Chapter 3. Spline smoothing and semi-parametric Models (II) Part 2

March 14, 2007

1 The generalized additive model

Suppose we have response variable Y and a number of predictors (independent variables) $\mathbf{x}_1, ..., \mathbf{x}_p$. We are interested in

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

The goal is to estimate $m(.)$. Because of the “curse of dimensionality”, the estimation is very unreliable if p is large (> 2).

One way to approximate $m(.)$ is by the summation of functions of each variable

$$m(x_1, ..., x_p) \approx g_1(x_1) + ... + g_p(x_p)$$

If the equality hold, we call the model additive model,

$$Y = g_1(\mathbf{x}_1) + ... + g_p(\mathbf{x}_p) + \varepsilon$$

where $E(\varepsilon|\mathbf{x}_1 = x_1, ..., \mathbf{x}_p = x_p) = 0$.

Identification of the model Up to a constant difference, each component is identifiable. That is if there is another functions

$$Y = f_1(\mathbf{x}_1) + ... + f_p(\mathbf{x}_p) + \varepsilon$$

Then there is a constant c_k such that

$$f_k(\mathbf{x}_k) = c_k + g_k(\mathbf{x}_k), \quad k = 1, 2, ..., p$$

We can rewrite the model as

$$Y = \beta_0 + g_1(\mathbf{x}_1) + ... + g_p(\mathbf{x}_p) + \varepsilon$$

where $E\{g_1(\mathbf{x}_1)\} = 0$.

Most of the time, we have some knowledge about the relation between Y and some independent variables. For example, we know the relation between Y and $\mathbf{x}_1, ..., \mathbf{x}_q$ are linear. Thus we have the following **Generalized Additive model**

$$Y = \beta_0 + \beta_1\mathbf{x}_1 + ... + \beta_q\mathbf{x}_q + g_{q+1}(\mathbf{x}_{q+1}) + ... + g_p(\mathbf{x}_p) + \varepsilon$$

For identification purpose, we further constrain that $E\{g_k(\mathbf{x}_k)\} = 0, k = q+1, ..., p$. In the model $\mathbf{x}_1, ..., \mathbf{x}_q$ are the linear part, and $g_{q+1}(\mathbf{x}_{q+1}), ..., g_p(\mathbf{x}_p)$ are the nonlinear components.

Note that the partially linear regression model is a special case of GAM.

1.1 Estimation of the GAM model

One way to estimate the GAM model is assuming the nonlinear components have the spline form, i.e.

$$g_k(x) = \sum_{j=1}^{J_k+4} \theta_{k,j} B_{k,j}(x)$$

where $B_{k,j}, j = 1, ..., J_k + 4$ is the spline basis for function g_k. Thus the model can be written as

$$Y = \beta_0 + \beta_1\mathbf{x}_1 + ... + \beta_q\mathbf{x}_q + \sum_{k=q+1}^{p} \sum_{j=1}^{J_k+4} \theta_{k,j} B_{k,j}(\mathbf{x}_k) + \varepsilon$$

Suppose that $(\mathbf{x}_{i1}, ..., \mathbf{x}_{ip}, Y_i), i = 1, ..., n$ are samples from the model. (How to estimate the model?)

R package: gam (please install it in your computer)

Example 1.1 (simulation) *100 samples are drawn from the following model*

$$Y = 2.5 + 0.5\mathbf{x}_1 - 0.4\mathbf{x}_2 + \sin(2\pi\mathbf{x}_3) + \exp(-20(\mathbf{x}_4 - 0.5)^2) + 0.2 * \varepsilon$$

where $\mathbf{x}_1, \mathbf{x}_2$ and ε are IID N(0, 1) and $\mathbf{x}_3, \mathbf{x}_4$ IID uniformly on [0, 1]

The estimated coefficients are

$$\hat{\beta}_0 = 3.8281922, \hat{\beta}_1 = 0.4876495, \hat{\beta}_2 = -0.3898373$$

and the estimated nonlinear components are shown in figure 1

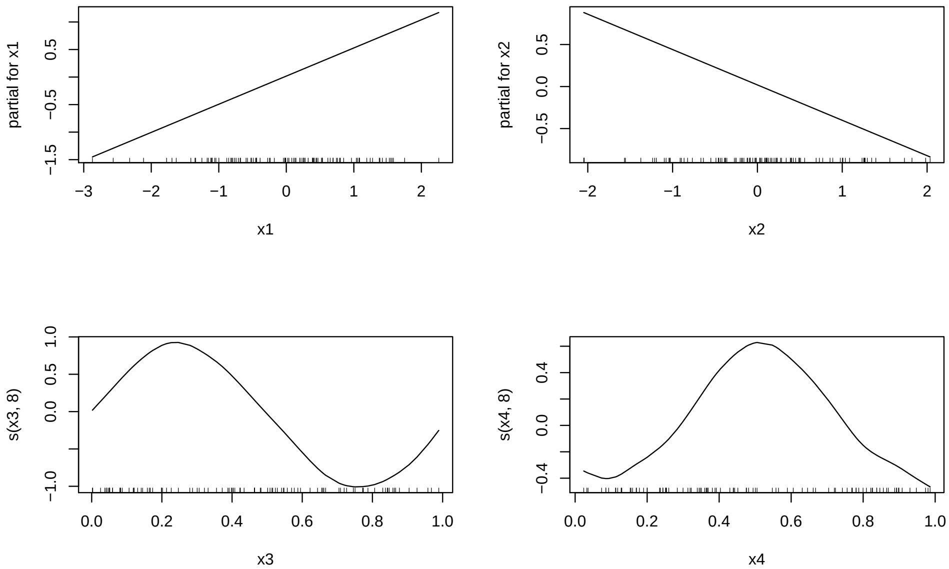

Figure 1: The estimated GAM model **(code)**

Example 1.2 (ozone) **(data)** *The level of ozone might be affected by radiation, temperature and wind. consider model*

$$ozone^{1/3} = g_1(rad.) + g_2(temp.) + g_3(wind) + \varepsilon$$

there are 111 observations

If we need to select one model amongst the following 5 models

$$
\begin{aligned}
(0) \qquad & ozone^{1/3} = g_1(rad.) + g_2(temp.) + g_3(wind) + \varepsilon \\
(I) \qquad & ozone^{1/3} = \beta_0 + \beta_1 * rad + g_2(temp.) + g_3(wind) + \varepsilon \\
(II) \qquad & ozone^{1/3} = \beta_0 + g_1(rad) + \beta_2 * temp + g_3(wind) + \varepsilon \\
(III) \qquad & ozone^{1/3} = \beta_0 + g_1(rad) + g_2(temp) + \beta_3 * wind + \varepsilon \\
(I) \qquad & ozone^{1/3} = \beta_0 + \beta_1 * rad + \beta_2 * temp + \beta_3 * wind + \varepsilon
\end{aligned}
$$

Their CV values are 0.2380925, 0.2390885, 0.2496370, 0.2531100, 0.2730964 respectively. thus, model (0) is selected

For a new set of predictors $rad = 100, temp = 80, wind = 10$, *predict its ozone level.*

The predicted ozone is $exp(2.95195) = 19.14325$

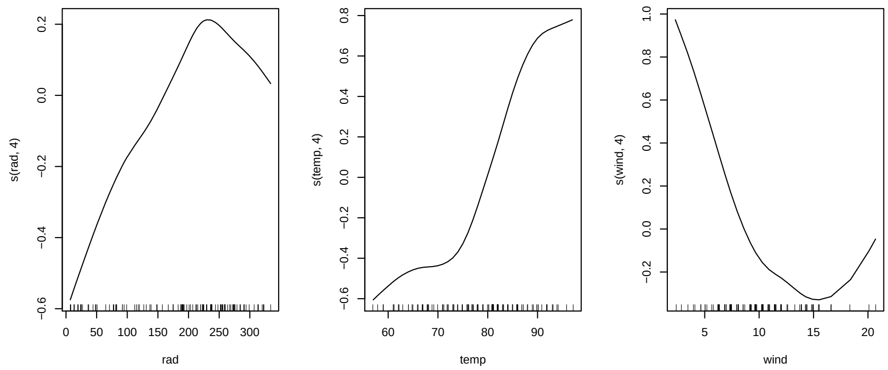

Figure 2: The estimated Additive model **(code)**

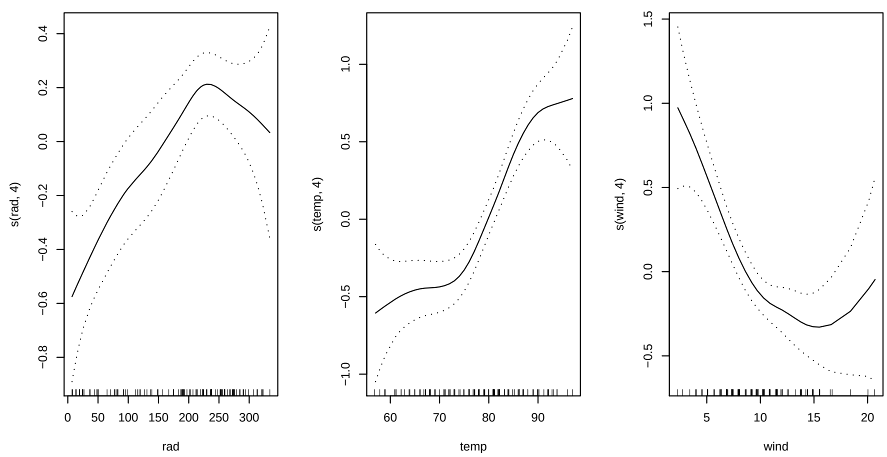

Figure 3: The estimated GAM model and its 95% confidence bands. **(code)**

References

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models* London: Chapman and Hall.