Chapter 2. Semi-parametric Models (I) Part 5

March 6, 2007

1 Projection Pursuit Regression

Suppose that we have response variable Y and covraiates $\mathbf{x}_1, ..., \mathbf{x}_p$, we are interested in the conditional expectation function

$$m(x_1, ..., x_p) = E(Y|\mathbf{x}_1 = x, ..., \mathbf{x}_p = x_p)$$

in order to make prediction. As we have discussed that, the function $m(x_1, ..., x_p)$ is difficult to estimate because of the so-called "curse of dimensionality". Projection pursuit tries to approach the function m by a set of (univariate) ridge functions of the projection $\alpha^\top x$, i.e.

$$m(x) \approx g_1(\alpha_1^\top x) + ... + g_k(\alpha_k^\top x).$$

We call k the number of components, g_k the kth component

It was further proved that

$$\lim_{k\to\infty} E\{m(X) - [g_1(\alpha_1^\top X) + ... + g_k(\alpha_k^\top X)]\}^2 = 0.$$

2 estimation of PPR

Suppose (X_i, Y_i) are the observations. Consider the first component

$$Y_i = g_1(\alpha_1^\top X_i) + \xi_i$$

it is a single index model. We can estimate it by the method mentioned above. Suppose the estimate is $\hat{g}_1(\hat{\alpha}_1 x)$. Now consider

$$r_{1,i} = Y_1 - \hat{g}_1(\hat{\alpha}_1 X_i)$$

and fit the second component

$$r_{1,i} = g_2(\alpha_2 X_i) + \eta_i.$$

Suppose the estimate is $\hat{g}_2(\hat{\alpha}_2 x)$. Now consider

$$r_{2,i} = r_{1,i} - \hat{g}_2(\hat{\alpha}_2 X_i)$$

and fit the third component

$$r_{2,i} = g_3(\alpha_3 X_i) + \epsilon_i$$

Keep doing this, we can estimate all the components.

Example 2.1 (simulation) *Consider model*

$$Y = 4 * \mathbf{x}_1 * \mathbf{x}_2 + \varepsilon$$

where $\mathbf{x}_1, \mathbf{x}_2$ and ε are IID N(0,1). 100 observations are taken from the model. The estimated model is shown in Figure 1.

Example 2.2 (ozone data) *We fit PPR model with component 2 to the data. The estimated model is shown in Figure 2.*

A simple question is whether we need the second component?

Example 2.3 *For* data *about the baseball's players and their performance, we consider the following model*

$$Y = g_1(\alpha_1^\top X) + ... + g_k(\alpha_k^\top X) + \varepsilon.$$

The estimated model is shown in figure 3

A simple question is whether we need the third component?

2.1 prediction based on projection pursuit regression

Suppose we have estimated the model

$$\hat{Y} = \hat{g}_1(\hat{\alpha}_1^\top X) + ... + \hat{g}_k(\hat{\alpha}_k^\top X).$$

For a new data, $X' = (\mathbf{x}'_1, ..., \mathbf{x}'_p)$, we predict its response by

$$\hat{Y}' = \hat{g}_1(z'_1) + ... + \hat{g}_k(z'_k)$$

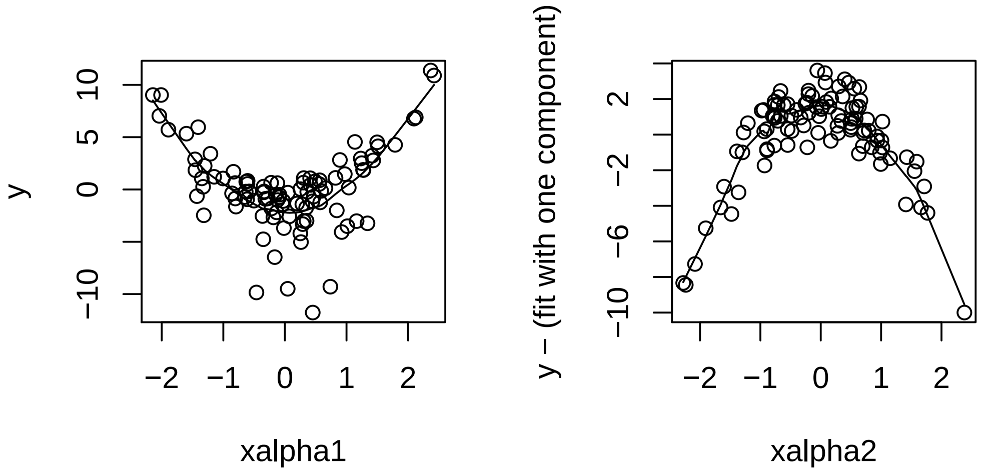

Figure 1: The first 3 components **(c2e1.R)**

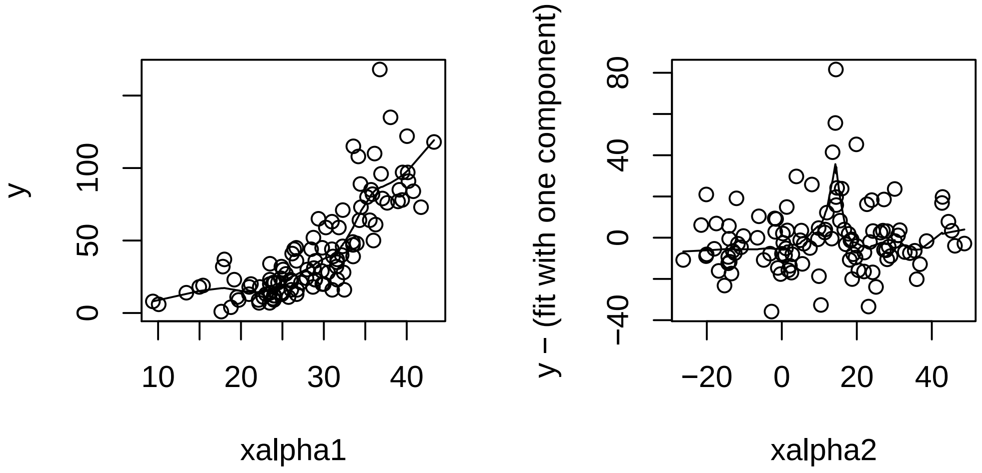

Figure 2: The first 2 components **(c2e2.R)**

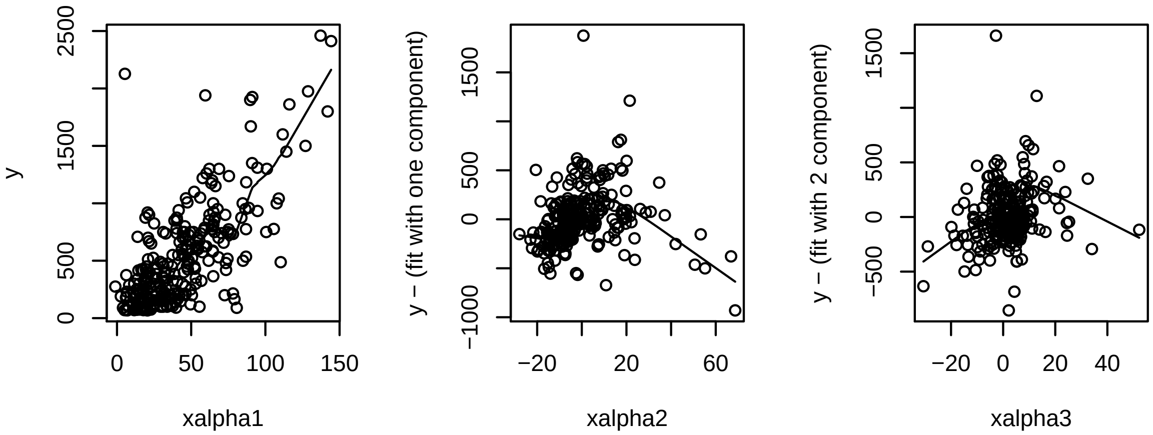

Figure 3: The first 3 components **(c2e3.R)**

where $z_1' = \hat{\alpha}_1^\top X', ..., z_k' = \hat{\alpha}_k^\top X'$ and

$$\hat{g}_1(z_1') = \frac{\sum_{i=1}^n K_h(\hat{\alpha}_1^\top X_i - z')Y_i}{\sum_{i=1}^n K_h(\hat{\alpha}_1^\top X_i - z')}$$

$$\hat{g}_1(z_2') = \frac{\sum_{i=1}^n K_h(\hat{\alpha}_1^\top X_i - z')r_{1i}}{\sum_{i=1}^n K_h(\hat{\alpha}_1^\top X_i - z')}$$

$$...$$

$$\hat{g}_k(z_k') = \frac{\sum_{i=1}^n K_h(\hat{\alpha}_1^\top X_i - z')r_{k-1,i}}{\sum_{i=1}^n K_h(\hat{\alpha}_1^\top X_i - z')}$$

where r_{1i} is the fitted residuals of PPR with one component, ..., and $r_{k-1,i}$ is the fitted residuals of PPR with k-1 component.

Example 2.4 (ozone data) *We fit PPR model with component 2 to the data. The estimated model is shown in Figure 2. Suppose we have a new set of covariate*

$$[184.8, 77.8, 9.9]$$

If we use two component PPR, the predicted value is 37.72731. **(c2e4.R)**

Example 2.5 *Consider* **data** *about the baseball's players and their performance. Suppose we have a new player with performance indicators*

$$[403, 107, 11, 54, 51, 41, 7, 2657, 722, 69, 361, 330, 260, 290, 118, 8]$$

If we use 3 component PPR, the predicted value is 523.8938. **(c2e5.R)**

3 model selection based on cross-validation

Up to now, we have introduced a number of models including Linear regression model, partially linear regression model, varying coefficient regression model; single-index model and projection pursuit regression. (More will be introduced later). A simple question is: which model shall we use? In other word, we need a criterion for this purpose

3.1 RSS cannot be used as the criterion

The residual sum of squares (RSS or SSR) is the sum of squares of residuals. RSS is determined by the model as well as its complexity. Here is an example.

Example 3.1 *Suppose the try model is*

$$Y = 0.5 - \mathbf{x}_1 + \mathbf{x}_2 + 0\mathbf{x}_3 + 0.5 * \varepsilon$$

where $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ *and* ε *are IID Normal N(0,1). The true model is*

$$Y = 0.5 - \mathbf{x}_1 + \mathbf{x}_2 + 0.5 * \varepsilon$$

50 samples are drawn from the model.

Suppose we consider models

$$\begin{array}{rl} I & Y = \beta_0 + \beta_1\mathbf{x}_1 + \varepsilon \\ II & Y = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \varepsilon \\ III & Y = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \varepsilon \\ IV & Y = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_1^2 + \varepsilon \\ V & Y = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_1^2 + \beta_5\mathbf{x}_1\mathbf{x}_2 + \varepsilon \\ \ldots & \end{array}$$

Let $RSS_I, RSS_{II}, \ldots$ *be the residual sum of squares. Then, we always have*

$$RSS_I > RSS_{II} > RSS_{III} > RSS_{IV} > RSS_V > \ldots$$

The more complicated the model is, the smaller RSS the model has. **(c2e21.R)**

One way to choose the correct model is to penalize the complexity. In the above example, the complexity is the number of parameters. One of the methods is the AIC (Akaike's Information Criterion) method.

$$AIC = \log(RSS/n) + 2\frac{p}{n}$$

where p is the number of covariates used in the model and n is the number of observations. The model with smallest AIC is the preferable model

Example 3.2 *(continued)*

Let $AIC_I, AIC_{II}, ...$ be the AICs for model I, II, ... respectively. Then, most likely

$$AIC_I > AIC_{II} < AIC_{III} < AIC_{IV} < AIC_V < ...$$

(c2e21.R)

3.2 Cross-validation as the criterion

The best statistical model is the model that has the best prediction amongst all possible models. After we fit a model by a (training) data set, **if we have a validation set**, then prediction can be calculated based on the validation set.

In practice, we usually dont have a validation set. In that case, we can partition the data into validation set and training set.

The (leave-one-out) Cross-validation each time partition the data (n observations) into training set with n-1 observations and validation set with 1 observation. Therefore there are n possible cases of partitioning. The overall prediction error is defined as the CV value. The model with smallest CV is the preferable model

For linear regression model,

$$Y = \beta_0 + \beta_1 \mathbf{x}_1 + ... + \beta_k \mathbf{x}_k + \varepsilon$$

Suppose we have n observations $(Y_i, \mathbf{x}_{i1}, ..., \mathbf{x}_{ik}), i = 1, 2, ..., n$. For each i, we estimate the model based on sample $1, ..., i-1, i+1, ..., n$. suppose the estimated model is

$$\hat{Y} = \hat{\beta}_{0,i} + \hat{\beta}_{1,i} \mathbf{x}_1 + ... + \hat{\beta}_{k,i} \mathbf{x}_k$$

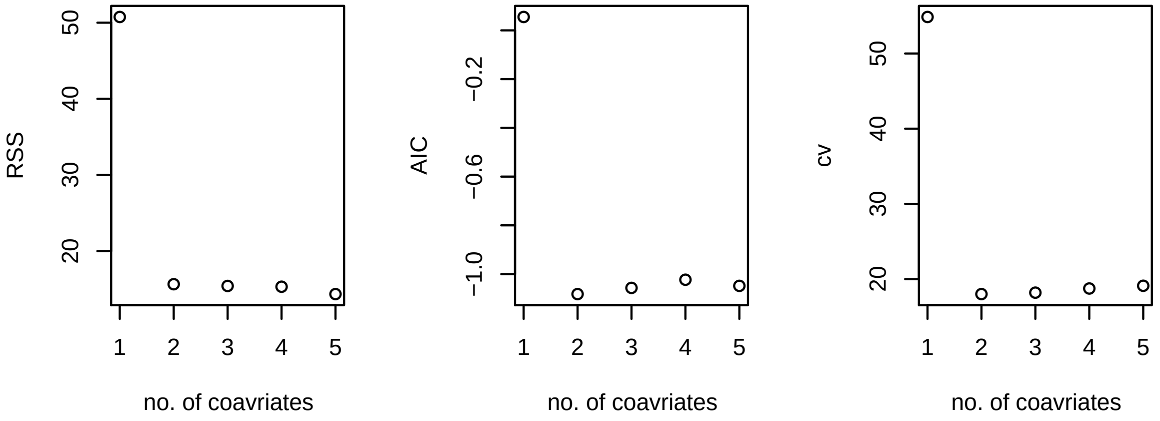

Figure 4: plot of RSS, AIC and CV against the number of covariates in a linear regression model.

The prediction of Y_i is $\hat{Y}_i$

$$\hat{Y}_i = \hat{\beta}_{0,i} + \hat{\beta}_{1,i}\mathbf{x}_{i1} + ... + \hat{\beta}_{k,i}\mathbf{x}_{ik}$$

Then the CV value is defined as

$$CV = n^{-1}\sum_{i=1}^{n}\{Y_i - \hat{Y}_i\}^2$$

Example 3.3 *(example 2.2 continued) Variable selection in Linear regression model.*

Let $CV_I, CV_{II}, ...$ be the CVs for model I, II, ... respectively. Then, most likely

$$CV_I > CV_{II} < CV_{III} < CV_{IV} < CV_V < ...$$

(c2e21.R)

References

Breiman L., Friedman J. H., Olshen R. A., and Stone, C. J. (1984) *Classification and Regression Trees.* Wadsworth.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge.