1 Estimation in functional varying-coefficient models

1.1 Overview

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

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

Blah blah blah, motivate extending classical linear models to VC models by elaborating the demands of longitudinal data encountered in natural and life sciences, biomedicine, and other social and life sciences, particularly clinical trials, like the AIDS cohort CD4 data.

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

$$Y_{ij}(t) = x_{ij}^T \beta + \mu(t) + W_i(t) + \epsilon_{ij} \tag{2}$$

where $\mu(t)$ is a smooth function of time, and β is a $p \times 1$ vector of regression coefficients. The $\{W_i(t) : \quad i = 1, \ldots, n\}$ 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.

They carry out estimation of $\mu(t)$ and β 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.

Varying coefficient models extend 1, allowing the effect of covariates as specified by model parameters to change with the value of the covariates themselves. They adopt the same easy of interpretability of the classical linear model and are inherently nonparametric; the general class of varying coefficient models is very flexible, including generalized additive models as a special case. Two general methods of constructing varying coefficient models have been employed in previous work; the first of which specifies a model such that all coefficients are dependent on a single common covariate. The mean function of the response Y take the form

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

where $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$ and Z are covariates and $\boldsymbol{\beta}(z) = (\beta_0(z), \beta_1(z), \ldots, \beta_p(z))^T$ 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. Hoover, Rice, Wu and Yang (1998) considered the following model:

$$Y(t) = \boldsymbol{X}^T(t) \boldsymbol{\beta}(t) + \epsilon(t) \tag{4}$$

proposing estimation of the coefficient functions via smoothing splines and local polynomials. $\epsilon(t)$ is defined as in 2 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.

The second approach in specifying varying coefficient models is by generalizing 3 to allow each covariate's coefficient function to depend on different covariates, $Z = (Z_1, Z_2, \ldots, Z_p)^T$. This leads to modeling the mean response as follows:

$$E(Y|\boldsymbol{X} = \boldsymbol{x},\ Z = z) = x_1\beta_1(z_1) + \cdots + x_p\beta_p(z_p) \tag{5}$$

There are many proposed extensions of 3 and 5, including models that allow a covariate to play both the roles of the linear effect covariate (X_j) in addition to the roles of the *smoothing variables* (Z_j). One can see that by letting the $\{\beta_j\}$ be constant for $j = 1, \ldots, p$, this reduces to 4 proposed by Hoover, Rice, Wu and Yang.

2 Model estimation

In the case of a single common smoothing variable, estimation of 3 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,

2.1 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

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{6}$$

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:

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

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

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

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

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

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

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

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

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

Estimating the denominator of 7 in similar fashion, we have

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

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

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

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 3; the Nadaraya-Watson (NW) estimator of $\boldsymbol{\beta}(z_0) = (\beta_0(z_0,),\beta_1(z_0,),\ldots,\beta_p(z_0,))^T$ minimizes

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

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

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

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

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

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.

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 $\{\tilde{b}_{0j}, \tilde{b}_{1j}\}$, $j = 1, \ldots, p-1$ and $\tilde{b}_{0p}, \tilde{b}_{1p}, \tilde{b}_{2p}, \tilde{b}_{3p}$ be the minimizers of the weighted sums of squares:

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

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.

2.2 Kernel estimation with multiple smoothing variables

A proposed extension of model 3 permits each coefficient function to depend on its own smoothing variable:

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

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

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

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 5. 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).

2.3 Smoothing spline estimation and penalized likelihood techniques

A general representation of models 3, 5, and 13 may be written

2.4 Smoothing methods

Models 3, 5, and 13 can be written as follows:

$$Y(t) = \sum_{j=1}^{q}\boldsymbol{X}_j^T\boldsymbol{f}(T) + \epsilon(T) \tag{14}$$

where $\boldsymbol{f} = (f_1,\ldots,f_q)^T$ is the vector of coefficient functions of interest and $\epsilon(t)$ is a mean zero stochastic process. Both the response and covariates are assumed to be observed at subject-specific times, which may be irregularly spaced. Let $\boldsymbol{X}_{ij} = \boldsymbol{X}_i(T_{ij})$ and $Y_{ij} = Y_i(T_{ij})$ denote the observed covariates and responses on subject i at random time points $\{T_{ij}\}$, $j = 1,\ldots,n_i$. Given this structure, model 14 can be written

$$Y_{ij} = \boldsymbol{f}\left(T_{ij}\right)^T \boldsymbol{X}_{ij} + \epsilon_{ij} \tag{15}$$

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

$$N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \left(Y_{ij} - \sum_{k=1}^{q} \alpha_k X_{ijk} \right)^2 K_h\left(t, T_{ij}\right) \tag{16}$$

with respect to $\alpha = \left(\alpha_1, \ldots, \alpha_q\right)^T$, where $N = \sum_{i=1}^{n} n_i$. The specification in 17 places equal weights on all subjects; to assign individual weights to each subject's contribution to the loss function, one may instead minimize

$$n^{-1} \sum_{i=1}^{n} w_i \sum_{j=1}^{n_i} \left(Y_{ij} - \sum_{k=1}^{q} \alpha_k X_{ijk} \right)^2 K_h\left(t, T_{ij}\right) \tag{17}$$

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 17 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.