ANALYSIS OF LONGITUDINAL DATA USING CUBIC SMOOTHING SPLINES

Tapio Nummi

Department of Mathematics, Statistics and Philosophy

33014 University of Tampere

Finland

1. Spline smoothing

Suppose that our aim is to model

$$y_i = d(x_i) + \epsilon_i, \ i = 1, \ldots, n,$$

where d is a smooth function and ϵ_i are iid with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma_\epsilon^2$.

The linear spline estimator is

$$d(x_i) = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k (x - \kappa_k)_+,$$

$$(x - \kappa_k)_+ = \begin{cases} 0, & x \le \kappa_k \\ x - \kappa_k, & x > \kappa_k \end{cases}$$

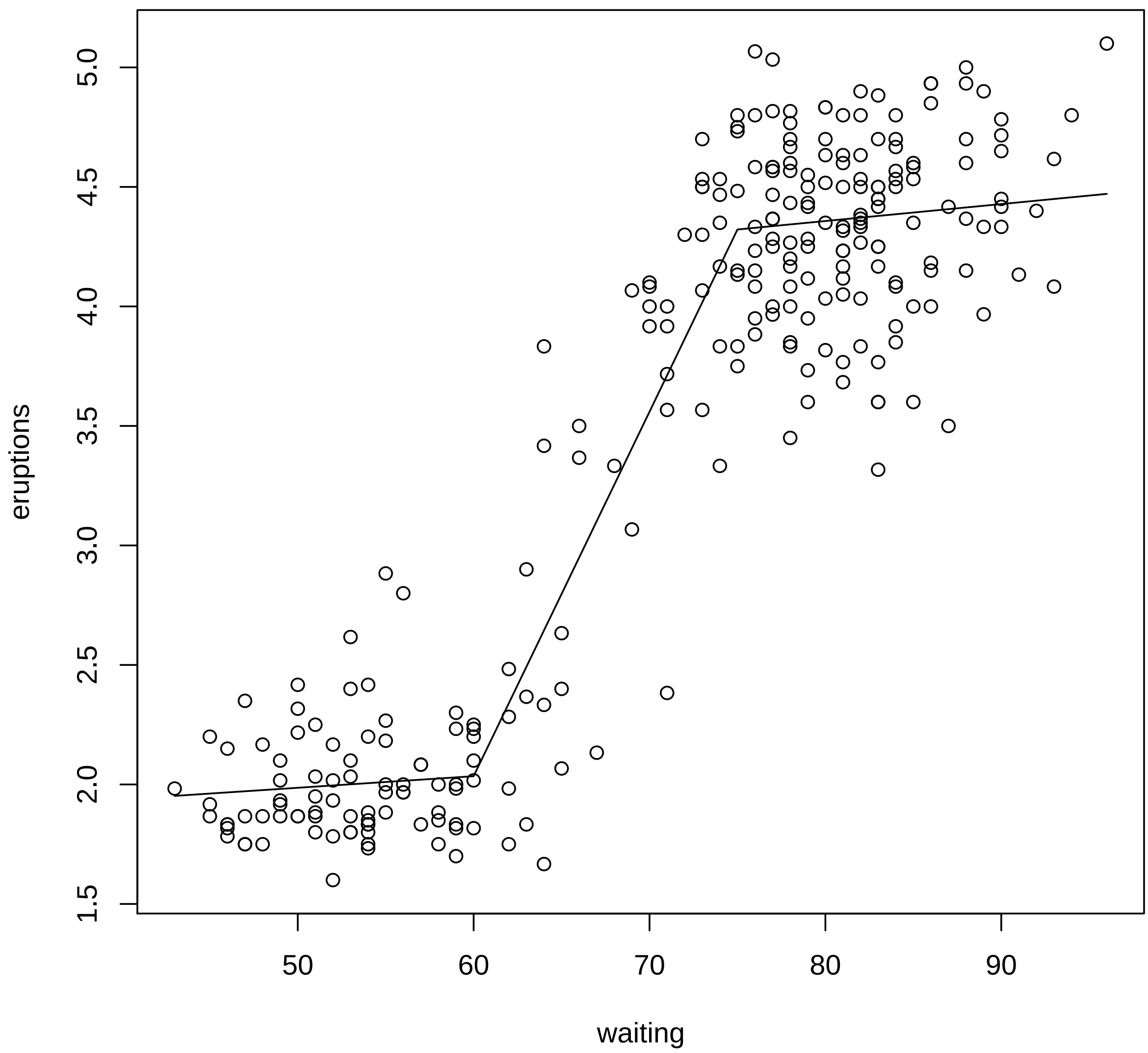
and $\kappa_1, \ldots, \kappa_K$ are $knots$.

The curve d is now modeled by piecewise line segments tied together at knots $\kappa_1, \ldots, \kappa_K$.

Example

```
> library(MASS)
> data(faithful)
> names(faithful)
[1] "eruptions" "waiting"
> plot(faithful)
> faithful<-faithful[order(faithful$waiting),]
> attach(faithful)
> knots<-c(0,60,75) % knots 60, 75
> rhs<-function(x,c) ifelse (x>c,x-c,0)
> dm<-outer(waiting, knots, rhs)
> dm
       [,1] [,2] [,3]
  [1,]   43    0    0
  [2,]   45    0    0
  ...
 [83,]   60    0    0
 [84,]   62    2    0
 ...
[134,]   75   15    0
[135,]   76   16    1
...
```

```
> g<-lm(eruptions~dm)
> plot(eruptions~waiting)
> lines(waiting, predict(g))
>
```



We can generalize the above equation to a piecewise polynomial of degree p, but the most common choices in practice are quadratic ($p = 2$) and cubic ($p = 3$) splines.

For cubic splines we have

$$d(x_i; \boldsymbol{\beta}; \boldsymbol{u}) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \sum_{k=1}^{K} u_k (x - \kappa_k)_+^3,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$, $\boldsymbol{u} = (u_1, \ldots, u_k)'$ and 1, x, x^2, x^3, $(x - \kappa_1)_+^3, \ldots, (x - \kappa_K)_+^3$ are called basis functions. Other possible choices of **basis functions** include B-splines, wavelet, Fourier Series and polynomial bases etc.

A **natural cubic spline** is obtained by assuming that the function is linear beyond the boundary knots.

The number (K) and location of knots $\kappa_1, \ldots, \kappa_K$ must be specified in advance.

Coefficients $\boldsymbol{\beta}$ and $\boldsymbol{u}$ can be estimated using standard least squares procedures.

However, in some cases the estimated curve tends to be a very rough estimate.

Our approach is to apply smoothing splines, where the smoothing is controlled by a smoothing parameter α.

Smoothing splines have a knot at each unique value of x and the fitting is carried out by least squares with a roughness penalty term.

2. Penalized smoothing

If $x_1, \ldots, x_n$ are points in $[a, b]$ satisfying $a < x_1, \ldots, x_n < b$ the penalized sum of squares (PSS) is given as

$$\sum_{i=1}^{n} \{y_i - d(x_i)\}^2 + \alpha \int_a^b \{d''(x)\}^2 dx,$$

where

$$\alpha \int_a^b \{d''(x)\}^2 dx$$

is the roughness penalty (RP) term with $\alpha > 0$.

Note that here α represents the rate of exchange between residual error and local variation.

If α is very large the main component of PSS will be RP and the estimated curve will be very smooth.

If α is relatively small the estimated curve will track the data points very closely.

If we define a non-negative definite matrix

$$\boldsymbol{K} = \boldsymbol{\nabla}\boldsymbol{\Delta}^{-1}\boldsymbol{\nabla}',$$

where non-zero elements of $n \times (n-2)$ matrix $\boldsymbol{\nabla}$ and $(n-2) \times (n-2)$ matrix $\boldsymbol{\Delta}$ are defined as

$$\nabla_{ii} = \frac{1}{h_i},\ \nabla_{i+1,i} = -\left(\frac{1}{h_i} + \frac{1}{h_{i+1}}\right),\ \nabla_{i+2,i} = \frac{1}{h_{i+1}}$$

and

$$G_{i,i+1} = G_{i+1,i} = \frac{h_{i+1}}{6},\ G_{ii} = \frac{h_i + h_{i+1}}{3},$$

where $h_j = x_{j+1} - x_j,\ j = 1, 2, \ldots, n-1$.

Now PSS becomes as

$$PSS(\boldsymbol{K}) = (\boldsymbol{y} - \boldsymbol{d})'(\boldsymbol{y} - \boldsymbol{d}) + \alpha \boldsymbol{d}'\boldsymbol{K}\boldsymbol{d}$$

and its minimum is obtained at

$$\widehat{\boldsymbol{d}} = (\boldsymbol{I} + \alpha \boldsymbol{K})^{-1}\boldsymbol{y}.$$

It can be shown (e.g. Green and Silverman, 1994) that $\widehat{\boldsymbol{d}}$ is a natural cubic smoothing with knots at the points $x_1, \ldots, x_n$.

Note that the special form $\widehat{\boldsymbol{d}}$ follows from the chosen RP term

$$\alpha \int_a^b \{d''(x)\}^2 dx.$$

If we, for example, would use a discrete approximation

$$\mu_{i+1} - 2\mu_i + \mu_{i-1}$$

of the second derivative the PSS would be (Demidenko, 2004)

$$PSS(\boldsymbol{QQ}') = (\boldsymbol{y} - \boldsymbol{d})'(\boldsymbol{y} - \boldsymbol{d}) + \alpha \boldsymbol{d}'\boldsymbol{QQ}'\boldsymbol{d},$$

where ($n = 6$)

$$\boldsymbol{Q} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ 0 & 1 & -2 & 1 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Then the minimizer is

$$\tilde{\boldsymbol{d}} = (\boldsymbol{I} + \alpha \boldsymbol{Q}\boldsymbol{Q}')^{-1}\boldsymbol{y}.$$

Note that for fixed α the spline fit

$$\widehat{\boldsymbol{d}} = (\boldsymbol{I} + \alpha \boldsymbol{K})^{-1}\boldsymbol{y} = \boldsymbol{S}_\alpha \boldsymbol{y}$$

is linear in $\boldsymbol{y}$ and the matrix $\boldsymbol{S}_\alpha$ is known as the *smoother matrix*.

The smoother matrix $\boldsymbol{S}_\alpha$ has many interesting properties discussed e.g. in Hastie, Tibshirami and Friedman (2001), but here I briefly mention only the following :

1. Choosing the smoothing parameter:

$$CV(\alpha) = \sum_{i=1}^{n} \left(\frac{y_i - \widehat{d}_\alpha(x_i)}{1 - S_\alpha(i,i)}\right)^2,$$

   where $S_\alpha(i,i)$ are diagonal elements of $\boldsymbol{S}_\alpha$.

2. Estimation of the effective degrees of freedom

$$df_\alpha = tr(\boldsymbol{S}_\alpha).$$

   This can be compared to matrix

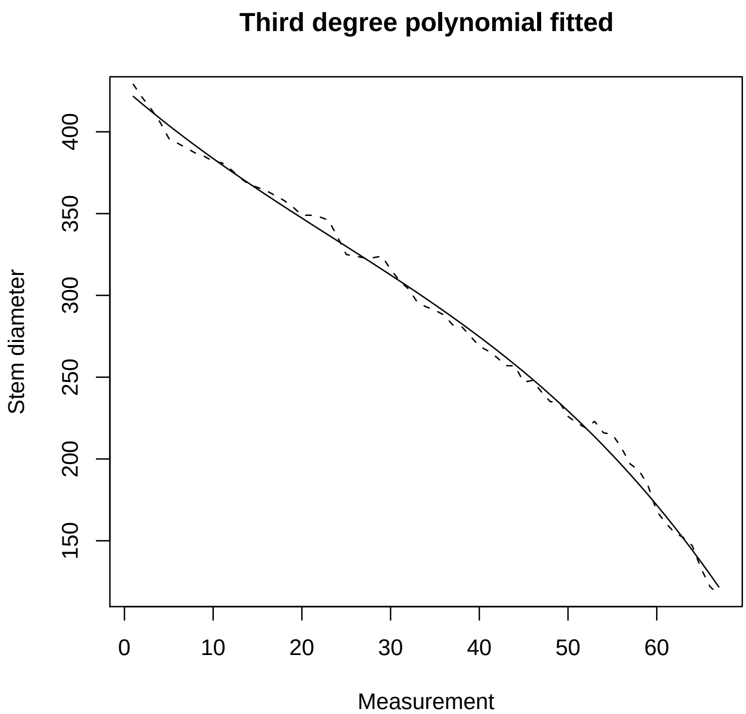
$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$$

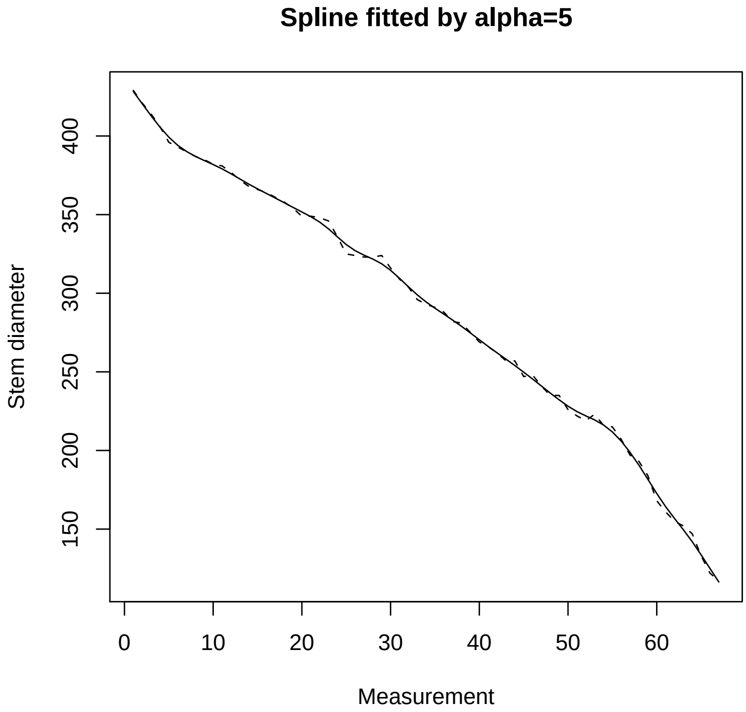
in regression analysis (or in regression splines) in a sense that

$$tr(\boldsymbol{H})$$

gives the number of estimated parameters (or the number of basis functions utilized).

Example: Stem curve model - modelling the degrease of stem diameter as a function stem height.





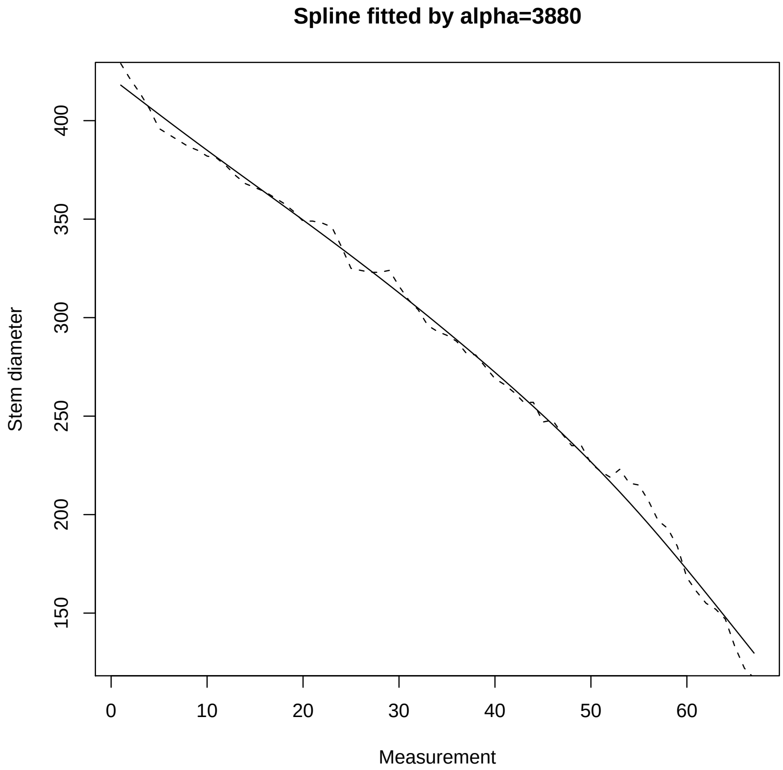
The effective number of degrees of freedom

$df_\alpha = tr(S_{\alpha=5}) = 16.79628.$

Note that if

$$\alpha \to 0,\ \ df_\alpha \to n$$

$$\alpha \to \infty,\ \ df_\alpha \to 2$$



Since $df_\alpha = tr(S_\alpha)$ is monotone in α, we can invert the relationship and specify α by fixing df. For $df = 4$ this gives $\alpha = 3880$.

This yields to model selection with different values for df, where more traditional criteria developed for regression models maybe used.

3. Connection to mixed models

If we let

$$\boldsymbol{X} = [\mathbf{1}, \boldsymbol{x}],$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)'$ and by the special form of $\boldsymbol{\nabla}$ we note that

$$\boldsymbol{X}'\boldsymbol{\nabla} = \mathbf{0}$$

and

$$(\boldsymbol{I}+\alpha\boldsymbol{K})^{-1} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'+\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z}+\alpha\Delta^{-1})\boldsymbol{Z}',$$

where $\boldsymbol{Z} = \boldsymbol{\nabla}(\boldsymbol{\nabla}'\boldsymbol{\nabla})^{-1}$.

Then the solution of $PSS(\boldsymbol{K})$ can be written as

$$\widehat{\boldsymbol{d}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\widehat{\boldsymbol{u}},$$

where

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y}$$

and

$$\widehat{\boldsymbol{u}} = (\boldsymbol{Z}'\boldsymbol{Z} + \alpha\boldsymbol{\Delta}^{-1})^{-1}\boldsymbol{Z}'\boldsymbol{y}.$$

These estimates can be seen as (BLUP) solutions of the mixed model

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon},$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are defined before and

$$\boldsymbol{u} \sim N(\mathbf{0}, \sigma_u^2\boldsymbol{\Delta}) \text{ and } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$$

with smoothing parameter as a variance ratio $\alpha = \frac{\sigma^2}{\sigma_u^2}$.

Note that we may always rewrite

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}_*\boldsymbol{u}_* + \boldsymbol{\epsilon},$$

where $\boldsymbol{Z}_* = \boldsymbol{Z}\Delta^{1/2}$ and $\boldsymbol{u}_* = \Delta^{-1/2}\boldsymbol{u}$ with

$$\boldsymbol{u}_* \sim N(\boldsymbol{0}, \sigma_u^2\boldsymbol{I}) \text{ and } \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

We can now use standard statistical software for parameter estimation (e.g. LME in R or Proc Mixed in SAS).

4. Growth Curves

- The growth curve model (GCM) of Potthoff & Roy (1964)

$$\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{B}\boldsymbol{A}' + \boldsymbol{E},$$

where $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n)$ is a matrix of obs.,
$\boldsymbol{T}$ and $\boldsymbol{A}$ are design matrices (within and between individual),
$\boldsymbol{B}$ is a matrix of unknown parameters, and
$\boldsymbol{E}$ is a matrix of random errors.

- The columns of $\boldsymbol{E}$ are independently distributed as

$$\boldsymbol{e}_i \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

- Here I assume that

$$\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{R},$$

where $\boldsymbol{R}$ takes certain parsimonious covariance structure with covariance parameters $\boldsymbol{\theta}$.

- Now we may write

$$\boldsymbol{Y} = \boldsymbol{G}\boldsymbol{A}' + \boldsymbol{E},$$

  where $\boldsymbol{G} = (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_m)$ is the matrix of mean curves.

- The GCM is a linear approximation

$$\begin{aligned} \boldsymbol{G} &= (\boldsymbol{g}_1, \ldots, \boldsymbol{g}_m) \\ &= (\boldsymbol{T}\boldsymbol{\beta}_1, \ldots, \boldsymbol{T}\boldsymbol{\beta}_m) \\ &= \boldsymbol{T}\boldsymbol{B}. \end{aligned}$$

- The aim here is to develop the methods needed when $\boldsymbol{G}$ is approximated by more flexible cubic smoothing splines.

- Penalized log-likelihood function

$$2l = -\frac{1}{\sigma^2}\text{tr}[(\boldsymbol{Y}' - \boldsymbol{A}\boldsymbol{G}')\boldsymbol{R}^{-1}(\boldsymbol{Y}' - \boldsymbol{A}\boldsymbol{G}')' +$$

$$\alpha(\boldsymbol{A}\boldsymbol{G}')\boldsymbol{K}(\boldsymbol{A}\boldsymbol{G}')'] - n\log|\sigma^2\boldsymbol{R}| - c.$$

- For given α, σ^2 and $\boldsymbol{R}$, the maximum is obtained at

$$\tilde{\boldsymbol{G}} = (\boldsymbol{R}^{-1} + \alpha\boldsymbol{K})^{-1}\boldsymbol{R}^{-1}\boldsymbol{Y}\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}.$$

- If $\boldsymbol{R}$ satisfies

$$\boldsymbol{R}\boldsymbol{K} = \boldsymbol{K},$$

this simplifies to

$$\widehat{\boldsymbol{G}} = (\boldsymbol{I} + \alpha\boldsymbol{K})^{-1}\boldsymbol{Y}\boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}.$$

- It is easily seen that

  $\boldsymbol{R} = \boldsymbol{I}$ (Independent),

  $\boldsymbol{R} = \boldsymbol{I} + \sigma_d^2 \mathbf{1}\mathbf{1}'$ (Uniform),

  $\boldsymbol{R} = \boldsymbol{I} + \sigma_{d'}^2 \boldsymbol{X}\boldsymbol{X}'$ (Linear1),

  $\boldsymbol{R} = \boldsymbol{I} + \boldsymbol{X}\boldsymbol{D}\boldsymbol{X}'$ (Linear2)

  satisfies the condition $\boldsymbol{R}\boldsymbol{K} = \boldsymbol{K}$.

- This result can be compared to estimation in linear models, when BLUE coinsides with OLSE.