Semiparametric mixed models based in P-splines

Mariola Sánchez González
CIFOR-INIA
Outline

• Introduction
• P-splines
• Mixed models and P-splines
• Extensions
• Applications
Introduction

Semiparametric mixed models based in P-splines

• Semiparametric regression
  • Smoothing
  • Mixed effects
  • Additive models
  • Splines
Semiparametric regression

- Flexible incorporation of nonlinear functional relationships in regression analyses
- Non-linear relationships can be handled without the restrictions of parametric models
Smoothing

• Scatterplot smoothing: exploratory data analysis method

\[ f(x) = \mathbb{E}(y \mid x) \]

\[ y = f(x) + \varepsilon, \quad \mathbb{E}(\varepsilon_i) = 0 \]

• Methods for smoothing:
  
  ✓ Kernel smoothers (nonparametric models)
  
  ✓ Spline smoothers (Semiparametric models)
Additive models

• Several covariates that may have nonlinear relationships with the response
• Each covariate affects independently of one another
• Understand the effect of each predictor variable on the response variable.
Mixed models

• Models that contain fixed effects and random effects

• Useful tools for analyzing:
  ✓ Longitudinal data
  ✓ Data with grouping structure

• Benefits depend on the application:
  ✓ Suitable variance-covariance matrix
  ✓ Correlation structure
  ✓ Prediction at different level of grouping
Semiparametric mixed models based in P-splines

- Flexibilize lineal hypothesis
- Include complex structures, e.g. interaction between variables in additive models
Outline

• Introduction
• P-splines
• Mixed models and P-splines
• Extensions
• Applications
• SAS: practical session
Semiparametric models with splines

• Spline function: polynomial pieces connected by knots

• Possible approaches to spline estimation:
  ✓ Smoothing splines
  ✓ Regression splines
  ✓ Penalized splines
P-splines methodology

1. Use a regression **basis** $B=B(x)$

   $$ y = f(x) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I) $$

   $$ y = f(x) + \varepsilon = Ba + \varepsilon $$

2. Modify the likelihood function by adding a **penalty** term

   $$ S(a; y, \lambda) = (y - B(x)a)^2 + PENALTY $$
Basis: B-splines

• Several alternatives: truncated functions, thin plate splines, B-splines

• The smooth function is a sum of local basis functions

\[ f(x_i) = a_1 B_1(x_i) + a_2 B_2(x_i) + \ldots + a_K B_K(x_i) \]

• A B-spline of order \( p \) has these properties:

  ✓ \( p + 1 \) polynomial pieces, each of degree \( p \);

  ✓ the polynomial pieces join at \( p \) inner knots;
Basis: B-splines

- B-spline of 1 degree
Basis: B-splines

• B-spline of 3 degree
Basis: Knots

- Specific values of $x$
- Knots number and location is not fixed
  - Equally-spaced
  - Knots number $= \min\{40, \text{unique values of } x/4\}$
- The knots divide the interval of $x$
  \[ x_{\min} = k_1 < k_2 < \ldots < k_{m-1} < k_m = x_{\max} \]
- Number of B-splines: $c = m + p$
Penalties

• Modify the likelihood function by adding a penalty term over adjacent regression coefficients

\[ S(a;y,\lambda) = (y - Ba)'(y - Ba) + \lambda aP_d a \]

\[ \hat{a} = (B'B + \lambda P_d)^{-1} B'y \]

• The order of the penalty \( d \) controls the changes between adjacent coefficients

\[ P_d = \lambda D_d' D_d \]

\[ D_2 = \begin{pmatrix} 1 & -2 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{pmatrix} \]
Penalties

- Force coefficients to follow a smooth pattern
Advantages of P-splines

- Low-Rank (c<n)
- Computationally efficient (m<=40)
- Selection of number and location of knots is not longer crucial
- Discrete penalties over the regression coefficients
- Easy extension to:
  - Mixed models
  - Non-gaussian data (GLM’s)
  - Multidimensional smoothing
  - Spatial and Spatio-temporal smoothing
Outline

• Introduction
• P-splines
• **Mixed models and P-splines**
• Extensions
• Applications
Reformulation

• Based in represent the smooth function as the sum of a fixed effect and a random effect

\[ y = X\beta + Zu + \varepsilon \quad u \sim N(0, \sigma_u^2 G) \]
\[ \varepsilon \sim N(0, \sigma_e^2 I) \]

• Reparametrization of the original model where we transform the model B-spline basis

\[ Ba = X\beta + Zu \]
\[ X = [1 | x] \]
\[ Z = BU\Sigma^{-1/2} \]
P-splines and mixed models

REML

- Variance estimation
\[-\frac{1}{2} \log |V| - \frac{1}{2} \log |X' V^{-1} X| - \frac{1}{2} y' \left( V^{-1} - V^{-1} X (X' V^{-1} X)^{-1} X' V^{-1} \right) y \]

\[ V = \sigma_u^2 ZZ' + \sigma_e^2 I \]

- Parameters estimation
\[ \hat{\beta} = (X' V^{-1} X)^{-1} X' V^{-1} y \]
\[ \hat{u} = \sigma_u^2 Z' \hat{V}^{-1} \left( y - X \hat{\beta} \right) \]

- The smooth parameter becomes the ratio between the variance of the residual and the variance of the random effects
\[ \lambda = \frac{\sigma_e^2}{\sigma_u^2} \]
Advantages

• Flexibility

• Mixed model theory

\[ \hat{f}(x) = X\hat{\beta} + Z\hat{u} \]

• Software implementation
  ✓ proc mixed de SAS
  ✓ lme() de S-Plus y R

• Extensions
Outline

• Introduction
• P-splines
• Mixed models and P-splines
• Extensions
• Applications
Non-Gaussian data

- Generalized linear models (GLMs)
- Basic structure
  \[ \eta = g(\mu) \quad \eta = E[\mu] \]
- Penalized Generalized linear models (P-GLMs)
  \[ \eta = Ba \]
  ✔ Penalized log-likelihood
  \[ l_p(a) = l(a) - \frac{1}{2} \lambda' Pa \]
- Penalized Generalized linear mixed models (P-GLMMs)
Multidimensional smoothing

- Two-dimensional case (e.g. spatial data)

\[ y = f(x_1, x_2) + \varepsilon \]

✓ Basis

\[ B = B_2 \boxtimes B_1 = (B_2 \otimes 1'_c) \circ (1'_c \otimes B_1) \]

✓ Mixed model matrices

\[ X = \begin{pmatrix} 1_n \mid x_1 \boxtimes 1_n \mid 1_n \boxtimes x_2 \mid x_1 \boxtimes x_2 \end{pmatrix} \]

\[ Z = \begin{pmatrix} Z_2 \boxtimes 1_n \mid Z_2 \boxtimes x_1 \mid 1_n \boxtimes Z_1 \mid Z_2 \boxtimes Z_1 \end{pmatrix} \]
Smoothing additive models

• Two-regressors case

\[ y = f(x_1) + f(x_2) + \varepsilon \]

✓ Basis

\[ B = [B_1 | B_2] \]

✓ Mixed model matrices

\[ X = [1_n | x_1 | x_2] \]
\[ Z = (Z_1 | Z_2) \]
Smoothing additive models

- Two-regressors case with interaction
  \[ y = f(x_1) + f(x_2) + f_{1,2}(x_1, x_2) + \varepsilon \]

  ✓ Basis and penalties
  \[
  B = \begin{bmatrix} B_1 & B_2 & B_{[1,2]} \end{bmatrix} \quad B_{[1,2]} = \left( B_1 \otimes 1_{c2} \right) \circ \left( 1_{c1} \otimes B_2 \right)
  \]

  ✓ Mixed model matrices
  \[
  X = \begin{pmatrix} 1_n \mid x_1 \boxdot 1_n \mid 1_n \boxdot x_2 \mid x_1 \boxdot x_2 \end{pmatrix} \\
  Z = \begin{pmatrix} Z_2 \boxdot 1_n \mid Z_2 \boxdot x_1 \mid 1_n \boxdot Z_1 \mid Z_2 \boxdot Z_1 \end{pmatrix}
  \]

Lee et al., (2013)
Outline

• Introduction
• P-splines
• Mixed models and P-splines
• Extensions
• Applications
NWFP modelling

• Data from NWFP
  ✓ Longitudinal
  ✓ Multidimensional: spatial, spatio-temporal
  ✓ Heterogeneous (biodiversity, climate change)
Longitudinal data

• More flexible models that allow to have different curves for each factor (factor by curve interaction)

Source: Jordan et al. 2008
Multidimensional data

Source: Ignacio Barbeito
Multidimensional data

Source: Ignacio Barbeito
Heterogeneous data

- Allow to use different kind of data with a suitable function

Source: Martinez et al. 2005

Source: Guan et al. 2006
Any question?