{1j}(Z_i - z_0)\right) X_{ij}\right]^2 K_{h_z}(z_0, Z_i) \tag{41}$$

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 $1 \leq j \leq p$, 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:

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

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:

$$E(Y|\boldsymbol{X} = \boldsymbol{x},\ \boldsymbol{Z} = \boldsymbol{z}) = x_1 \sum_{j=1}^{q} \beta_{1j}(z_1) + \cdots + x_p \sum_{j=1}^{q} \beta_{pj}(z_p) \qquad (42)$$

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

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

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 42 may be represented as follows:

$$E\left(Y|\boldsymbol{X},\boldsymbol{Z}\right)=\sum_{i=1}^{q}\boldsymbol{X}_i^T\boldsymbol{\beta}_i\left(\boldsymbol{Z}_i\right) \tag{43}$$

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 42 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

$$\begin{aligned} E\left(Y|\boldsymbol{X},\boldsymbol{Z}\right) &= \sum_{i=1}^{d_1} X_i\beta_i\left(\boldsymbol{Z}\right) \qquad (44)\\ \beta_i\left(\boldsymbol{Z}\right) &= \sum_{j=1}^{d_2}\beta_{ij}\left(Z_j\right) \end{aligned}$$

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

4.4 Smoothing methods with longitudinal data

Models 2, 3, and 42 can be written as follows:

$$Y\left(t\right)=\sum_{j=1}^{q}\boldsymbol{X}_j^T\boldsymbol{f}\left(T\right)+\epsilon\left(T\right) \tag{45}$$

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 $\boldsymbol{X}_{ij}=\boldsymbol{X}_i\left(T_{ij}\right)$ and $Y_{ij}=Y_i\left(T_{ij}\right)$

denote the observed covariates and responses on subject i at random time points $\{T_{ij}\}$, $j = 1, \ldots, n_i$. Given this structure, model 45 can be written

$$Y_{ij} = \boldsymbol{f}\left(T_{ij}\right)^T \boldsymbol{X}_{ij} + \epsilon_{ij} \tag{46}$$

where $\epsilon_{ij} = \epsilon\left(T_{ij}\right)$. 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

$$N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(Y_{ij} - \sum_{k=1}^{q} \alpha_k X_{ijk} \right)^2 K_h\left(t, T_{ij}\right) \tag{47}$$

with respect to $\alpha = \left(\alpha_1, \ldots, \alpha_q\right)^T$, where $N = \sum_{i=1}^{n} n_i$. The specification in 48 places equal weights on all subjects; to assign individual weights to each subject's contribution to the loss function, one may instead minimize

$$n^{-1} \sum_{i=1}^{n} w_i \sum_{j=1}^{n_i} \left(Y_{ij} - \sum_{k=1}^{q} \alpha_k X_{ijk} \right)^2 K_h\left(t, T_{ij}\right) \tag{48}$$

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 48 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.

[Craven and Wahba, 1978] Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403.

[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.

[Eubank, 1999] Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. CRC press.

[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.