

1 Nonparametric extensions of the classical linear model

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

$$Y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (1)$$

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 & Tsay 1993a, Hastie & Tibshirani 1990, Sperlich, Tjostheim & Yang 2002, Stone 1985), partially linear models (Hardle, Liang & Gao 2000, Zeger and Diggle 1994), varying coefficient models (Hastie & 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 and the ‘curse of dimensionality’ can be avoided. Another advantage of this model is its interpretability. It 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

$$E(Y|\mathbf{X} = \mathbf{x}, Z = z) = x_1 \beta_1(z) + \dots + x_p \beta_p(z) \quad (2)$$

where $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$ and Z are covariates and $\boldsymbol{\beta}(z) = (\beta_0(z), \beta_1(z), \dots, \beta_p(z))^T$ are unknown coefficient functions, assumed to be smooth functions of Z . The cases for which Z is both univariate as well as multivariate have been considered. 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 & J. T. Zhang, 2000).

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, $\mathbf{Z} = (Z_1, Z_2, \dots, Z_p)^T$. This leads to modeling the mean response as follows:

$$E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = x_1\beta_1(z_1) + \dots + x_p\beta_p(z_p) \quad (3)$$

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 $j = 1, \dots, p$, this reduces to 5 proposed by Hoover, Rice, Wu and Yang. The class of models having the form as specified in 3 is quite extensive. 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.

Zeger and Diggle (1994) present a partially linear model motivated by the MACS data. They consider data of the form $\{(x_{ij}, y_{ij}(t_{ij})) : j = 1, \dots, m_i; i = 1, \dots, n\}$, where x_{ij} denotes a $p \times 1$ vector of covariates corresponding to $y_{ij}(t_{ij})$, the j th measurement on the i th subject at time t_{ij} . They propose the semiparametric model

$$Y_{ij}(t) = x_{ij}^T \beta + \mu(t) + W_i(t) + \epsilon_{ij} \quad (4)$$

where $\mu(t)$ is a smooth function of time, and β is a $p \times 1$ vector of regression coefficients. The $\{W_i(t) : i = 1, \dots, n\}$ 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 $\{Z_{ij} : j = 1, \dots, m_i; i = 1, \dots, n\}$ 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:

$$Y(t) = \mathbf{X}^T(t) \boldsymbol{\beta}(t) + \epsilon(t) \quad (5)$$

proposing estimation of the coefficient functions via smoothing splines and local polynomials. $\epsilon(t)$ is defined as in 4 and is assumed to be independent of $\mathbf{X}(t)$. 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.

1.1 Nonparametric approaches to modeling nonlinear time series data

For nonlinear time series applications, Chen & Tsay (1993) 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.

2 Model estimation

Zeger and Diggle (1994) carry out estimation of $\mu(t)$ and β as defined in model 4 iteratively via kernel smoothing and generalized least squares. While more flexible than the classical linear model, this is 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 $\beta_0(z), \beta_1(z), \dots, \beta_p(z)$ may need varying degrees of smoothing in the z dimension. To address this,

2.1 Kernel estimation with a single smoothing variable

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

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, \dots, n \quad (6)$$

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 :

$$\begin{aligned} f(x) = E(Y|X=x) &= \int yp(y|x) dy \\ &= \frac{\int yp(y|x) dy}{\int p(y|x) dy} \end{aligned} \quad (7)$$

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

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

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

$$\begin{aligned}\hat{p}(x, y) &= \frac{1}{nh_x h_y} \sum_{i=1}^n K_{h_x} \left(\frac{x - x_i}{h_x} \right) K_{h_y} \left(\frac{y - y_i}{h_y} \right) \\ &= \frac{1}{n} \sum_{i=1}^n K_{h_x}(x - x_i) K_{h_y}(y - y_i)\end{aligned}\tag{8}$$

Then, substituting 8 for $p(x, y)$ in the numerator of 7, we can write

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

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

$$\int y \hat{p}(x, y) dy = \frac{1}{n} \sum_{i=1}^n K_{h_x}(x - x_i) y_i\tag{9}$$

Estimating the denominator of 7 in similar fashion, we have

$$\begin{aligned}\int \hat{p}(x, y) dy &= \frac{1}{n} \sum_{i=1}^n K_{h_x}(x - x_i) \int K_{h_y}(y - y_i) dy \\ &= \frac{1}{n} \sum_{i=1}^n K_{h_x}(x - x_i) \\ &= \hat{f}_x(x)\end{aligned}\tag{10}$$

Using 9 and 10 as plug-in estimators in 7, then

$$\hat{f}(x) = \sum_{i=1}^n W_{h_x}(x, x_i) y_i\tag{11}$$

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 $\beta(z_0) = (\beta_0(z_0), \beta_1(z_0), \dots, \beta_p(z_0))^T$ 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 $\alpha = (\alpha_1, \dots, \alpha_p)^T$ 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 $\mathbf{Z} = (Z_1, \dots, Z_n)^T$. Further, let $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, then the NW estimator has form

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

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

$$\sum_{i=1}^n \left[Y_i - \sum_{j=1}^p (\alpha_{0j} + \alpha_{1j} (Z_i - z_0)) X_{ij} \right]^2 K_{h_z}(z_0, Z_i) \quad (12)$$

with respect to $\boldsymbol{\alpha}_0 = (\alpha_{01}, \dots, \alpha_{0p})^T$, and $\boldsymbol{\alpha}_1 = (\alpha_{11}, \dots, \alpha_{1p})^T$. 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

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

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

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

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.

2.1.1 Kernel bandwidth selection with a single smoothing variable

2.1.2 Asymptotic properties of kernel estimators with a single smoothing variable

2.1.3 Two-step estimation for multiple bandwidths

Model selection as described in 2.1.1 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 $\{\tilde{b}_{0j}, \tilde{b}_{1j}\}$, $j = 1, \dots, p - 1$ and $\tilde{b}_{0p}, \tilde{b}_{1p}, \tilde{b}_{2p}, \tilde{b}_{3p}$ be the minimizers of the weighted sums of squares:

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

If we take $\tilde{\beta}_p^{os}(z_0) = \tilde{b}_{0p}$, then they show that the bias of the *one-step estimator* is $O(h_0^2)$ and the variance is $O((nh_0)^{-1})$. 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 $Y_i - \sum_{j \neq k} X_{ik} \tilde{\beta}_j(z_0)$ 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.

2.2 Kernel estimation with multiple smoothing variables

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

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

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:

$$E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = x_1 \sum_{j=1}^q \beta_{1j}(z_1) + \dots + x_p \sum_{j=1}^q \beta_{pj}(z_p) \quad (13)$$

which allows for inclusion of all interaction terms $X_j\beta_{jk}(Z_k)$, $j = 1, \dots, p$, $k = 1, \dots, q$. Applying multivariate kernel smoothing locally to each point $\mathbf{z} = (z_1, \dots, z_p)^T$ 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 $\{\beta_j(\mathbf{z})\}$, 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 \cdot . 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 $n \times h_1 \times \cdots \times h_p \rightarrow \infty$, 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

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

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

2.3 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

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

with $M \rightarrow \infty$. 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 β_1, \dots, β_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 13 may be represented as follows:

$$E(Y|\mathbf{X}, \mathbf{Z}) = \sum_{i=1}^q \mathbf{X}_i^T \boldsymbol{\beta}_i(\mathbf{Z}_i) \quad (14)$$

where \mathbf{X}_i is a $d_i \times 1$ vector, $d_i \geq 1$; \mathbf{X} is the collection of all covariates contained in $\{\mathbf{X}_i\}$, $i = 1, \dots, q$. For example, model 13 may be written as above by letting $\mathbf{X}_i \equiv \mathbf{X} = (X_1, \dots, X_p)^T$ 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

$$\begin{aligned} E(Y|\mathbf{X}, \mathbf{Z}) &= \sum_{i=1}^{d_1} X_i \beta_i(\mathbf{Z}) \\ \beta_i(\mathbf{Z}) &= \sum_{j=1}^{d_2} \beta_{ij}(Z_j) \end{aligned} \quad (15)$$

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

2.4 Smoothing methods with longitudinal data

Models 2, 3, and 13 can be written as follows:

$$Y(t) = \sum_{j=1}^q \mathbf{X}_j^T \mathbf{f}(T) + \epsilon(T) \quad (16)$$

where $\mathbf{f} = (f_1, \dots, f_q)^T$ 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 $\mathbf{X}_{ij} = \mathbf{X}_i(T_{ij})$ and $Y_{ij} = Y_i(T_{ij})$ denote the observed covariates and responses on subject i at random time points $\{T_{ij}\}$, $j = 1, \dots, n_i$. Given this structure, model 16 can be written

$$Y_{ij} = \mathbf{f}(T_{ij})^T \mathbf{X}_{ij} + \epsilon_{ij} \quad (17)$$

where $\epsilon_{ij} = \epsilon(T_{ij})$. The $\{T_{ij}\}$ are assumed to be independent for all i, j ; \mathbf{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

$$N^{-1} \sum_{i=1}^n \sum_{j=1}^{n_i} \left(Y_{ij} - \sum_{k=1}^q \alpha_k X_{ijk} \right)^2 K_h(t, T_{ij}) \quad (18)$$

with respect to $\alpha = (\alpha_1, \dots, \alpha_q)^T$, where $N = \sum_{i=1}^n n_i$. The specification in 19 places equal weights on all subjects; to assign individual weights to each subject's contribution to the loss function, one may instead minimize

$$n^{-1} \sum_{i=1}^n w_i \sum_{j=1}^{n_i} \left(Y_{ij} - \sum_{k=1}^q \alpha_k X_{ijk} \right)^2 K_h(t, T_{ij}) \quad (19)$$

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 19 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.