

1 Estimation in functional varying-coefficient models

1.1 Overview

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

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

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 $\{(x_{ij}, y_{ij}(t_{ij})) : j = 1, \dots, m_i; i = 1, \dots, n\}$, 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

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

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

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

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

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 . 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:

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

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 $\mathbf{X}(t)$. Hoover et al (1998) propose the same model, using smoothing splines and kernel smoothing to estimate the components of $\beta(t)$ and develop asymptotic properties of kernel estimators.

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

$$E(Y|\mathbf{X} = \mathbf{x}, Z = z) = x_1\beta_1(z_1) + \dots + x_p\beta_p(z_p) \quad (5)$$

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 $j = 1, \dots, p$, 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 $\beta_0(z), \beta_1(z), \dots, \beta_p(z)$ 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

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

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 3; 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 $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_p)^T$ 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

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

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

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

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.

- 2.2 Kernel estimation with multiple smoothing variables
- 2.3 Smoothing spline estimation and penalized likelihood techniques