# Simple fitting of subject-specific curves for longitudinal data

Maria Durbán y Matt Wand

Dept. Estadística y Econometría, Universidad Carlos III de Madrid

Department of Statistics University of New South Wales

IX Conferencia Española de Biometría, A Coruña, Mayo 2003 .

# What is this talk about?

# What is this talk about?

- Introduction
  - ⋆ The data
  - ⋆ P-splines
  - ⋆ Mixed model representation of P-splines

## What is this talk about?

- Introduction
  - ⋆ The data
  - ⋆ P-splines
  - ⋆ Mixed model representation of P-splines
- P-spline models for longitudinal data.
  - ⋆ Subject-specific Curves

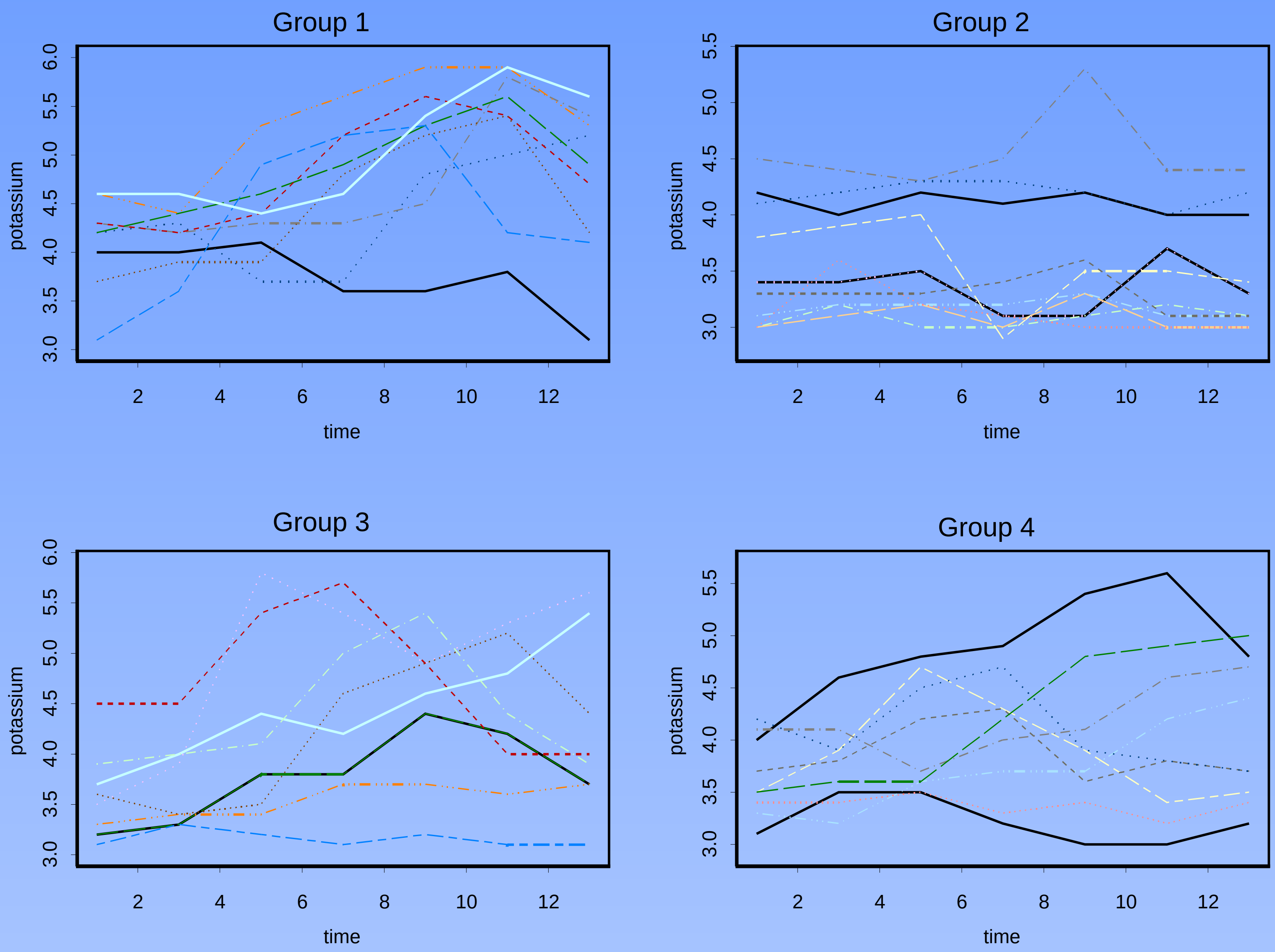
# What is this talk about?

- Introduction
  - ⋆ The data
  - ⋆ P-splines
  - ⋆ Mixed model representation of P-splines
- P-spline models for longitudinal data.
  - ⋆ Subject-specific Curves
- Analysis of dogs data.

# The data

Objetive: Determine the effect of 4 surgical treatments on coronary sinus potasium in dogs

- 36 dogs
- 4 treatments
- 7 measurements per dog


Group 1
Group 2
Group 3
Group 4
potassium
time
2
4
6
8
10
12
3.0
3.5
4.0
4.5
5.0
5.5
6.0

# Smoothing

- For each dog, potassium concentration varies smoothly along time

# Smoothing

- For each dog, potassium concentration varies smoothly along time

- A suitable model for these data could be:

$$\boldsymbol{y} = f(\boldsymbol{t}) + \boldsymbol{\epsilon}$$

where $t$ is the covariate (Time) and $f$ is a smooth function of $t$ which depends on $\lambda$ =smoothing parameter

# Smoothing

- For each dog, potassium concentration varies smoothly along time

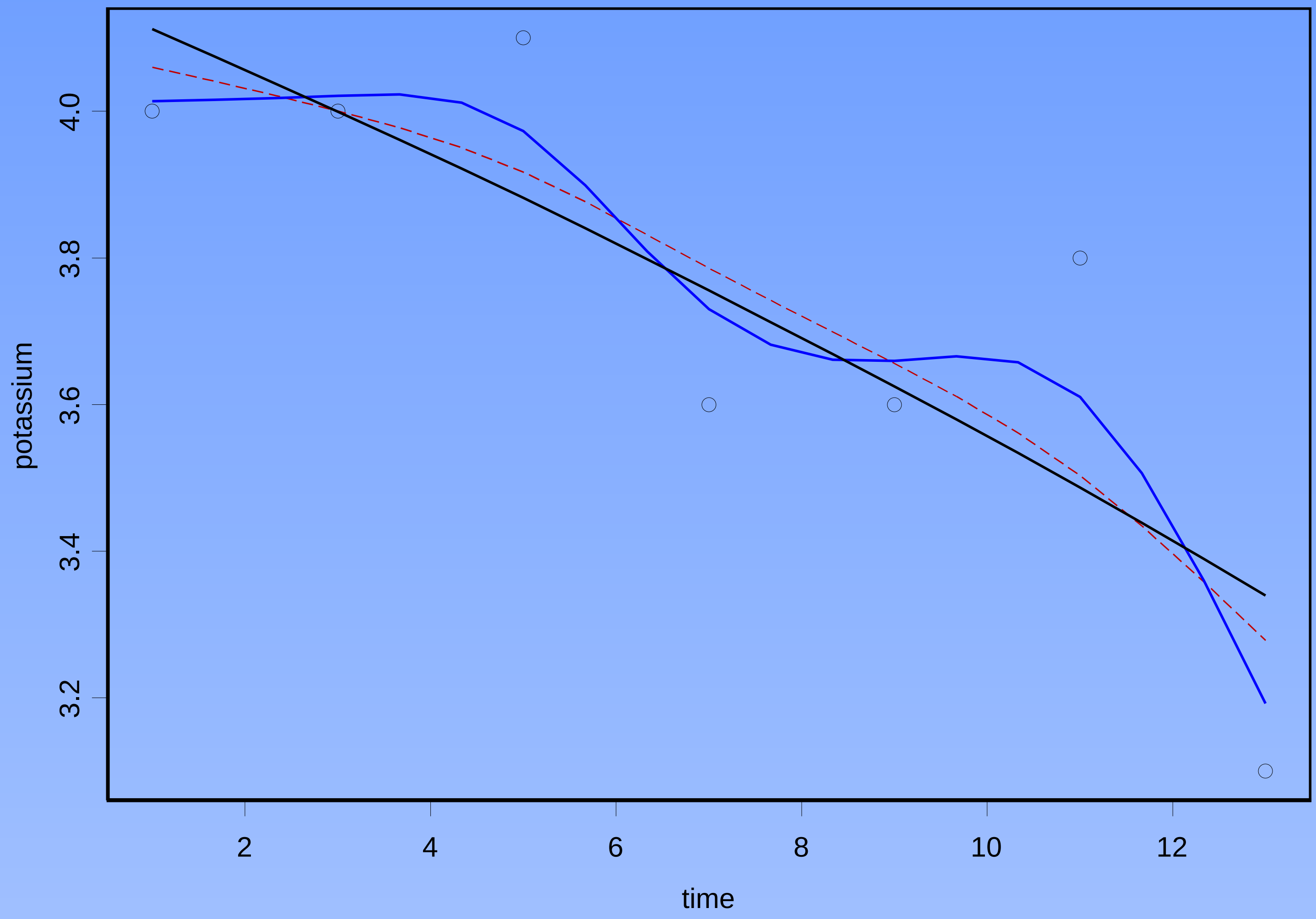
- A suitable model for these data could be:

$$\boldsymbol{y} = f(\boldsymbol{t}) + \boldsymbol{\epsilon}$$

  where $t$ is the covariate (Time) and $f$ is a smooth function of $t$ which depends on $\lambda =$smoothing parameter

- Smoothing methods fall into two groups:

  ⋆ Specified by the fitting procedure: **Kernels**
  ⋆ Solution of a minimisation problem: **Splines**

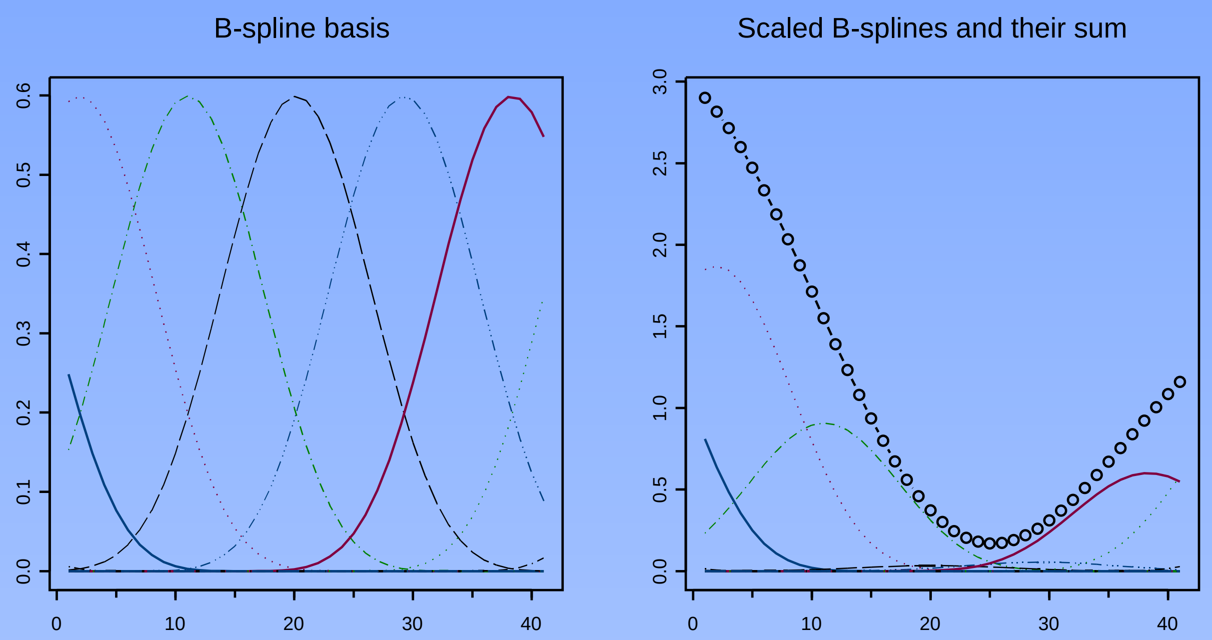

potassium
4.0
3.8
3.6
3.4
3.2
2
4
6
8
10
12
time

# B-splines

- We write $f = \boldsymbol{Ba}$, $\boldsymbol{B}$ is a B-spline basis and $\boldsymbol{a}$ are coefficients
- B-spline: bell-shaped like Gauss curve
- Used as the basis for the regression
- Polynomial pieces smoothly joining at the knots

# B-splines

- We write $f = \boldsymbol{B}\boldsymbol{a}$, $\boldsymbol{B}$ is a B-spline basis and $\boldsymbol{a}$ are coefficients
- B-spline: bell-shaped like Gauss curve
- Used as the basis for the regression
- Polynomial pieces smoothly joining at the knots



# Penalized splines (P-splines)

# Penalized splines (P-splines)

- Regress on $k$ B-splines
- Difference penalty on the regression coefficients

# Penalized splines (P-splines)

- Regress on $k$ B-splines
- Difference penalty on the regression coefficients
- Minimise penalised sum of squares

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon} \quad f(\boldsymbol{x}) \approx \boldsymbol{B}\boldsymbol{a} \quad S = (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a})'(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}) + \lambda \boldsymbol{a}'\boldsymbol{D}'\boldsymbol{D}\boldsymbol{a}$$

$$\hat{\boldsymbol{a}} = (\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\boldsymbol{y}$$

# Penalized splines (P-splines)

- Regress on $k$ B-splines
- Difference penalty on the regression coefficients
- Minimise penalised sum of squares

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon} \quad f(\boldsymbol{x}) \approx \boldsymbol{Ba} \quad S = (\boldsymbol{y} - \boldsymbol{Ba})'(\boldsymbol{y} - \boldsymbol{Ba}) + \lambda \boldsymbol{a}'\boldsymbol{D}'\boldsymbol{Da}$$

$$\hat{\boldsymbol{a}} = (\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\boldsymbol{y}$$

- Moderate dimension ($k$ between 10 and 40)
- Computationally faster than smoothing splines

# Mixed model representation of P-splines

# Mixed model representation of P-splines

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{a} + \boldsymbol{\epsilon}$$

$\Downarrow$ $\boldsymbol{a}$ treated as mixed effects

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \quad \boldsymbol{u} \sim N(0, \sigma_u^2)$$

# Mixed model representation of P-splines

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{a} + \boldsymbol{\epsilon}$$

$\Downarrow$ $\boldsymbol{a}$ treated as mixed effects

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \quad \boldsymbol{u} \sim N(0, \sigma_u^2)$$

$$\lambda = \sigma^2/\sigma_u^2$$

# Mixed model representation of P-splines

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{a} + \boldsymbol{\epsilon}$$

$\Downarrow$ $\boldsymbol{a}$ treated as mixed effects

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \quad \boldsymbol{u} \sim N(0, \sigma_u^2)$$

$$\lambda = \sigma^2/\sigma_u^2$$

**Best Linear Unbiased Predictor = Penalized likelihood fit**

# Advantages

# Advantages

- No need for backfitting in additive models

# Advantages

- No need for backfitting in additive models
- Fast with large datasets, not posible with smoothing splines

# Advantages

- No need for backfitting in additive models
- Fast with large datasets, not posible with smoothing splines
- Easily implemented in Splus and R

```
lme(y~X-1,random=pdIdent(~Z-1))
```

# Advantages

- No need for backfitting in additive models
- Fast with large datasets, not posible with smoothing splines
- Easily implemented in Splus and R

```
lme(y~X-1,random=pdIdent(~Z-1))
```

- GAM $\Rightarrow$ GLMM

# P-splines with truncared lines basis

$$y_{ij} = f(t_{ij}) + \epsilon_{ij} \quad 1 \leq j \leq 7$$

$$f(t_{ij}) = \alpha_0 + \alpha_1 t_{ij} + \sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+ \quad u_k \sim N(0, \sigma_u^2)$$

$$\Downarrow$$

$$\boldsymbol{y} = \boldsymbol{X\alpha} + \boldsymbol{Zu} + \boldsymbol{\epsilon} \quad Cov \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \sigma_u^2 \boldsymbol{I} & 0 \\ 0 & \sigma^2 \boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{Z} = \begin{bmatrix} (t_{i1} - \kappa_1)_+ & \dots & (t_{i1} - \kappa_K)_+ \\ \vdots & \ddots & \vdots \\ (t_{i7} - \kappa_1)_+ & \dots & (t_{i7} - \kappa_K)_+ \end{bmatrix}$$

# P-spline models for longitudinal data

Basic Model $\quad y_{ij} = \alpha_0 + \alpha_1 t_{ij} + \beta_{i0} + \epsilon_{ij} \quad 1 \leq j \leq 7 \quad 1 \leq i \leq 36$

# P-spline models for longitudinal data

Basic Model $y_{ij} = \alpha_0 + \alpha_1 t_{ij} + \beta_{i0} + \epsilon_{ij} \quad 1 \leq j \leq 7 \quad 1 \leq i \leq 36$

$\Downarrow$ Relax linearity assumption

Model A $y_{ij} = f_{gr(i)}(t_{ij}) + \beta_{i0} + \epsilon_{ij} \quad 1 \leq gr(i) \leq 4$

# P-spline models for longitudinal data

Basic Model $y_{ij} = \alpha_0 + \alpha_1 t_{ij} + \beta_{i0} + \epsilon_{ij} \quad 1 \leq j \leq 7 \quad 1 \leq i \leq 36$

$\Downarrow$ Relax linearity assumption

Model A $y_{ij} = f_{gr(i)}(t_{ij}) + \beta_{i0} + \epsilon_{ij} \quad 1 \leq gr(i) \leq 4$

$\Downarrow$ Add random slope + general covariance matrix

Model B $y_{ij} = f_{gr(i)}(t_{ij}) + \beta_{i0} + \beta_{i1} t_{ij} + \epsilon_{ij}$

# P-spline models for longitudinal data

Basic Model $y_{ij} = \alpha_0 + \alpha_1 t_{ij} + \beta_{i0} + \epsilon_{ij} \quad 1 \le j \le 7 \quad 1 \le i \le 36$

$\Downarrow$ Relax linearity assumption

Model A $y_{ij} = f_{gr(i)}(t_{ij}) + \beta_{i0} + \epsilon_{ij} \quad 1 \le gr(i) \le 4$

$\Downarrow$ Add random slope + general covariance matrix

Model B $y_{ij} = f_{gr(i)}(t_{ij}) + \beta_{i0} + \beta_{i1} t_{ij} + \epsilon_{ij}$

$\Downarrow$ Subject specific curves

Model C $y_{ij} = f_{gr(j)}(t_{ij}) + g_i(t_{ij}) + \epsilon_{ij}$

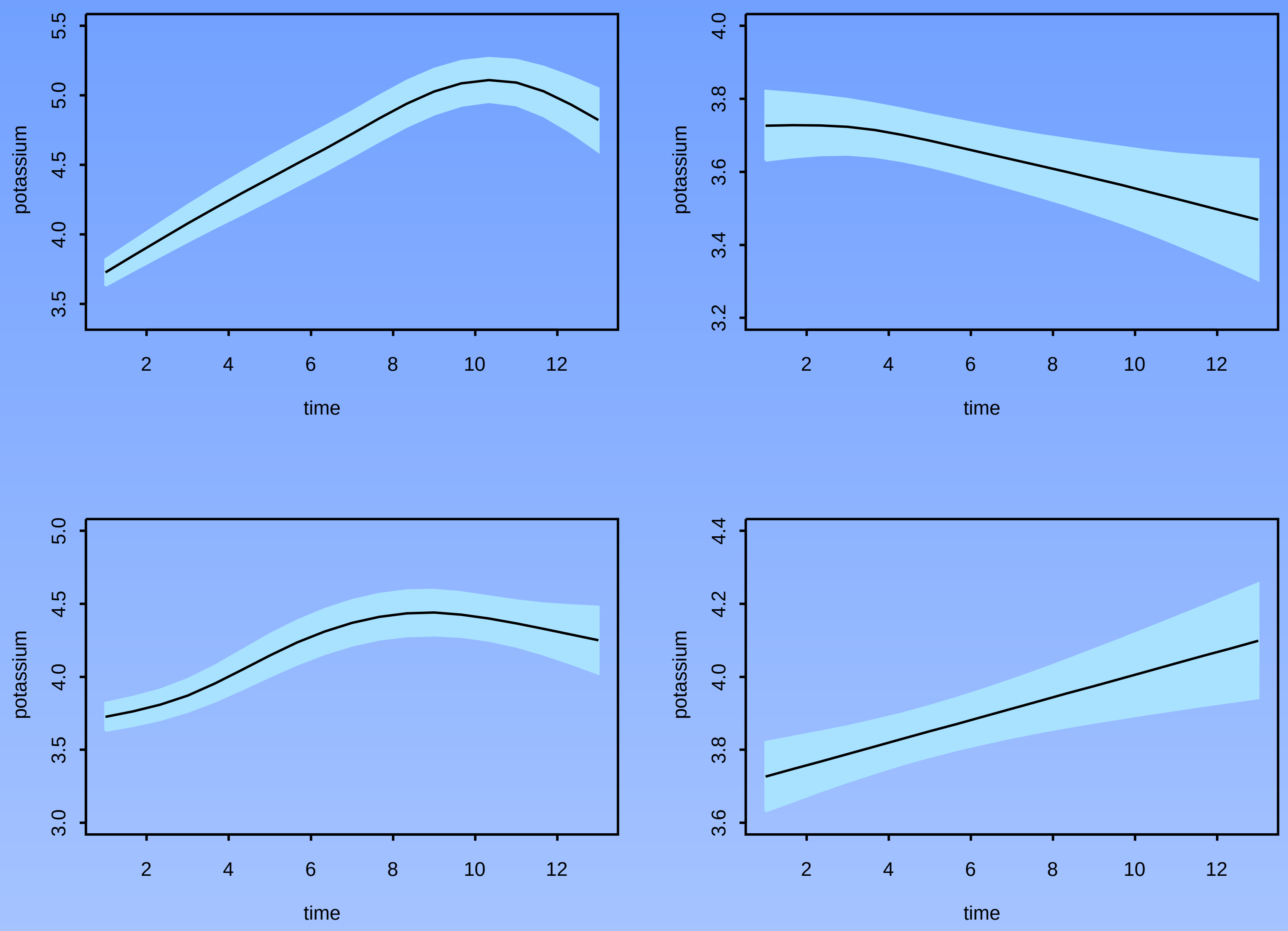
The mixed model associated to Model A is:



$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \quad Cov \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \sigma^2_{\beta_0}\boldsymbol{I} & 0 & 0 \\ 0 & \Sigma_{gr} & 0 \\ 0 & 0 & \sigma^2\boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_{\text{time}} \\ \vdots \\ \boldsymbol{X}_{\text{time}} \end{bmatrix} \quad \boldsymbol{X}_{\text{time}} = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_7 \end{bmatrix} \quad \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & & & & 1 & 0 & \cdots & 0 \\ & & & & \vdots & \vdots & \ddots & \vdots \\ & & & & 1 & 0 & \cdots & 0 \\ & \boldsymbol{Z}_2 & & & 0 & 1 & \cdots & 0 \\ & & & & \vdots & \vdots & \ddots & \vdots \\ & & & & 0 & 1 & \cdots & 0 \\ & & \boldsymbol{Z}_3 & & \vdots & \vdots & \vdots & \vdots \\ & & & & 0 & 0 & \cdots & 1 \\ & & & & \vdots & \vdots & \ddots & \vdots \\ & & & \boldsymbol{Z}_4 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$\boldsymbol{Z}_{\text{gr(i)}} = \begin{bmatrix} \boldsymbol{Z}_{\text{time}} \\ \vdots \\ \boldsymbol{Z}_{\text{time}} \end{bmatrix} \quad \Sigma_{gr} = \begin{bmatrix} \sigma^2_1\boldsymbol{I} & & & \\ & \sigma^2_2\boldsymbol{I} & & \\ & & \sigma^2_3\boldsymbol{I} & \\ & & & \sigma^2_4\boldsymbol{I} \end{bmatrix}$$


potassium
5.5
5.0
4.5
4.0
3.5
2
4
6
8
10
12
time
potassium
4.0
3.8
3.6
3.4
3.2
2
4
6
8
10
12
time
potassium
5.0
4.5
4.0
3.5
3.0
2
4
6
8
10
12
time
potassium
4.4
4.2
4.0
3.8
3.6
2
4
6
8
10
12
time

The mixed model associated to Model B is:

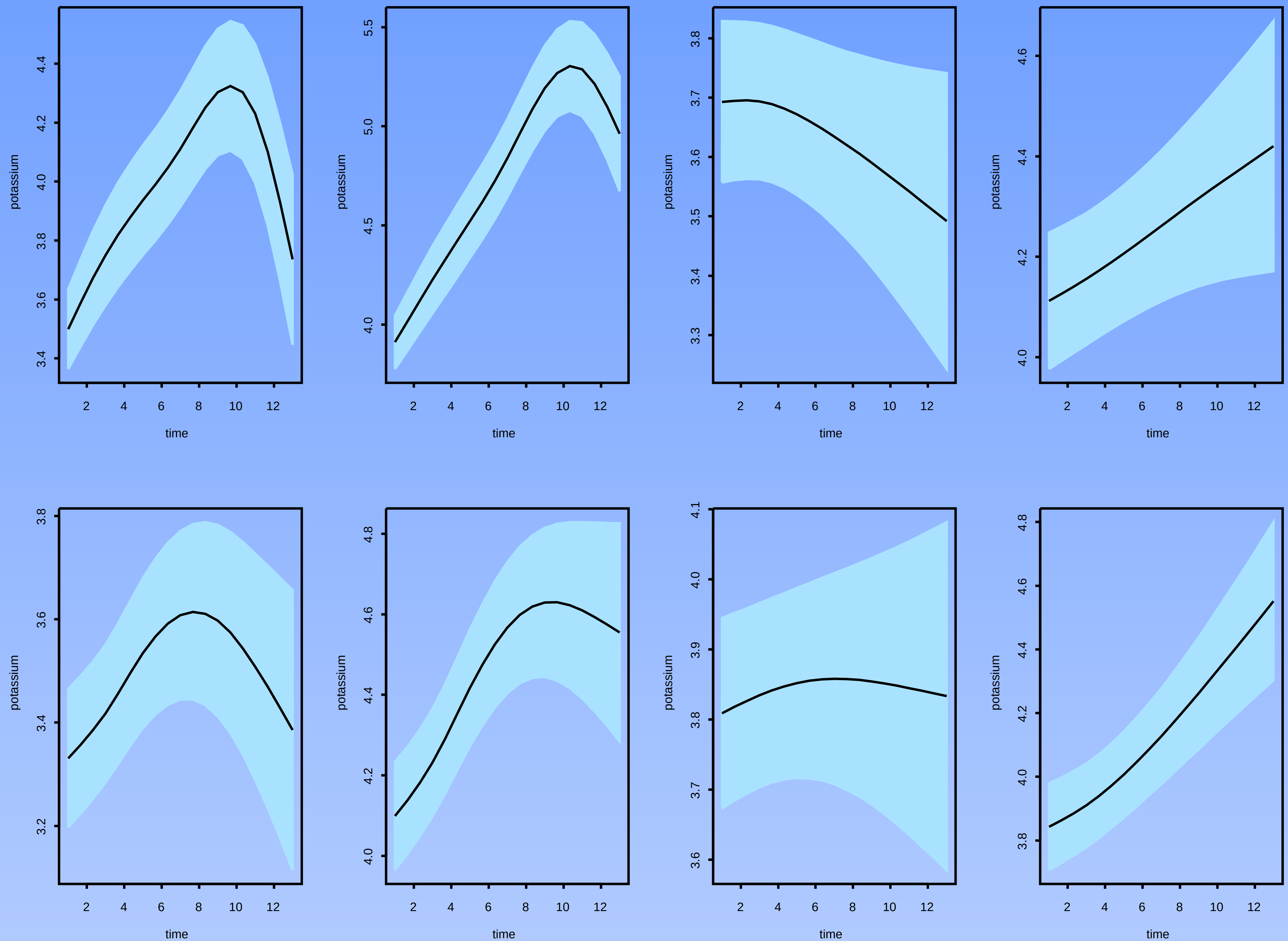
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \quad Cov\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \Sigma_{gr} & 0 & 0 \\ 0 & blockdiag(\Sigma) & 0 \\ 0 & 0 & \sigma^2 \boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{Z} = \left[\begin{array}{cccccccc} \boldsymbol{Z}_1 & & & & \boldsymbol{X}_{\text{time}} & 0 & \cdots & 0 \\ & & & & \vdots & \vdots & \ddots & \vdots \\ & & & & \boldsymbol{X}_{\text{time}} & 0 & \cdots & 0 \\ & \boldsymbol{Z}_2 & & & 0 & \boldsymbol{X}_{\text{time}} & \cdots & 0 \\ & & & & \vdots & \vdots & \ddots & \vdots \\ & & & & 0 & \boldsymbol{X}_{\text{time}} & \cdots & 0 \\ & & \boldsymbol{Z}_3 & & \vdots & \vdots & \vdots & \vdots \\ & & & & 0 & 0 & \cdots & \boldsymbol{X}_{\text{time}} \\ & & & & \vdots & \vdots & \ddots & \vdots \\ & & & \boldsymbol{Z}_4 & 0 & 0 & \cdots & \boldsymbol{X}_{\text{time}} \end{array}\right]$$

The mixed model associated to Model C is:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon} \quad Cov\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{bmatrix} = \begin{bmatrix} \Sigma_{gr} & 0 & 0 & 0 \\ 0 & blockdiag(\Sigma) & 0 & 0 \\ 0 & 0 & \sigma_c^2\boldsymbol{I} & 0 \\ 0 & 0 & 0 & \sigma^2\boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 & & & & \boldsymbol{X}_{\mathsf{time}} & 0 & \cdots & 0 & \boldsymbol{Z}_{\mathsf{time}} & 0 & \cdots & 0 \\ & & & & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ & & & & \boldsymbol{X}_{\mathsf{time}} & 0 & \cdots & 0 & \boldsymbol{Z}_{\mathsf{time}} & 0 & \cdots & 0 \\ & \boldsymbol{Z}_2 & & & 0 & \boldsymbol{X}_{\mathsf{time}} & \cdots & 0 & 0 & \boldsymbol{Z}_{\mathsf{time}} & \cdots & 0 \\ & & & & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ & & & & 0 & \boldsymbol{X}_{\mathsf{time}} & \cdots & 0 & 0 & \boldsymbol{Z}_{\mathsf{time}} & \cdots & 0 \\ & & \boldsymbol{Z}_3 & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & & & & 0 & 0 & \cdots & \boldsymbol{X}_{\mathsf{time}} & 0 & 0 & \cdots & \boldsymbol{Z}_{\mathsf{time}} \\ & & & & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ & & & \boldsymbol{Z}_4 & 0 & 0 & \cdots & \boldsymbol{X}_{\mathsf{time}} & 0 & 0 & \cdots & \boldsymbol{Z}_{\mathsf{time}} \end{bmatrix}$$


potassium
time

# Conclusions and work in progress

# Conclusions and work in progress

- $P$-splines are useful tool to model longitudinal data
- P-splines as mixed models
- Easy to implement in standard sorfware
- Model selection

# References

# References


- Eilers, P.H.C. & Marx, B.D. (1996). Flexible smoothing with B-splines ans penalties. *Statist. Sci.* 11.

- Currie, I. and Durban, M. (2002). Flexible smoothing with P-splines: a unified approach. *Statistical Modelling* 2.

- Wand, M.P. (2003). Smoothing and mixed models. *Comput. Stat.* 18.

- Ruppert, D., Wand, M.P., Carroll, R.J. (2003). Semiparametric Regression. Cambridge University Press.