

LTCC Advanced Course: Introduction to Semiparametric Modelling Lecture 5

Clifford Lam

Department of Statistics
London School of Economics and Political Science

Degree of freedom

- Linear terms have 1 degree of freedom for each predictor variable, whereas nonlinear terms have some number greater than 1, depending on the curviness of the function.
- Since for additive model $y_i = \beta_0 + \sum_{j=1}^d f_j(x_{ji}) + \epsilon_i$ we have

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}} = \mathbf{C}(\mathbf{C}^T\mathbf{C} + \Lambda)^{-1}\mathbf{C}^T\mathbf{y},$$

hence the smoother matrix is $\mathbf{L} = \mathbf{C}(\mathbf{C}^T\mathbf{C} + \Lambda)^{-1}\mathbf{C}^T$, where

$$\mathbf{C} = [\mathbf{X} \ \mathbf{Z}], \quad \text{and} \quad \Lambda = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma^2 \text{cov}(\mathbf{u})^{-1} \end{pmatrix}.$$

Here $\text{cov}(\mathbf{u}) = \text{diag}(\sigma_{u_1}^2, \dots, \sigma_{u_d}^2)$.

- The total degree of freedom is

$$df_{\text{fit}} = \text{tr}\{(\mathbf{C}^T\mathbf{C} + \Lambda)^{-1}\mathbf{C}^T\mathbf{C}\}.$$

- To define degree of freedom for each component, we can let P denotes the number of columns of \mathbf{C} , and let

$$\{I_0, I_1, \dots, I_d\}$$

be a partition of the column indices $\{1, \dots, P\}$ such that I_0 corresponds to the intercept β_0 and I_j corresponds to $f_j(\cdot)$ for each $1 \leq j \leq d$.

Degree of freedom

- In the temperature data example, $K_s = K_t = 20$, hence we have $P = 43$ and we would have

$$I_0 = \{1\}, \quad I_1 = \{2, 4, 5, \dots, 23\}, \quad I_2 = \{3, 24, 25, \dots, 43\}.$$

- Let \mathbf{A}_I denotes the submatrix of \mathbf{A} consisting of columns with indices in I . Then

$$\{\mathbf{C}_{I_0}, \mathbf{C}_{I_1}, \dots, \mathbf{C}_{I_d}\}$$

represents a partition of the columns of \mathbf{C} corresponding to the terms of the spline formulation of the additive model $y_i = \beta_0 + \sum_{j=1}^d f_j(x_{ji}) + \epsilon_i$.

- We define \mathbf{E}_i to be the $P \times P$ diagonal matrix with ones in the diagonal positions with indices in I_i and zero elsewhere. Then the fitted values for the j th term are

$$\begin{pmatrix} \hat{f}_j(x_{j1}) \\ \vdots \\ \hat{f}_j(x_{jn}) \end{pmatrix} = \{\mathbf{C}\mathbf{E}_j(\mathbf{C}^T\mathbf{C} + \Lambda)^{-1}\mathbf{C}^T\}\mathbf{y}.$$

- The corresponding degree of freedom may be computed as

$$df_j = \text{tr}\{\mathbf{C}\mathbf{E}_j(\mathbf{C}^T\mathbf{C} + \Lambda)^{-1}\mathbf{C}^T\} = \text{tr}\{\mathbf{E}_j(\mathbf{C}^T\mathbf{C} + \Lambda)^{-1}\mathbf{C}^T\mathbf{C}\},$$

which is the sum over the indices of I_j of the diagonal elements of the matrix $(\mathbf{C}^T\mathbf{C} + \Lambda)^{-1}\mathbf{C}^T\mathbf{C}$. Hence

$$df_{\text{fit}} = df_0 + \dots + df_d.$$

Smoothing parameter selection

- The REML approach to finding variance components in a mixed model framework is attractive, since it is easy and automatic. Then $\hat{\lambda}_j = \hat{\sigma}^2 / \hat{\sigma}_{u_j}^2$.
- Yet this automatic selection can be somehow erratic. In particular it can be very sensitive to the number of knots chosen for one component. Sometimes it even grossly oversmooth the component.
- Hence it is important to try other degree of smoothness for a component and inspect the resulting fits.

Smoothing parameter selection

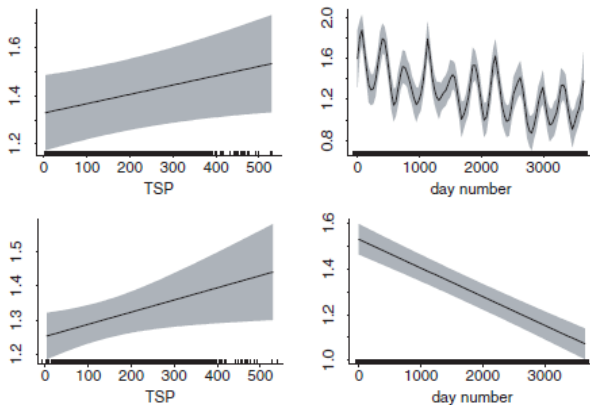


Figure: Additive model of Milan mortality data with smoothing parameters chosen by REML. TSP is a parametric component. Upper row: 35 knots for two other additive components and 60 knots for day number. Lower: 30 knots instead of 35 knots for the two other components. Still 60 knots for day number.

Hypothesis testing for additive model

- The linearity of the effect of a general predictor s can be assessed through a test of the hypotheses

$$H_0 : \sigma_s^2 = 0 \longleftrightarrow H_1 : \sigma_s^2 > 0,$$

where σ_s^2 is the variance for the spline basis function coefficients for estimating the effect of s .

- Likelihood ratio test can be used, but finding p-value is not an easy task. One can use the chi-square approximation, where degree of freedom under the null is the change in the degree of freedom between the models.
- This is well a crude approximation in penalized spline regression. But it is a **good** approximation if it is not penalized, since effectively unpenalized spline regression is just GLM, and the generalized likelihood ratio test (GLRT) applies.
- Hence if hypothesis testing is a key aim, it is sometimes preferable to carefully construct basis and knots so that unpenalized spline regressions are compared.

Hypothesis testing for additive model

- In general for semiparametric regression, if we are testing **two nested models**, we can construct the usual F-ratio

$$F = \frac{(RSS_0 - RSS_1)/(df_{res,1} - df_{res,0})}{RSS_1/(n - df_{res,1})} = \frac{(R_1^2 - R_0^2)/(df_{res,1} - df_{res,0})}{(1 - R_1^2)/(n - df_{res,1})},$$

where RSS_0 is the residual sum of squares for the smaller model, and RSS_1 is that for the larger model.

- Also $df_{res,0}$ is the degree of freedom of the RSS for the smaller model, and similarly for $df_{res,1}$.
- We also have $R_i^2 = 1 - \frac{RSS_i}{TSS}$, where TSS is the total corrected sum of squares.
- Under the null hypothesis, the F ratio will have an approximate F-distribution with degree of freedom

$$df_{res,0} - df_{res,1} \quad \text{and} \quad df_{res,1}$$

- These degrees of freedom will not be integer in general.

Exponential family and GLM

- The 1-parameter exponential family of distribution for the response y has density of the form

$$f(y; \theta) = \exp\left(\frac{y\theta - c(\theta)}{\phi} + d(y, \phi)\right),$$

where θ is called the canonical parameter and ϕ the dispersion parameter. Many well known distributions belong to this family.

- Can prove easily that $E(y) = c'(\theta)$, $\text{var}(y) = \phi c''(\theta)$.
- We usually model the parameter of interest with a linear predictor $\eta = \mathbf{x}^T \boldsymbol{\beta}$, where \mathbf{x} is a vector of predictors. The **canonical link function** is the function that links the canonical parameter to the linear predictor η .
- E.g. for $y \sim \text{Binomial}(n, \pi)$, the canonical parameter is $\theta = \log(\pi/(1 - \pi))$. Setting $\theta = \eta$, we immediately know that the canonical link function is the **logit** function.
- We fit the model by numerical methods of finding solutions of the score equation $\mathbf{U}(\boldsymbol{\beta}) := \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \mathbf{0}$, like the **Newton's method**. Updating equation is

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k - \mathbf{H}(\boldsymbol{\beta}_k)^{-1} \mathbf{U}(\boldsymbol{\beta}_k),$$

where $\mathbf{H}(\boldsymbol{\beta}) = \frac{\partial \mathbf{U}}{\partial \boldsymbol{\beta}^T}$ is called the **Hessian matrix**. The **Fisher scoring** method replaces $-\mathbf{H}(\boldsymbol{\beta})$ by $I(\boldsymbol{\beta})$, the information matrix. The **iterative reweighted least square** (IRLS) algorithm is based on this.

Generalized linear mixed model (GLMM)

- An extension of GLM to GLMM involves replacing $\mathbf{X}\beta$ by $\mathbf{X}\beta + \mathbf{Z}\mathbf{u}$ as in the introduction to mixed model.
- If $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}_\theta)$, where θ is a vector of (unknown) parameters, then for a random sample of size n from the exponential family with $\phi = 1$ using the canonical link function, we have

$$f(\mathbf{y}|\mathbf{u}) = \exp\{\mathbf{y}^T(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T c(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}) + \mathbf{1}^T c(\mathbf{y})\},$$
$$f(\mathbf{u}) = (2\pi)^{-q/2} |\mathbf{G}_\theta|^{-1/2} \exp\left(-\frac{1}{2} \mathbf{u}^T \mathbf{G}_\theta^{-1} \mathbf{u}\right).$$

- This gives the likelihood

$$L(\beta, \theta) = \int_{\mathbb{R}^q} f(\mathbf{y}|\mathbf{u}) f(\mathbf{u}) d\mathbf{u}$$
$$= (2\pi)^{-q/2} |\mathbf{G}_\theta|^{-1/2} \exp\{\mathbf{1}^T c(\mathbf{y})\} J(\beta, \theta),$$

where

$$J(\beta, \theta) = \int_{\mathbb{R}^q} \exp\{\mathbf{y}^T(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T c(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}) - \frac{1}{2} \mathbf{u}^T \mathbf{G}_\theta^{-1} \mathbf{u}\} d\mathbf{u}.$$

Penalized Quasilikelihood (PQL)

- The above likelihood is not easy to deal with because of the integral $J(\beta, \theta)$. The easier way is to treat θ as known first, then treating \mathbf{u} as fixed parameters, (β, \mathbf{u}) is obtained by maximizing the penalized log-likelihood

$$\log\{f(\mathbf{y}|\mathbf{u})\} - \frac{1}{2}\mathbf{u}^T \mathbf{G}_\theta^{-1} \mathbf{u}.$$

- We still assume that $\phi = 1$ for our exponential family and canonical link is used. Differentiating w.r.t. $(\beta^T, \mathbf{u}^T)^T$ and set to $\mathbf{0}$, we get

$$\begin{pmatrix} \mathbf{X}^T(\mathbf{y} - \mu) \\ \mathbf{Z}^T(\mathbf{y} - \mu) - \mathbf{G}_\theta^{-1} \mathbf{u} \end{pmatrix} = \mathbf{0}.$$

- Differentiating LHS further w.r.t. (β^T, \mathbf{u}^T) , we obtain the Hessian

$$-\begin{pmatrix} \mathbf{X}^T \mathbf{W} \mathbf{X} & \mathbf{X}^T \mathbf{W} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{W} \mathbf{X} & \mathbf{Z}^T \mathbf{W} \mathbf{Z} + \mathbf{G}_\theta^{-1} \end{pmatrix},$$

where $\mathbf{W} = \text{diag}(c''(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}))$. This is independent of the data \mathbf{y} , and hence the Fisher's scoring and the Newton's method are identical.

Penalized Quasilikelihood (PQL)

- To estimate θ or \mathbf{G}_θ , one can use cross-validation (not covered).
- Or define the pseudodata

$$\mathbf{y}_{\text{pseudo}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{W}^{-1}(\mathbf{y} - \boldsymbol{\mu}),$$

where $\boldsymbol{\epsilon}_{\text{pseudo}} = \mathbf{W}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ and $\text{var}(\mathbf{y}_{\text{pseudo}}) = \mathbf{R} = \mathbf{W}^{-1}$.

- One can then use profile likelihood or the REML to find \mathbf{G}_θ using current estimates of $(\boldsymbol{\beta}, \mathbf{u})$, where the profile log-likelihood is the one used in mixed model formulation for estimating $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$ when $\mathbf{y}|\mathbf{u} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R})$, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$. In our case, the data is replaced by the pseudo data, with $\mathbf{R} = \mathbf{W}^{-1}$, and $\mathbf{G} = \mathbf{G}_\theta$.
- We can then iterate between estimating $(\boldsymbol{\beta}, \mathbf{u})$ and \mathbf{G}_θ until convergence.

Generalized Additive Model

- GLM is nonlinear, but still a parametric model. It can be extended to nonparametric functions for each independent variable, resulting in the generalized additive model (GAM).
- Again assume that we have $\phi = 1$ and we are using the canonical link function. Hence if we have two independent variables s and t , say, we have

$$\eta(s, t) = \beta_0 + f(s) + g(t).$$

A linear spline formulation is then

$$\eta(s, t) = \mathbf{X}_x \boldsymbol{\beta} + \mathbf{Z}_x \mathbf{u},$$

where $\mathbf{X}_x = (1, s, t)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$, $\mathbf{u} = (u_1^s, \dots, u_{K_s}^s, u_1^t, \dots, u_{K_t}^t)^T$,

$$\mathbf{Z}_x = ((s - \kappa_1^s)_+ \dots (s - \kappa_{K_s}^s)_+ (t - \kappa_1^t)_+ \dots (t - \kappa_{K_t}^t)_+).$$

- Suppose we have sample of size n , then the log-likelihood of the data is

$$\mathbf{y}^T (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T c(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) + \mathbf{1}^T d(\mathbf{y}),$$

where $\mathbf{X} = [\mathbf{X}_{x_i}]_{1 \leq i \leq n}$, $\mathbf{Z} = [\mathbf{Z}_{x_i}]_{1 \leq i \leq n}$, with $x_i = (s_i, t_i)$.

Generalized Additive Model

- A penalty on the roughness can be thought of as penalizing on the second derivative of the η w.r.t. s and t , hence in the end the penalized log-likelihood is to be maximized w.r.t. (β, \mathbf{u}) is

$$\mathbf{y}^T (\mathbf{X}\beta + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T c(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}) - \lambda_s \int_{\mathbb{R}} (f''(s))^2 ds - \lambda_t \int_{\mathbb{R}} (g''(t))^2 dt.$$

- Since $f''(s) = \eta_{ss}(s, t) = \mathbf{X}_{x,ss}\beta + \mathbf{Z}_{x,ss}\mathbf{u}$ and similarly $g''(t) = \mathbf{X}_{x,tt}\beta + \mathbf{Z}_{x,tt}\mathbf{u}$, the above penalized log-likelihood can be written as

$$\mathbf{y}^T (\mathbf{X}\beta + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T c(\mathbf{X}\beta + \mathbf{Z}\mathbf{u}) - \begin{pmatrix} \beta \\ \mathbf{u} \end{pmatrix}^T \mathbf{B} \begin{pmatrix} \beta \\ \mathbf{u} \end{pmatrix},$$

where $\mathbf{B} = \lambda_s \int_{\mathbb{R}} \mathbf{C}_{x,ss}^T \mathbf{C}_{x,ss} ds + \lambda_t \int_{\mathbb{R}} \mathbf{C}_{x,tt}^T \mathbf{C}_{x,tt} dt$, and $\mathbf{C}_x = (\mathbf{X}_x, \mathbf{Z}_x)$.

Generalized Additive Mixed Model

- If we assume a mixed model formulation with $\log\{f(\mathbf{y}|\mathbf{u})\}$ given as in page 12 and

$$\mathbf{u} \sim N\left(\mathbf{0}, \begin{bmatrix} \sigma_s^2 \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \sigma_t^2 \mathbf{I} \end{bmatrix}\right),$$

then (to be shown in exercise 2) if given σ_s^2 and σ_t^2 , we are in effect maximising

$$\mathbf{y}^T (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \mathbf{1}^T c(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}) - \frac{1}{2} \mathbf{u}^T \text{cov}(\mathbf{u})^{-1} \mathbf{u}.$$

- This is the PQL like before, and hence fitting can be done by maximising this PQL through the IRLS on a properly defined pseudo data, or directly evaluate a Fisher scoring updating equation for the iterations. Note that the expression immediately after (11.6) in the book is WRONG.
- Under normality, we can obtain the profile log-likelihood for the variance components in closed form as before.

Degree of freedom approximations

- Let $\mu(\eta)$ and $V(\eta)$ be the conditional mean and variance of y given $\eta = \mathbf{x}^T \boldsymbol{\beta} + \mathbf{z}^T \mathbf{u}$. Define $\mathbf{W} = \text{diag}(\mu'(\mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \mathbf{z}_i^T \hat{\mathbf{u}}))$.
- A generalized hat matrix can be defined as

$$\mathbf{H} = \mathbf{W}\mathbf{X}(\mathbf{X}^T \mathbf{W}\mathbf{X} + \Lambda/2)^{-1} \mathbf{X}^T,$$

where $\Lambda = \text{diag}(\mathbf{0}, \text{cov}(\mathbf{u})^{-1})$.

- Degree of freedom of the fit is

$$\text{tr}(\mathbf{H}) = \text{tr}((\mathbf{X}^T \mathbf{W}\mathbf{X} + \Lambda/2)^{-1} \mathbf{X}^T \mathbf{W}\mathbf{X}).$$

- We also have

$$\text{cov} \left(\begin{array}{c} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{array} \right) \Big|_{\mathbf{u}} \approx (\mathbf{X}^T \mathbf{W}\mathbf{X} + \Lambda/2)^{-1} (\mathbf{X}^T \mathbf{W}\mathbf{X}) (\mathbf{X}^T \mathbf{W}\mathbf{X} + \Lambda/2)^{-1}.$$


Varying Coefficient Models

- It is a special class of interaction models, where interaction between two continuous variables is considered.
- For two variables x_1 and x_2 , in classical linear regression an interaction between x_1 and x_2 is usually captured by adding a term $\gamma x_1 x_2$, so that the model becomes $E(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \gamma x_1 x_2 = \alpha + (\beta_1 + \gamma x_2) x_1 + \beta_2 x_2$. That is, **the effect** of changing x_1 on the average response y **is not constant**, and depends on the value of x_2 **linearly**.
- The linearity of the effect on x_2 is not always true. It is sometimes better to consider a more general model. If (x_i, s_i, y_i) , $1 \leq i \leq n$, a **varying coefficient model** for the data is

$$y_i = \alpha(s_i) + \beta(s_i)x_i + \epsilon_i.$$

- The penalized linear spline version of this model is

$$y_i = \alpha_0 + \alpha_1 s_i + \sum_{k=1}^{K_1} u_k^\alpha (s_i - \kappa_k^1)_+ + \left(\beta_0 + \beta_1 s_i + \sum_{k=1}^{K_2} u_k^\beta (s_i - \kappa_k^2)_+ \right) x_i + \epsilon_i.$$

- The book set $K_1 = K_2$ with same set of κ_k 's for both $\alpha(s)$ and $\beta(s)$. There is NO reason and certainly wrong in some cases to use the same set of knots. 

Varying Coefficient Models

- We can obtain a mix model representation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$ by setting

$$\mathbf{Z} = [(s_i - \kappa_1^1)_+, \dots, (s_i - \kappa_{K_1}^1)_+, x_i(s_i - \kappa_1^2)_+, \dots, x_i(s_i - \kappa_{K_2}^2)_+],$$

$$\mathbf{X} = [1, s_i, x_i, s_i x_i]_{1 \leq i \leq n}, \quad \mathbf{u} = [u_1^\alpha, \dots, u_{K_1}^\alpha, u_1^\beta, \dots, u_{K_2}^\beta]^T,$$

$$\text{cov}(\mathbf{u}) = \text{diag}(\sigma_\alpha^2, \dots, \sigma_\alpha^2, \sigma_\beta^2, \dots, \sigma_\beta^2).$$

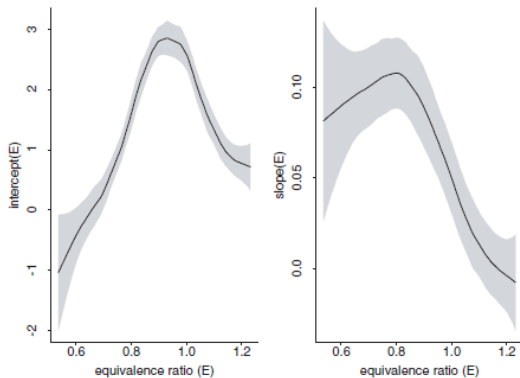


Figure: Fit of varying coefficient model to ethanol data