Additive Mixed Models with P-Splines

Sonja Greven[1,2], Helmut Küchenhoff[1] and Annette Peters[2] for the AIRGENE study group

[1] LMU, Department of Statistics, Akademiestr. 1, D-80799 Munich, Germany
[2] GSF National Research Center for Environment and Health, Institute of Epidemiology, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany

Presenting author: Sonja Greven, sonja.greven@stat.uni-muenchen.de
Student Oral Presentation

1 Overview

We present an approach to additive mixed models using P-Splines where the spline coefficients are estimated in the mixed model framework. We have implemented this approach in a SAS macro which can be used in a wide variety of contexts. We apply our methodology to an epidemiological study assessing the health effects of ambient air pollution.

2 Motivating data set - the AIRGENE Study

Background: Epidemiological studies clearly link ambient air pollution, especially particulate matter (PM), to morbidity and mortality due to cardio-pulmonary diseases. This lead to new regulations being introduced in the EU in 2005, limiting average PM_{10} ($\varnothing < 10\,\mu m$) concentrations in cities. However, research on causal pathways linking air pollution to outcomes such as myocardial infarctions is still ongoing. Pro-inflammatory and pro-thrombotic processes are thought to be involved. There might also be genetic dispositions for susceptibility to air pollutants.
The study: The AIRGENE study is an EU-funded epidemiological study conducted in the six European cities Athens, Augsburg, Barcelona, Helsinki, Rome and Stockholm between May 03 and July 04. It aims at assessing inflammatory responses in association with ambient air pollution concentrations in myocardial infarction (MI) survivors (see also Ruckerl et al., 2005) and at defining susceptible subgroups of MI survivors based on genotyping.
Data structure: Three inflammatory blood markers (C-reactive protein, Fibrinogen and Interleukin-6) were measured every month repeatedly up to 8 times in over 1,000 MI survivors, resulting in about 6,000 samples

per marker in total. Ten air pollution and several weather variables were measured hourly throughout the study period. Patient characteristics were collected at baseline, including the determination of 114 SNPs on 13 inflammatory pathway genes hypothesized to modify the pollutant effects.

3 Methods

To analyze the AIRGENE data, we have to account for the longitudinal data structure and the potential non-linearity of weather and trend variables. We decided on a mixed model approach using a random patient effect. Short half-times of the blood markers render an additional correlation structure unnecessary. Trend and weather variables are potentially included as additive terms modeled by P-Splines, where coefficients are estimated in the mixed model framework. To assess the possible non-linearity of pollutant effects, dose-response-functions are also modeled as smooth functions in a second step.

The models for the AIRGENE data can be embedded into a more general framework of additive mixed models, where the additive components are modeled by P-Splines (penalized splines with a B-Spline basis, see Eilers and Marx, 1996), and the spline coefficients are estimated in a mixed model framework.
We use P-Splines rather than the truncated power basis usually used in this context, as in Ruppert, Wand and Carroll, 2003, or Ngo and Wand, 2004. This not only results in better numerical properties, but also allows us to independently choose the degree of the B-Splines and the order of the penalization, contrary to them being linked in the truncated powers approach. For the AIRGENE data we can thus use cubic B-Splines for smooth curves with 2nd order difference penalties, penalizing deviations from linearity as the natural default assumption.
We will use notation relating to our longitudinal data, but the extension to the more general additive mixed model case is obvious. A typical model for y_{ij}, the j^{th} blood marker value of the i^{th} patient, would be

$$y_{ij} = u_i + \sum_{l=1}^{w} x_{ijl}\beta_l + \sum_{k=1}^{t} f_k(s_{ijk}) + \varepsilon_{ij} \qquad \text{with} \tag{1}$$

- $u = (u_1, \ldots, u_n) \sim N(\mathbf{0}, \sigma_u^2 I_n)$, the random person effects
- $x_{ij1}, \ldots, x_{ijw}$ values of the linear effect variables $x_1, \ldots, x_w$ for the ij^{th} observation
- $\varepsilon = (\varepsilon_{11}, \ldots, \varepsilon_{1n_1}, \ldots, \varepsilon_{n1}, \ldots, \varepsilon_{nn_n}) \sim N(\mathbf{0}, \sigma_\varepsilon^2 I_N)$, the error terms

- $f_1, \ldots, f_t$ smooth functions of continuous variables with

$$f_k(.) = \sum_{\nu=1}^{K(k)} \gamma_{k\nu} B^d_{k\nu}(.), \tag{2}$$

where the $B^d_{k\nu}(.)$ are $K(k)$ B-splines of degree d.

Omitting the u_i for the moment, the penalized least squares problem for model (1) can then be written as

$$\min \quad \|y - X\beta - \sum_{k=1}^{t} B_k\gamma_k\|^2 + \sum_{k=1}^{t} \lambda_k \gamma_k' {D^p_{K(k)}}' D^p_{K(k)} \gamma_k \qquad \text{with} \tag{3}$$

- D^p_h the p^{th} order difference matrix of size $(h-p) \times h$:

$D^0_h = I_h$, $D^1_h = \begin{bmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}$ and $D^p_h = D^1_{h-p+1} D^{p-1}_h$, $p > 1$.

- λ_k, $k = 1, \ldots, t$, smoothing parameters
- $y = (y_{ij})$, $X = (x_{ijl})$, $\beta = (\beta_l)$, $B_k = (b_{ij\nu}) = (B^d_{k\nu}(s_{ijk}))$, $\gamma_k = (\gamma_{k\nu})$ with ij ordered as in ε, $l = 1, \ldots, w$, $\nu = 1, \ldots, K(k)$, and $k = 1, \ldots, t$.

We can now split the γs into an unpenalized and a penalized part (see Fahrmeir, Kneib and Lang, 2004, for this idea for Bayesian P-Splines)

$$\gamma_k = \Psi^{p,unp}_k \gamma^{p,unp}_k + \Psi^{p,pen}_k \gamma^{p,pen}_k \qquad \text{with} \tag{4}$$

- $\Psi^{p,unp}_k = E^p_k = (e^0_k | \ldots | e^{p-1}_k)$ with $e^l_k = (1^l, \ldots, K(k)^l)'$, the columns of E^p_k spanning the kernel of $D^p_{K(k)}$.
- $\Psi^{p,pen}_k = {D^p_{K(k)}}' (D^p_{K(k)} {D^p_{K(k)}}')^{-1}$.

The penalty terms then reduce nicely and and our problem can be rewritten as

$$\min \quad \|y - \tilde{X}\tilde{\beta} - \tilde{Z}\tilde{\gamma}\|^2 + \sum_{k=1}^{t} \lambda_k \gamma^{p,pen\prime}_k I_{K(k)} \gamma^{p,pen}_k \quad \text{with} \tag{5}$$

- $\tilde{X} = (X | B_1\Psi^{p,unp}_1 | \ldots | B_s\Psi^{p,unp}_s)$, $\tilde{\beta} = (\beta' | \gamma^{p,unp\prime}_1 | \ldots | \gamma^{p,unp\prime}_s)'$,
- $\tilde{Z} = (B_1\Psi^{p,pen}_1 | \ldots | B_s\Psi^{p,pen}_s)$, $\tilde{\gamma} = (\gamma^{p,pen\prime}_1 | \ldots | \gamma^{p,pen\prime}_s)'$.

Divided by σ^2_ε, (5) is equivalent to BLUP-estimation of $\tilde{\beta}$ and $\tilde{\gamma}$ in the mixed model

$$y = \tilde{X}\tilde{\beta} + \tilde{Z}\tilde{\gamma} + \varepsilon \tag{6}$$

with fixed effects $\tilde{\beta}$ and random effects $\tilde{\gamma}$ with a block diagonal covariance matrix with fixed variances $\sigma^2_{\gamma_k} = \sigma^2_\varepsilon / \lambda_k$. We can now easily re-include the other random effects in the model, appending $\tilde{\gamma}$ and $\tilde{Z}$ accordingly.

We use the approach of Ruppert, Wand and Carroll, 2003, to estimate our model (1) as the mixed model (6), including the estimation of the smoothing parameters λ_k as variance ratios $\sigma^2_\varepsilon/\sigma^2_{\gamma_k}$.
Centered plots for f_k can be constructed setting up the design matrices for a grid of s_k, the mean values of the other continuous variables and the reference category of the categorical variables. After reparametrizing, resulting in the matrices $\tilde{X}_{f_k}$ and $\tilde{Z}_{f_k}$, we can estimate the vector of BLUPs for the grid as

$$\hat{f}_k = \tilde{X}_{f_k}\hat{\beta} + \tilde{Z}_{f_k}\hat{\gamma}. \tag{7}$$

where $\hat{\beta}$ and $\hat{\gamma}$ are the estimated BLUPs for $\tilde{\beta}$ and $\tilde{\gamma}$.
Variability bands for the linear and smooth components can be computed using

$$C := \mathrm{Cov}\left(\begin{bmatrix} \hat{\beta} \\ \hat{\gamma} - \tilde{\gamma} \end{bmatrix}\right) = \sigma^2_\varepsilon \begin{bmatrix} \tilde{X}'\tilde{X} & \tilde{X}'\tilde{Z} \\ \tilde{Z}'\tilde{X} & \tilde{Z}'\tilde{Z} + F \end{bmatrix}^{-1} \tag{8}$$

where $F = \mathrm{blockdiag}(\frac{\sigma^2_\varepsilon}{\sigma^2_u} I_n, \frac{\sigma^2_\varepsilon}{\sigma^2_{\gamma_1}} I_{K(1)}, \ldots, \frac{\sigma^2_\varepsilon}{\sigma^2_{\gamma_t}} I_{K(t)})$ (see Ruppert, Wand and Carroll, 2003, for a derivation). An approximate $100(1-\alpha)\%$ confidence interval for the centered f_k at a specific point t_k can then for large numbers of observations be calculated as

$$\begin{aligned} & \hat{f}_k(t_k) \pm z_{1-\frac{\alpha}{2}} \widehat{\mathrm{std}}\left(\hat{f}_k(t_k) - f_k(t_k)\right) \\ = \quad & \hat{f}_k(t_k) \pm z_{1-\frac{\alpha}{2}} \sqrt{l_{t_k} \widehat{C}\, l'_{t_k}} \end{aligned} \tag{9}$$

where l_{t_k} is the corresponding row in $[\tilde{X}_{f_k}|\tilde{Z}_{f_k}]$ and $\hat{C}$ is constructed using the estimated variances. Partial residuals can be added to the plots as an additional diagnostic tool by summing the residuals and the componentwise fitted values.

4 Implementation

To our knowledge, this approach is not implemented in standard software as yet, although the implementation is described in Ngo and Wand, 2004, using a truncated lines basis. In our SAS macro, random intercepts, smooth, linear and categorical components can be named and the degree of the B-Splines, the order of the differences for the penalization and the number of knots can be chosen. Plots of the smooth components with variability bands and partial residuals as well as tests of the linear and categorical covariates as implemented in SAS proc mixed are available. The macro could potentially be used in a wide variety of contexts and can be obtained from the presenting author. The approach showed good results in the simulations we conducted.

TABLE 1. Effect estimates for log(IL-6) [pg/ml] in Stockholm.

Variable		Estimate	Std	p-Value
Intercept		0.1904	0.4304	0.6588
log(BNP)	[pg/ml]	0.1336	0.0351	0.0001
BMI	[kg/m^2]	0.0241	0.0085	0.0047
HDL	[mg/dl]	-0.0064	0.0027	0.0188
Temperature	[°C]	-0.0066	0.0051	0.1971
COPD/	no	-0.3045	0.1516	
chronic	some indication	-0.0996	0.1638	
bronchitis	yes	0		0.0177
Reinfarction	no	-0.1557	0.1063	
	yes	0		0.1434

5 Results

We decided on seperate models for each blood marker and city, as climate and study period differ considerably. We built confounder models first without air pollutants to allow for meaningful tests of pollutant effects, and then added one pollutant at a time to avoid collinearity, testing for a linear effect. Results were pooled subsequently using meta-analysis methodology.
The air pollution results of the AIRGENE study will be presented elsewhere. As an example for the analyses conducted, we here present the selected confounder model for one of the blood markers and one of the cities - for the log-transformed Interleukin-6 in Stockholm. This model was built in a forward step-wise procedure using the AIC to compare models. We selected chronic obstructive pulmonary disease / chronic bronchitis and reinfarction indicators (categorical), log(BNP) (a heart failure blood marker), body mass index (BMI), high density cholesterol (HDL) and average apparent temperature in the last 48 hours (linear), and time trend (smooth) as potential confounders. Results for the effect estimates of the linear and categorical predictors are shown in table 1. Higher values of BMI and BNP, lower values of HDL (the "good" cholesterol) and temperature as well as having COPD / chronic bronchitis or a reinfarction correspond to higher IL-6 values. Time trend was measured in days since start of the study, corresponding in Stockholm to September 03, 2003, to June 24, 2004. Figure 1 shows the estimated smooth time trend with 95% variability bands and partial residuals, estimated using cubic B-Splines with second order difference penalties.

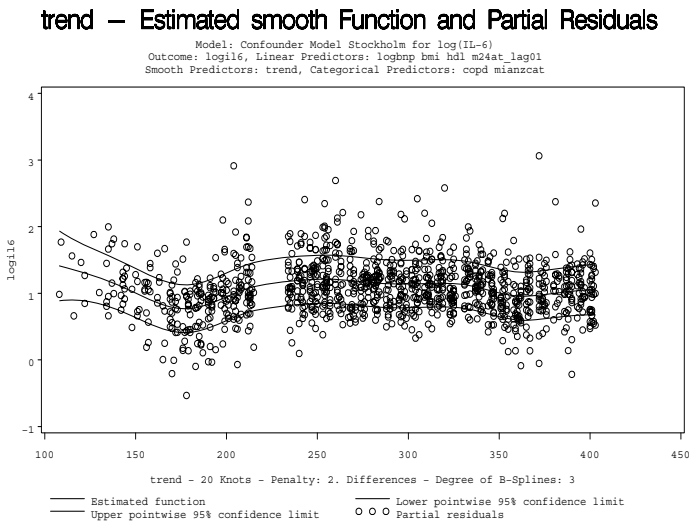

FIGURE 1. Estimated smooth time trend of log(IL-6) in Stockholm.

6 Summary and Outlook

We have shown that additive mixed models with penalized splines estimated in the mixed model framework are extendable to P-splines using an additional reparametrization. This allows an independent choice of the B-Spline degree and the order of the penalty. We implemented this approach in a SAS macro suitable for many applications and used it to analyze the longitudinal AIRGENE study.

We will next focus on testing of smooth components and pooling of smooth components across cities, which is motivated by estimation and testing of dose-response-functions for environmental factors in AIRGENE. A bootstrap test analogously to the truncated lines case in Coull, Schwartz, and Wand, 2001, will be implemented first. Afterwards we plan an extension of the exact likelihood ratio tests for penalized splines developed by Crainiceanu et al., 2005, using the truncated power basis, and a comparison of the two tests.

An extension of our approach to generalized models is also in progress, to allow models for additional diary data in AIRGENE where presence or absence of symptoms as well as health status in five categories was noted down daily. All extensions will be implemented and made available in SAS.

Acknowledgments: The AIRGENE study is funded by the EU contract QLK4-CT-2002-O2236. The study group comprises the following partners: 1. GSF, EPI (Neuherberg, Germany): A. Peters (PI), S. Greven, R. Rueckerl, A. Schneider, S. von Klot, T. Illig, M. Kolz, I. Brueske-Hohlfeld, A. Ibald-Mulli, A. Schaffrath

Rosario, H.E. Wichmann; 2. Univ. of Ulm Cardiology (Germany): W. Koenig; 3. ASL RM E, Dep. of Epi. (Rome, Italy): F. Forastiere (PI), S. Picciotto, M. Stafoggia, R. Pistelli (Catholic Univ.); 4. KTL (Kuopio & Helsinki, Finland): J. Pekkanen (PI), T. Lanki, V. Salomaa; 5. KI IMM (Stockholm, Sweden): T. Bellander (PI), N. Berglind, E. Lampa, P. Ljungman, F. Nyberg, G. Pershagen; 6. IMIM (Barcelona, Spain): J. Sunyer (PI), J. Marrugat, B. Jacquemin; 7. Univ. of Athens Medical School, Dep. of Hygiene a. Epi. (Greece): K. Katsouyanni (PI), A. Chalamandaris; 8. UHEL (Helsinki, Finland): M. Kulmala, P. Aalto, P. Paatero

Special thanks to Thomas Kneib and Susanne Breitner for fruitful discussions regarding additive mixed models.

Coull, B.A., Schwartz, J., and Wand, M.P. (2001). Respiratory health and air pollution: additive mixed model analyses. *Biostatistics*, **2**(3), 337-349.

Crainiceanu, C., Ruppert, D., Claeskens, G., and Wand, M.P. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika*, **92**, 91-103.

Eilers, P.H.C., and Marx, B.D. (1996). Flexible Smoothing with B-Splines and Penalties. *Statistical Science*, **11**, 89-121.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, **14**, 731-761.

Ngo, L., and Wand, M.P. (2004). Smoothing with Mixed Model Software. *Journal of Statistical Software*, **9**(1), 1-54.

Ruckerl, R., Ibald-Mulli, A., Koenig, W., Schneider, A., et al. (2006). Air Pollution and Markers of Inflammation and Coagulation in Patients with Coronary Heart Disease. *American Journal of Respiratory and Critical Care Medicine*, **173**(4).

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