Smoothing Samples of Curves

aka

Longitudinal Analysis

STAT 526

Mike Danilov

based on paper by Brumback and Rice, 1998

Overview

- The topic
- My data
- Smoothing splines and Mixed-Effects Models

  uni-subject case

  sample of curves

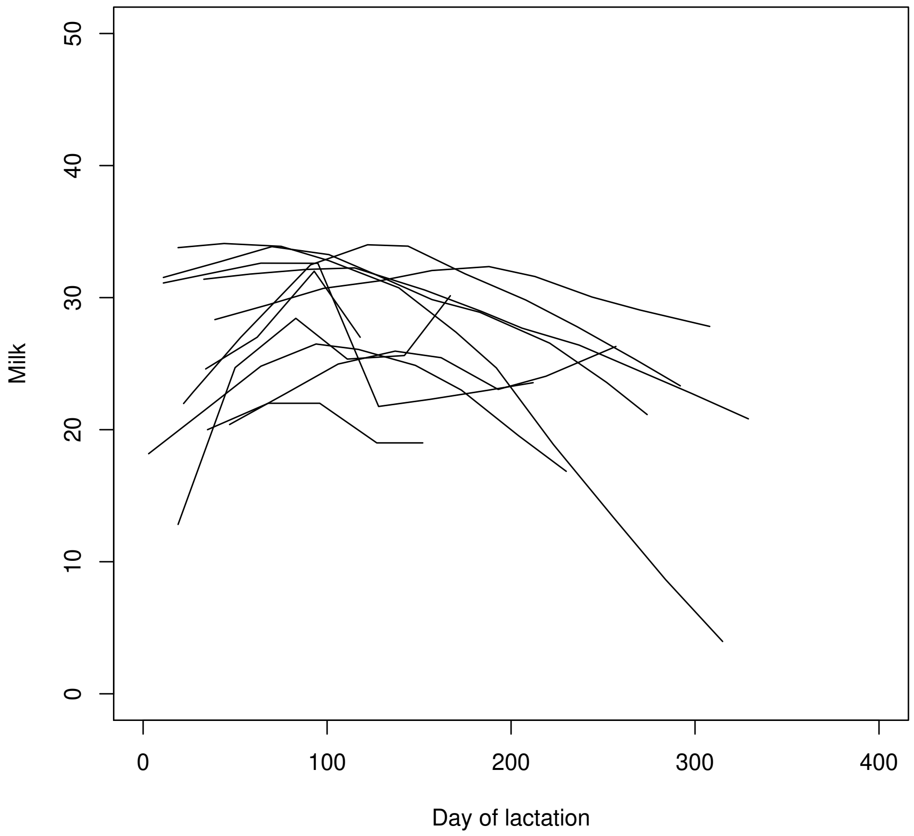
The topic

- Set of curves (usually subjects)
- Borrow knowledge from one curve to another

Mike's data

- Lactation curve – amount of milk given by a cow as a function of time since the last calving
- Many cows, many lactations for each
- Suspect: the form of lactation curve depends on the amount of milk and on the number of lactation
- Other covariates (not to be considered now)

Visualization



...if there are too many subjects it makes sense to plot only some of them

Model

$$y_{ij} = s_\mu(t_{ij}) + s_{g(i)}(t_{ij}) + s_{l(i)}(t_{ij}) + \varepsilon_{ij} \tag{1}$$

- ANOVA-like
- s_μ – general curve for everyone
- $g(i)$ – productivity group, e.g. $g(i) = \begin{cases} 0, & y_{i1} \le 10 \\ 1, & y_{i1} \in (10, 15] \\ 2, & y_{i1} \in (15, 25] \\ 3, & y_{i1} > 25 \end{cases}$

  (as we want to predict the curve after beginning of lactation)
- $l(i)$ – lactation number

Mixed models in one-curve case

If the model is

$$y_i = s(t_i) + \varepsilon_i \tag{2}$$

then we know that the minimizer of

$$\sum_{i=1}^{n}(y(t_i) - s(t_i))^2 + \lambda \int (s''(t))^2 dt \tag{3}$$

is called *the cubic smoothing spline* to the data.

It is equivalent to the BLUP solution to a Mixed Effect model constructed as follows.

Construction of the Mixed Effects Model

- $\hat{\mathbf{s}} = (\hat{s}(t_1), \ldots, \hat{s}(t_n))^T$
- $B_i, i = 1, \ldots, n$ — basis of natural cubic spines with knots at $t_2, \ldots, t_{n-1}$, such that B_1, B_2 are constant-linear.
- $Q = \left(\int B_i''(t) B_j''(t) dt\right)_{ij}$, $i = 3, \ldots, n$, $j = 3, \ldots, n$
- $Q = UDU^T$ — singular value decomposition
- $X = (B_j(t_i))_{ij}$, $i = 1, \ldots, n$, $j = 1, 2$
- $B = (B_j(t_i))_{ij}$, $i = 1, \ldots, n$, $j = 3, \ldots, n$
- $Z = BUD^{-1/2}$

then $\hat{\mathbf{s}}$ is the BLUP solution $\mathbf{y} = X\hat{\beta} + Z\hat{\mathbf{u}}$ to

The Mixed Model

$$\mathbf{y} = X\beta + Z\mathbf{u} + \varepsilon, \tag{4}$$

where $\mathbf{u} \sim MV\mathcal{N}(\mathbf{0}, (\sigma^2/\lambda)I_{n-2})$, $\varepsilon \sim MV\mathcal{N}(\mathbf{0}, \sigma^2 I_n)$ independent of $\mathbf{u}$.

Solution is given by

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}\mathbf{y} \tag{5}$$

$$\hat{\mathbf{u}} = (\sigma^2/\lambda)Z'V^{-1}(\mathbf{y} - X\hat{\beta}), \tag{6}$$

where $V = \mathbf{Var}\ (\mathbf{y}) = (\sigma^2/\lambda)ZZ' + \sigma^2 I_n$ is assumed to be known.

- We model linear part by fixed effects and matrix X
- The wiggle part is represented by random effects and by matrix Z. Columns of Z correspond to different smoothing functions and it somehow happens that smoother ones have more weight than wiggle ones.

$\boxed{\sigma^2 \text{ AND } \lambda \text{ ISSUES}}$

In practice they must be estimated

- by cross-validation
- by likelihood methods (REML, GML)

Big part of the paper is concerned about fast enough method of estimating them (beyond the scope of the presentation)

Modeling a Crossed Sample of Curves

Recall model (1)

$$y_{ij} = s_\mu(t_{ij}) + s_{g(i)}(t_{ij}) + s_{l(i)}(t_{ij}) + \varepsilon_{ij} \tag{7}$$

Generalizing what was said before we can write for a single curve:

$$\mathbf{y}_i = X_i\beta_\mu + Z_i\mathbf{u}_\mu + X_i\beta_{g(i)} + Z_i\mathbf{u}_{g(i)} + X_i\beta_{l(i)} + Z_i\mathbf{u}_{l(i)} + \varepsilon_i \tag{8}$$

- Let $\{\nu_i\}_{i=1}^T$ be the ordered set of observation times from all the curves taken together. B_j, $j = 1, \ldots, T$.
- Subscripts at design matrices to indicate curve specific time points: $X_i = (B_j(t_il))_{jl}$, $j = 3, \ldots, T+2$, $l = 1, \ldots, n_i$; $Z_i = B_iUD^{-1/2}$, $\mathbf{u}_\mu \sim MV\mathcal{N}(\mathbf{0}, (\sigma^2/\lambda_\mu)I_{T-2})$; $\mathbf{u}_{g(i)} \sim MV\mathcal{N}(\mathbf{0}, (\sigma^2/\lambda_g)I_{T-2})$; $\mathbf{u}_{l(i)} \sim MV\mathcal{N}(\mathbf{0}, (\sigma^2/\lambda_l)I_{T-2})$;

Big model

- Stacking all such equations together, we get the big mixed model equation:

$$\mathbf{y} = X_\mu \beta_\mu + Z_\mu \mathbf{u}_\mu + X_g \beta_g + Z_l \mathbf{u}_g + X_l \beta_g + Z_l \mathbf{u}_g + \varepsilon \qquad (9)$$

  where all vectors are just stacked ones from small models, and matrices are block-diagonal stocked ones.

- Identifiability issues must be solved by putting some constrains on fixed factors.

- In the model, observations $\mathbf{y}_{ij}$ are not independent now — being inside same block increases correlation between them.

Equivalency to splines

Theorem The BLUP solutions to the system (9) are equivalent to the fitted curves given by a corresponding penalized regression:

$$\sum_i \sum_j (y_{ij} - (s_\mu(t_{il}) + s_{g(i)}(t_{il}) + s_{l(i)}(t_{il})))^2 + \lambda_\mu \int (s''_\mu(t))^2 dt + \lambda_g \sum_j \int (s_{b_j}(t))^2 dt + \lambda_l \sum_k \int (s_{l_k}(t))^2 dt \quad (10)$$

and they are natural cubic splines with knots at $(\nu_2, \ldots, \nu_{T-1})$

Other issues

- σ^2 and λ's need to be estimated and this is a big problem in terms of time and memory.
- Variability of those estimates is not easy to estimate as well.
- Once estimated, σ^2/λ_μ, σ^2/λ_g, σ^2/λ_l may serve as measure of relative importance of groups involved into analysis.