Advanced Data Analysis

Penalised Splines and Mixed Models

A very important fact in semiparametric regression is:

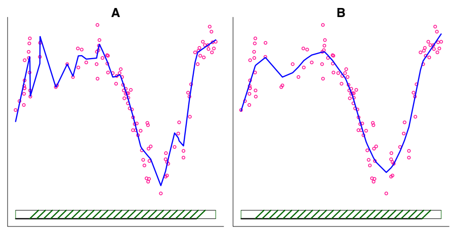
penalised splines

are a special case of

mixed models!

Tricking Mixed Models to Smooth Scatterplots

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



A: u_k's fixed B: u_k i.i.d. $N(0, \sigma_u^2)$

Penalised Splines as Mixed Models

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{fixed effects}} + \underbrace{\sum_{k=1}^{K} u_k (x_i - \kappa_k)_+}_{\text{random effects}} + \varepsilon_i$$

Matrix Notation

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

$$\boldsymbol{X} = [1\ x_i] \qquad \boldsymbol{Z} = [(x_i - \kappa_1)_+ \cdots (x_i - \kappa_K)_+]$$

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \sigma_u^2 \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_\varepsilon^2 \boldsymbol{I} \end{bmatrix} \right)$$

Expressions for Fitted Coefficients

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T \widehat{\boldsymbol{V}}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \widehat{\boldsymbol{V}}^{-1} \boldsymbol{y} \\ \widehat{\boldsymbol{u}} &= \widehat{\sigma}_u^2 \boldsymbol{Z}^T \widehat{\boldsymbol{V}}^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \end{aligned}$$

where

$$\widehat{\boldsymbol{V}} = \widehat{\text{Cov}(\boldsymbol{y})} = \widehat{\sigma}_u^2 \boldsymbol{Z}\boldsymbol{Z}^T + \widehat{\sigma}_\varepsilon^2 \boldsymbol{I}$$

More Compact Expression

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{bmatrix} = (\boldsymbol{C}^T \boldsymbol{C} + \lambda \boldsymbol{D})^{-1} \boldsymbol{C}^T \boldsymbol{y}$$

where

$$\boldsymbol{C} = [\boldsymbol{X}\ \boldsymbol{Z}], \quad \boldsymbol{D} = \text{diag}(0, 0, 1, \ldots, 1) \quad \text{and} \quad \lambda = \widehat{\sigma}_\varepsilon^2 / \widehat{\sigma}_u^2.$$

$$\Longrightarrow\ \widehat{\boldsymbol{y}} = \boldsymbol{C}(\boldsymbol{C}^T \boldsymbol{C} + \lambda \boldsymbol{D})^{-1} \boldsymbol{C}^T \boldsymbol{y}$$

This matches our original expressions for penalised splines.

See also the last section in the handout titled ‘Mixed Models’.

Binary Response Extension

$$\text{logit}\{P(y_i = 1)\} = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k z_k(x_i),$$

$$u \sim N(0, \sigma_u^2)$$

Generalised Additive Models Extension

$$y = X\beta + Zu + \varepsilon, \qquad X = [1 \; x_{1i} \; x_{2i} \; x_{3i}]$$

$$Z = [z_1(x_{1i}) \cdots z_{K_1}(x_{1i}) | z_1(x_{2i}) \cdots z_{K_2}(x_{2i}) | z_1(x_{3i}) \cdots z_{K_3}(x_{3i})]$$

$$\begin{bmatrix} u \\ \varepsilon \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u1}^2 I & 0 & 0 & 0 \\ 0 & \sigma_{u2}^2 I & 0 & 0 \\ 0 & 0 & \sigma_{u3}^2 I & 0 \\ 0 & 0 & 0 & \sigma_{\varepsilon}^2 I \end{bmatrix} \right)$$

R Demonstration

In this R demonstration we illustrate

generalised additive model analysis via
linear mixed model software

for the Californian ozone data.

The code for this demonstration is in the script ozoneCalifLMM.Rs

Advantages of Penalised Splines as Mixed Models

- The nonparametric regression parts can be done in the same way as the parametric parts.
- That is, all just use the notions of (restricted) maximum likelihood and best prediction.
- This means that standard software such as `lme()` in `R` can be used for semiparametric regression analyses.
- In the Bayesian case we can use MCMC and BRugs.

References for Further Reading

Wand, M. P. (2003). Smoothing and mixed models. *Computational Statistics*, **18**, 223–249.

Ngo, L. and Wand, M.P. (2004). Smoothing with mixed model software. *Journal of Statistical Software*, Volume 9, Article 1, 1–54.

Ruppert, D., Wand, M. P. & Carroll, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

Ruppert, D., Wand, M. P. & Carroll, R.J. (2009). Semiparametric regression during 2003-2007. *Electronic Journal of Statistics.* **3**, 1193–1256.

Marley, J.K. and Wand, M.P. (2010). Non-standard semiparametric regression via BRugs. *Journal of Statistics Software*, Volume 37, Issue 5, 1–30.

Note: All but the 2nd of these references is available (for free) on Wand papers web-site (www.uow.edu.au/˜mwand/papers.html)