Advanced Statistics-20
Semiparametric regression-I

Changliang Zou
Institute of Statistics, Nankai University
Email: nk.chlzou@gmail.com
Outline

- Single-index model
- Varying coefficient model
- Partially linear model
- Additive model
Single-index model: motivation

- Project all explanatory variables onto a linear space $\beta^T X$ and then fit a univariate nonparametric curve

$$E(Y | X) = m(X) = g(\beta^T X)$$

- Identification Issues:
  - Intercept of the linear combination is not identifiable
  - The linear coefficients can only be identified up to scale
(Weighted) Semiparametric Least Squares (SLS, WSLS): minimize an objective function based on the conditional distribution of $Y$ and the linear index

Average Derivative Methods: directly estimate the parametric coefficients via the average derivative of the regression function

Outer product of gradients method (Xia, 2006)
Single-index model: Semiparametric Least Squares

Semiparametric least squares

$$\min_{\beta} \sum_{i=1}^{n} \left[ Y_i - E \left\{ Y_i \mid \beta^\top X_i \right\} \right]^2 \omega(X_i)$$

$$\hat{\beta} = \min_{\beta} \sum_{i=1}^{n} \left\{ Y_i - \hat{m}_\beta(X_i) \right\}^2 \omega(X_i)$$

$\hat{m}_\beta(X_i)$ is a leave-one-out estimator of $m$ given $\beta$.

- An iterative procedure
- $\hat{\beta}$ is asymptotically unbiased and converges with $\sqrt{n}$-rate.
- By choosing the weight $\omega(X) = 1/\sigma^2(X)$, the WSLS reaches the efficiency bound for the parameter estimates in SIM
Average Derivatives Estimation (ADE): Identify $\beta$ as the average derivative.

The vector of average derivatives

$$\delta = E\{\nabla_m(X)\} = E \left\{ g'(\beta^\top X) \right\} \beta,$$

where $\nabla_m(t) = \left\{ \frac{\partial m(t)}{\partial t_1}, \ldots, \frac{\partial m(t)}{\partial t_q} \right\}^\top$

- Do not need any distributional assumption on $Y$
- The resulting estimator is direct, i.e. non-iterative
- The ADE estimators are inefficient
- The ADE method is only applicable to continuous explanatory variables
Define score vector

\[ s(t) = \frac{\nabla f(t)}{f(t)} \]

\[ \delta = \int \nabla m(t)f(t) dt \]

\[ = - \int \frac{\nabla f(t)}{f(t)} f(t)m(t) dt \]

\[ = - \int s(t)m(t)f(t) dt \]

\[ = -E \{s(t)m(t)\} = -E \{s(t)Y\} , \]

where we use an assumption that \( f(t)m(t) \to 0 \) for \( \|t\| \to \infty \)
ADE by sample analog

\[ \hat{\delta} = -n^{-1} \sum_{i=1}^{n} \hat{s}(X_i) Y_i, \]

\[ \hat{s}(X) = \left\{ \hat{f}(X) \right\}^{-1} \left\{ \frac{\partial \hat{f}(t)}{\partial t_1}, \ldots, \frac{\partial \hat{f}(t)}{\partial t_q} \right\}^T \]

\( \hat{\delta} \) reaches the parametric rate of convergence.
Observe that

$$E \left\{ \nabla m(X) \nabla^\top m(X) \right\} = E \left[ \{g'(\beta^\top X)\}^2 \right] \beta \beta^\top$$

has only one nonzero eigenvalue.

Therefore, index $\beta$ is the eigenvector corresponding to the largest eigenvalue of $E \left\{ \nabla m(X) \nabla^\top m(X) \right\}$. 
We locally approximate the link function by a Taylor expansion
\[
g(\beta^\top X_i) \approx g(\beta^\top X_l) + g'(\beta^\top X_l)\beta^\top X_{il},
\]
where \(X_{il} = X_i - X_l\).
Denote \(a_l = g(\beta^\top X_l)\), \(b_l^\top = g'(\beta^\top X_l)\beta^\top\).

Considered the local linear fitting in the form of the following minimization problem:
\[
\min_{a_l, b_l} \sum_{i=1}^{n} (Y_i - a_l - b_l^\top X_{il})^2 w_{il},
\]
where \(w_{il}\) is a weight depending on the distance between \(X_i\) and \(X_l\). \(w_{il}\) can be taken as \(K_h(\beta^\top X_{il})\)
Calculate \( \hat{\Sigma} = n^{-1} \sum_{l=1}^{n} \hat{b}_{l, \beta_0} \hat{b}_{l, \beta_0}^{\top} \), and the first eigenvector of \( \hat{\Sigma} \), \( \hat{\beta} \), is an estimator of \( \beta \).

It is worth noting that the weight \( w_{il} = K_h(\beta_0^\top X_{il}) \) involves unknown parameter \( \beta_0 \).

A repeated procedure: set an initial estimate \( \beta_0 \) and update \( w_{il} \) with the latest estimator of \( b_l \) until convergence.
A partially linear model is given by

\[ y_i = \mathbf{X}_i^\top \beta + g(Z_i) + \epsilon_i, \quad i = 1, \ldots, n, \]

where \( E(\epsilon_i \mid \mathbf{X}_i, Z_i) = 0, \ E(\epsilon_i^2 \mid \mathbf{X}_i, Z_i) = \sigma^2(x, z) \) and \( Z_i \in \mathbb{R}^q \).

Goal: Estimate the parametric part \( \beta \) and nonparametric function \( g \).

Identification Issues:
- Intercept of the linear part or constant in nonparametric part is not identified.
- \( \beta \) can be identified when none of the components of \( \mathbf{X} \) can be a deterministic function of \( Z \).
Objective functions:

\[ \sum_{i=1}^{n} [y_i - X_i^T \beta - g(Z_i)]^2 + \lambda J(g) \]

where \( J(g) \) is a penalty for roughness in the fitted \( g \).

**Insights:** Apply the conditional expectation given \( Z_i \), we have

\[ E(y_i - X_i\beta \mid Z_i) = g(Z_i) \]
For every given $\beta$, use nonparametric method to estimate $g(z)$.

- Kernel smoothing (Speckman 1988); Spline smoothing (Engle et al. 1986); Local linear when $g(z)$ is second order differentiable:

  $$g(Z_i) \approx g(z) + g'(z)(Z_i - z);$$

  and so on ...

Equivalently, $g$ can be parameterized as

$$\{g(Z_1), \ldots, g(Z_n)\}^\top = S(y - X\beta),$$

where $S$ is a smoother matrix.
Smoothing Method (Speckman 1988)

- Replace \( g(Z_1), \ldots, g(Z_n) \) by \( S(y - X\beta) \) in the objective function;
- Use LSE or PLS to estimate \( \beta \) by

\[
\hat{\beta} = \left\{ X^\top (I - S)X \right\}^{-1} X^\top (I - S)y
\]

\[
\{ \hat{g}(Z_1), \ldots, \hat{g}(Z_n) \}^\top = S(y - X\hat{\beta})
\]

- \( \hat{\beta} \) is \( \sqrt{n} \)-rate consistent.
Robinson Estimator (Robinson 1988)

- Rewrite model by

\[ y_i - E(y_i \mid Z_i) = [X_i - E(X_i \mid Z_i)]^\top \beta + \varepsilon_i \]

- Use NW method to estimate \( E(y_i \mid Z_i) \) and \( E(X_i \mid Z_i) \),

\[ \hat{y}_i = \hat{E}(y_i \mid Z_i) = n^{-1} \sum_{j=1}^{n} K_h(Z_j - Z_i) y_j / \hat{f}(Z_i) \]

\[ \hat{X}_i = \hat{E}(X_i \mid Z_i) = n^{-1} \sum_{j=1}^{n} K_h(Z_j - Z_i) X_j / \hat{f}(Z_i) \]

\[ \hat{f}(Z_i) = n^{-1} \sum_{j=1}^{n} K_h(Z_j - Z_i) \]
Trim out small $\hat{f}(Z_i)$ by $l_i = I(\hat{f}(Z_i) > b)$ where $b$ is a given constant.

Use LSE to get the $\hat{\beta}$

$$\hat{\beta} = \left\{ \sum_{i=1}^{n} (x_i - \hat{x}_i)(x_i - \hat{x}_i)^\top l_i \right\}^{-1} \sum_{i=1}^{n} (x_i - \hat{x}_i)(y_i - \hat{y}_i) l_i.$$
An additive model is

\[ y = \alpha + \sum_{j=1}^{p} g_j(X_j) + \varepsilon \]

where \( \alpha \) is a constant and \( X_j \) is the \( j \)th element of \( X \), \( E(\varepsilon | X) = 0 \) and \( E(\varepsilon^2 | X) = \sigma^2 \) and \( g_j \)'s are smooth and unknown univariate functions.

Additive models generalize the linear regression models and allow interpretation of marginal changes, i.e., the effect of \( X_j \).

Identification Issues:
- Need to assume \( E[g_j(X_j)] = 0 \) or \( \alpha = E(y) \).
- Easily we can get \( \hat{\alpha} = \bar{y} \) with \( \sqrt{n} \)-rate. Without loss of generality, assume \( \alpha = 0 \) in the model.
Intuitive idea: Create a set of basic functions, and then perform multiple regression on these basic functions.

- Approximate $g_j(x)$ by

$$g_j(x) \approx g_j(x, \beta_j) = \sum_{l=1}^{k_j} \beta_{jl} K_l(x)$$

where $\beta_j = (\beta_{j1}, \ldots, \beta_{jk_j})^\top$ and $K_l(x)$ are some basic functions, e.g., B-splines, polynomial basic functions ($1, x, x^2, x^3, \ldots$) and Fourier basis functions $\{\sin(kx), \cos(kx)\}_{k=1}^\infty$.

- (LSE or PLS) Choose $\beta_j$ to minimize the following objective function

$$n^{-1} \sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{p} g_j(X_{ji}, \beta_j) \right\}^2$$

- The estimated functions are $\hat{g}_j(x) = g_j(x, \hat{\beta}_j), j = 1, \ldots, p.$
Additive Models: Intuitive idea

**Drawbacks:**

- A large set of basic functions need to be chosen to approximate $g_j(x)$. The dimensionality is reduced by *shrinkage regression*, like ridge regression. This results in a biased regression.

- To solve the least squares problem, a large parametric problem with an inversion of matrix of high order is needed. But it may not exist.

**Solutions:** Consider the conditional expectation given $X_j$. For any $j$,

$$g_j(X_j) = E(y - \sum_{k \neq j} g_k(X_k) | X_j).$$

This suggests an iterative algorithm for computing $g_j$. 
Additive Models: Backfitting Algorithm

**Backfitting Algorithm** (Hastie and Tibshirani 1990)

**Algorithm**

**Step 1** Initialize the functions $\hat{g}_j = g_j^0$ for $j = 1, \ldots, p$

**Step 2** For $j = 1, \ldots, p$, update $\hat{g}_j = S_j(y - \sum_{k \neq j} \hat{g}_k)$

**Step 3** Repeat **Step 2** until the individual functions convergence.

- $S_j$: nonparametric estimator by regressing the partial residuals $\{y_i - \sum_{k \neq j} \hat{g}_k\}$ on $X_j$.

- Without prior knowledge, the starting values can be linear regression of $y$ on the predictors.
Additive Models: Backfitting Algorithm

- Convergence is guaranteed, when all the smoothers $S_j$ are projection operators.
  - For example, NW estimator
    \[
    S_j(x) = \frac{\sum_{i=1}^{n} K_h(X_{ji} - x)\{y_i - \sum_{k\neq j} \hat{g}_k(X_{ki})\}}{\sum_{i=1}^{n} K_h(X_{ji} - x)}.
    \]

- Backfitting algorithm is equivalent to the following equation,
  \[
  \begin{pmatrix}
  I & S_1 & \ldots & S_1 \\
  S_2 & I & \ldots & S_2 \\
  \vdots & \vdots & \ddots & \vdots \\
  S_p & S_p & \ldots & I
  \end{pmatrix}
  \begin{pmatrix}
  \hat{g}_1 \\
  \hat{g}_2 \\
  \vdots \\
  \hat{g}_p
  \end{pmatrix}
  =
  \begin{pmatrix}
  S_1 \\
  S_2 \\
  \vdots \\
  S_p
  \end{pmatrix}
  \begin{pmatrix}
  y_1 \\
  y_2 \\
  \vdots \\
  y_n
  \end{pmatrix}
  \]

- No iteration is needed.
- But this equation involves inverting a $np \times np$ square matrix.
Additive Models: Marginal Integration Method

Marginal Integration Estimator (Chen et al. 1995)

- Notations: $f_j(\cdot)$ is the marginal density of $X_{ij}$, 
  $j = (1, \ldots, j-1, j+1, \ldots, p)^\top$, 
  $X_{i,j} = (X_{i1}, \ldots, X_{i,j-1}, X_{i,j+1}, \ldots, X_{ip})^\top$ and its joint density is $f_j(X_{-j})$.

- Denote $m(X_i) = \sum_{j=1}^{p} g_j(X_{ij})$, for a fixed $X$ we have,
  $$g_j(X_j) = \int m(X) f_j(X_{-j}) dX_{-j}$$

- Use NW method to estimate $m(\cdot)$ and average over the observations to obtain $\hat{g}_j$ by
  $$\hat{g}_j(x) = \frac{1}{n} \sum_{i=1}^{n} \tilde{m}(X_{i1}, \ldots, X_{i,j-1}, x, X_{i,j+1}, X_{ip}) = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{\sum_{l=1}^{n} K_{h_1}(X_{l,j} - X_{i,j})K_{h_2}(X_{lj} - x) Y_l}{\sum_{l=1}^{n} K_{h_1}(X_{l,j} - X_{i,j})K_{h_2}(X_{lj} - x)} \right]$$
Additive partially linear models: \( Y_i = X_i^T \beta + \sum_{j=1}^{q} g_j(Z_{ij}) + \varepsilon_i. \)

Thinking: How to estimate these functions?
Varying coefficient model

- Varying coefficient regression models: structure and interpretability are similar to those for the linear models.
- They are more flexible because of the infinite dimensionality of the corresponding parameter spaces.
- Consider the response and covariate values \((y_i, X_i, u_i), i = 1, \ldots, n\), are collected from the model

\[
y_i = X_i^\top \beta(u_i) + \varepsilon_i, \quad i = 1, \ldots, n,
\]

where \(E(\varepsilon_i \mid X_i, u_i) = 0\) and \(\beta(u) = \{\beta_1(u), \ldots, \beta_p(u)\}^\top\).
For each given $u$, the local linear estimator $\hat{\beta}(u)$ is the part corresponding to $a$ of the minimizer of

$$\sum_{i=1}^{n} \left\{ y_i - \mathbf{X}_i^\top a - \mathbf{X}_i^\top b(u_i - u) \right\}^2 K_h(u_i - u).$$

where $K_h(t) = K(t/h)/h$.

$\tilde{\mathbf{X}} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^\top$, $\mathbf{U}_u = \text{diag}(u_1 - u, \ldots, u_n - u)$,

$\Gamma_u = (\tilde{\mathbf{X}}, \mathbf{U}_u \tilde{\mathbf{X}})$, $\mathbf{Y} = (y_1, \ldots, y_n)^\top$,

$\mathbf{W}_u = \text{diag} \{ K_h(u_1 - u), \ldots, K_h(u_n - u) \}$.

$$\hat{\beta}(u) = (\mathbf{I}_p, \mathbf{0}_p) \left( \Gamma_u^\top \mathbf{W}_u \Gamma_u \right)^{-1} \Gamma_u^\top \mathbf{W}_u \mathbf{Y}.$$
There is an interesting (unique) issue arising from the estimation.

When the components of $\beta(\cdot)$ have different degrees of smoothness, how to estimate $\beta(\cdot)$?

Intuitively, the smoother components need larger bandwidth whilst the less smooth components need smaller bandwidth.

This means it is impossible to optimally estimate all components simultaneously with a single choice of the bandwidth.