Additive Mixed Models for Correlated Functional Data

Fabian Scheipl

Institut für Statistik
Ludwig-Maximilians-Universität München

SuSTaIn Workshop
"High Dimensional and Dependent Functional Data"
September 10, 2012

Joint work with C. Crainiceanu, S. Greven, A. Ivanescu, P. Reiss, A.-M. Staicu
Flexible, multiple regression models with functional predictors for correlated functional responses:

\[ Y_{ij}(t) = \alpha(t) + \eta(t; x_{ij}) + \epsilon_{ij}(t) \]
Flexible, multiple regression models with functional predictors for correlated functional responses:

\[ Y_{ij}(t) = \alpha(t) + \eta(t; x_{ij}) + \epsilon_{ij}(t) \]

white noise errors: \( \epsilon_{ij}(t) \overset{iid}{\sim} \mathcal{N}(0, \sigma^2) \)
Flexible, multiple regression models with functional predictors for correlated functional responses:

\[ Y_{ij}(t) = \alpha(t) + \eta(t; x_{ij}) + \epsilon_{ij}(t) \]

white noise errors: \( \epsilon_{ij}(t) \sim iid N(0, \sigma^2) \)

additive predictor \( \eta(t; x_{ij}) \) is the sum of:

- (nonlinear, multivariate, index-varying) effects of scalar covariates:
  - \( x\beta, x\beta(t), f(x), f(x, t), f(x_1, x_2, t) \)
Flexible, multiple regression models with functional predictors for correlated functional responses:

\[ Y_{ij}(t) = \alpha(t) + \eta(t; x_{ij}) + \epsilon_{ij}(t) \]

white noise errors: \( \epsilon_{ij}(t) \overset{iid}{\sim} N(0, \sigma^2) \)

additive predictor \( \eta(t; x_{ij}) \) is the sum of:

- (nonlinear, multivariate, index-varying) effects of scalar covariates: 
  - \( x\beta, x\beta(t), f(x), f(x, t), f(x_1, x_2, t) \)
- effects of functional covariates \( X(s) \):
  - \( \int X(s)\beta(s, t)ds \)
  - for \( X_i(s) \approx \sum_{m=1}^{M} \phi_m(s)\xi_{m,i} : \sum_{m=1}^{M} \xi_m\beta_m(t), \sum_{m=1}^{M} f_m(\xi_m, t) \)
Flexible, multiple regression models with functional predictors for correlated functional responses:

\[ Y_{ij}(t) = \alpha(t) + \eta(t; x_{ij}) + \epsilon_{ij}(t) \]

white noise errors: \( \epsilon_{ij}(t) \overset{iid}{\sim} N(0, \sigma^2) \)

additive predictor \( \eta(t; x_{ij}) \) is the sum of:

- (nonlinear, multivariate, index-varying) effects of scalar covariates:
  - \( x \beta, x \beta(t), f(x), f(x, t), f(x_1, x_2, t) \)
- effects of functional covariates \( X(s) \):
  - \( \int X(s) \beta(s, t) \, ds \)
  - for \( X_i(s) \approx \sum_{m=1}^{M} \phi_m(s) \xi_{m,i} : \sum_{m=1}^{M} \xi_m \beta_m(t), \sum_{m=1}^{M} f_m(\xi_m, t) \)
- scalar or functional random effects for correlated data:
  - \( b_0i, b_1i u_{ij}, B_{0i}(t), B_{1i}(t)u_{ij} \)
This is just a kind of varying coefficient model...

... effects vary over index of response:

▶ write as **model for concatenated function values** $\text{vec} \left( Y_{n \times T} \right)$

▶ **reformat covariate data** accordingly & add index $t$ as covariate

⇒ can be fitted like **standard AMM for scalar responses** using penalized splines

▶ effects vary smoothly over $t \Rightarrow$ smoothness of $E(Y(t))$

▶ sparse/irregular $Y(t)$ possible
Model representation:

\[ \eta(t; x) = \sum_p f(t, x_p) \]

effects are weighted sums of basis functions \( B(t, x) \) in index of response and covariates:

\[ f(t, x_p) = \sum_d B_d(t, x_p) \theta_d = B_p \theta_p \]
Model representation:

\[ \eta(t; x) = \sum_{p} f(t, x_p) \]

effects are weighted sums of basis functions \( B(t, x) \) in index of response and covariates:

\[ f(t, x_p) = \sum_{d} B_d(t, x_p) \theta_d = B_p \theta_p \]

each associated with penalty term \( \text{pen}(\theta_p | \lambda_p) \) for regularization: enforce smoothness, similarity of correlated units, etc.

\[ \text{pen}(\theta | \lambda) = \theta^T P(\lambda) \theta \]

equivalent to distributional assumption:

\[ \theta \sim N(0, P(\lambda)^{-1}) \Rightarrow f(t, x) \sim GP(0, B P(\lambda)^{-1} B^T) \]
Inference is mostly a solved problem:

- use **penalized splines for smooth effects**, including linear functional effects and functional random effects:

- **standard penalized likelihood inference** with (G)CV or (RE)ML via mixed model representation to solve

\[
\min_{\theta, \lambda} \left( \left\| \text{vec}(Y) - \sum_p B_p \theta_p \right\|^2_2 + \sum_p \text{pen}(\theta_p | \lambda_p) \right)
\]

- approximate CIs, tests, diagnostics, etc. immediately available

⇒ **refund**'s `pffr()` as wrapper for **mgcv**:

- optimized, robust, well-tested algorithms
- versatile library of spline bases ready to use
- model effects with any given correlation structure via GMRFs

Scheipl, Staicu, Greven (2012) Function-on-Function AMMs
Tensor product basis representation of effects:

\[ f(t, \text{covariate}) \approx \left( \begin{bmatrix} B_c \otimes B_t \end{bmatrix} \right) \theta \]

\[ \text{pen}(\theta|\lambda_t, \lambda_c) = \theta^T (\lambda_t I_{K_c} \otimes P_t + \lambda_c P_c \otimes I_{K_t}) \theta \]

- **B_c**: marginal basis for covariate
- **B_t**: marginal basis on response’s index
- **P_t, P_c**: marginal penalty matrices
- **P \otimes I, I \otimes P**: repeated penalties that apply to each subvector of \( \theta \) associated with a specific marginal basis function (Wood, 2006)
- **construction valid for all types of effects we consider**: \( \int X(s)\beta(s, t)ds, \ x\beta(t), \ f(x, t), \text{ etc.} \)
Model

Tensor product basis representation of effects:

\[ f(t, \text{covariate}) \approx \left( \mathbf{B}_c \otimes \mathbf{B}_t \right)_{nT \times 1} \theta_{K_c K_t \times 1} \]

\[ \text{pen}(\theta|\lambda_t, \lambda_c) = \theta^T (\lambda_t \mathbf{I}_{K_c} \otimes \mathbf{P}_t + \lambda_c \mathbf{P}_c \otimes \mathbf{I}_{K_t}) \theta \]

- e.g. scalar random intercepts \( b_0 \sim N(0, \frac{1}{\lambda_c} \mathbf{S}) \):

\[
\begin{align*}
\mathbf{B}_c &= \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}_{n \times \#\text{units}} \\
\mathbf{P}_c &= \mathbf{S}^{-1}; \\
\mathbf{B}_t &= 1; \\
\mathbf{P}_t &= 0
\end{align*}
\]
Functional random effects with between-unit correlation:

\[ \mathbf{B}_0(t) = \left( \begin{array}{c} \mathbf{B}_c \\ \mathbf{B}_t \end{array} \right) = \left( \begin{array}{c} \mathbf{B}_c \otimes \mathbf{B}_t \end{array} \right) \begin{pmatrix} \theta \\ \theta' \end{pmatrix} \begin{pmatrix} \# \text{units} \\ \# \text{units} \times K_t \end{pmatrix} \]

\[ \text{pen}(\theta | \lambda_t, \lambda_c) = \theta^T \left( \lambda_t \mathbf{P}_t \otimes \mathbf{I}_{\# \text{units}} + \lambda_c \mathbf{I}_{K_t} \otimes \mathbf{S}^{-1} \right) \theta \]

- **\( \mathbf{B}_t \)**: spline basis over \( t \) with smoothness penalty \( \mathbf{P}_t \)
- **\( \mathbf{B}_c \)**: matrix of unit level indicators
- **\( \mathbf{S} \)**: correlation of subvectors of coefficients associated with the same marginal basis function in \( \mathbf{B}_t \)
  \[ \Rightarrow \text{induces similarity of } B_{0i}(t), B_{0j}(t) \text{ if } S_{ij} \text{ is large.} \]
- **model effects with any given marginal correlation \( \mathbf{S} \)**, for example:
  - \( \mathbf{S} = \mathbf{I} \) for independent random effects
  - \( \mathbf{S}^{-1} \): precision matrix of a GMRF, e.g. via adjacency matrix of a graph
  - \( S_{ij} = \rho^{|i-j|} \) for AR(1)-structure
Functional random effects with between-unit correlation:

\[
B_0(t) = \begin{pmatrix}
B_c \otimes B_t \\
_{nT \times 1} & _{n \times \# \text{units}} & _{T \times K_t}
\end{pmatrix} \begin{pmatrix}
\theta \\
_{\# \text{units} \cdot K_t \times 1}
\end{pmatrix}
\]

\[
\text{pen}(\theta | \lambda_t, \lambda_c) = \theta^T \left( \lambda_t P_t \otimes I_{\# \text{units}} + \lambda_c I_{K_t} \otimes S^{-1} \right) \theta
\]

⇒ fit correlated functional random effects with arbitrary, fixed inter-unit correlation structure

⇒ fit correlated smooth residual functions with arbitrary, fixed inter-observation correlation structure
Example: Canadian Weather Data
Empirical Evaluation

Canadian Weather Data: Matèrn-correlated smooth residuals

$$\log_{10}(\text{precipitation}_i(t)) = \alpha_{\text{region}(i)}(t) + \int \text{temperature}_i(s) \beta(t, s) ds + E_i(t) + \epsilon_i(t),$$

- region-specific functional intercepts $\alpha_{\text{region}}(t)$
- linear function-on-function effect of temperature profiles
- spatially correlated smooth residuals $E(t)$ with Matèrn-correlation function over station locations to capture deviations from regional, temperature-adjusted means
Canadian Weather Data: Effect Estimates Model

\[ \hat{\alpha}_g(t): \text{regional mean log(Precipitation)} \text{[mm]} \]

\[ \hat{\beta}(s,t) \]

Scheipl, Staicu, Greven (2012) Function-on-Function AMMs
Canadian Weather Data: Effect Estimates

Smooth, correlated residual functions $\hat{E}_i(t)$
Simulation Study Results

- estimation accuracy robust against small number of observations
- estimation accuracy for $E(Y(t))$ good and very stable:
  - relative IMSE stays $< 3\%$ even for $SNR = 1$, 30 observations in 10 groups.
- computation time increases quickly, but still doable for large data
  - 30 observations in 10 groups, 30 gridpoints, two functional random effects: $\approx 5\text{sec}$
  - 2000 observations in 100 groups, 100 gridpoints, two functional random effects: $\approx 12\text{h}$
- REML much more stable than GCV
- not feasible (yet): smooth residuals for large data sets
Open questions & further directions:

- Identifiability conditions for function-on-function terms: compare FPC- and spline-based approaches (Müller and Yao, 2008; Ivanescu et al., 2012; Scheipl et al., 2012)
- More efficient algorithms:
  - Array models (Currie et al., 2006)
  - Exploit sparse designs & penalties
- Non-i.i.d. errors:
  - GLS-type approaches: $\epsilon_{ij}(t) \overset{iid}{\sim} GP(0, K(t, t'))$ with known/fixed covariance $K(t, t')$ (Reiss et al., 2010)
  - Bootstrap-based inference
- Bayes inference: better uncertainty measures, variable selection, robustification (Morris and Carroll, 2006; Zhu et al., 2011; Goldsmith et al., 2011)
- Non-Gaussian data: binary data, count data, hazard rate models
Summary

In terms of **flexible, multiple** regression models for **correlated** functional responses and predictors, *refund’s* `pffr()` can already do (almost) everything that GAMMs can do for scalar responses.

**But:** Implementation more advanced than theoretical results and empirical evaluation

- framework pushes limits of underlying inference algorithms
  - validate with more case studies
- do asymptotics & robustness results carry over from scalar case?

Find the papers & supplementary material with code examples on my homepage:

  http://www.statistik.lmu.de/~scheipl/research.html
References


\( P \otimes I, I \otimes P \): repeated penalties that apply to each subvector of \( \theta \) associated with a specific marginal basis function (Wood, 2006):

\[
B_c \otimes B_t = \left( \begin{array}{c}
\perp_1 \nabla_1 \times_1 \\
\perp_2 \nabla_2 \times_2 \\
\perp_3 \nabla_3 \times_3 \\
\perp_4 \nabla_4 \times_4 \\
\end{array} \right) \otimes \left( \begin{array}{c}
\times_1 \star_1 \\
\times_2 \star_2 \\
\times_3 \star_3 \\
\end{array} \right) = \\
\left( \begin{array}{cccc}
\perp_1 \times_1 & \perp_1 \star_1 & \nabla_1 \times_1 & \nabla_1 \star_1 \\
\perp_1 \times_2 & \perp_1 \star_2 & \nabla_1 \times_2 & \nabla_1 \star_2 \\
\perp_1 \times_3 & \perp_1 \star_3 & \nabla_1 \times_3 & \nabla_1 \star_3 \\
\end{array} \right)
\]

\[
I_{Kc} \otimes P_t = \left( \begin{array}{c}
1 \\
1 \\
1 \\
\end{array} \right) \otimes \left( \begin{array}{c}
P_x \times \star \times \\
P_x \times \star \times \\
P_x \times \star \times \\
\end{array} \right) = \\
\left( \begin{array}{cccc}
P_x & P_x \star \times & P_x \times \star & P_x \times \star \times \\
P_x & P_x \star \times & P_x \times \star & P_x \times \star \times \\
P_x & P_x \star \times & P_x \times \star & P_x \times \star \times \\
\end{array} \right)
\]

\[
P_c \otimes I_{Kt} = \left( \begin{array}{c}
P_{\perp} & P_{\nabla \perp} & P_{\times \perp} \\
P_{\nabla \perp} & P_{\nabla \times} & P_{\times \nabla} \\
P_{\times \perp} & P_{\times \nabla} & P_{\times \times} \\
\end{array} \right) \otimes \left( \begin{array}{c}
1 \\
1 \\
1 \\
\end{array} \right) = \\
\left( \begin{array}{cccc}
P_{\perp} & P_{\perp \perp} & P_{\perp \times} & P_{\perp \times \times} \\
P_{\nabla \perp} & P_{\nabla \perp} & P_{\nabla \times} & P_{\nabla \times \times} \\
P_{\times \perp} & P_{\times \perp} & P_{\times \times} & P_{\times \times \times} \\
\end{array} \right)
\]
Canadian Weather Data: Effect Estimates without Smooth Residuals

\[ \hat{\alpha}_g(t) : \text{regional mean log(Precipitation) [mm]} \]

\[ \hat{\beta} \]

-0.010
-0.005
0.000
0.005
0.010
0.015

2 4 6 8 10 12
-0.4 −0.2 0.0 0.2 0.4

Months
Canadian Weather Data: Spatio-temporal smooth effect + uncorrelated scalar random effects

\[
\log_{10}(\text{precipitation}_i(t)) = \alpha(t) + b_{0i} + \\
\int \text{temperature}_i(s) \beta(t, s) ds + f(\text{latitude}_i, \text{longitude}_i, t) + \epsilon_i(t),
\]

- uncorrelated scalar random effects \(b_0\) to capture small-scale spatial variation in precipitation levels
- linear function-on-function effect of temperature profiles
- smooth spatio-temporal effect \(f(\text{latitude}, \text{longitude}, t)\) to capture large-scale spatial variation over time

R-Code:

```r
m2 <- pffr(l10precip ~ c(s(place, bs="re")) + 
  ff(temp, yind=month.t, xind=month.s, splinepars=list(bs=c("cc", "cc")) + 
  s(lat,lon),
  bs.int = list(bs="cc", k=10), bs.yindex = list(bs="cc"), data=dataM, yind = month.t)
```
Canadian Weather Data: Effect Estimates

\[ \hat{f}(\text{latitude, longitude, } t), \ t = 1 \]
Issues

Flexibility - Overfitting - Concurvity

very flexible models $\rightarrow$ danger of overfitting
$\rightarrow$ overparameterization? identifiability?

functional concurvity $\rightarrow$ useful analogies to scalar concurvity?

no good metrics/diagnostics for these issues yet
Identifiability for function-on-function terms

Functional Analysis:
“Coefficient surfaces $\beta(s, t)$ are not identifiable outside the span of the eigenfunctions of the cross-covariance of $Y(t)$ and $X(s)$.”

- **FPCA approach:**
  “Restrict $\hat{\beta}(s, t)$ to span(eigenfunctions).”

- **penalized splines approach:**
  “Not actually a problem for most finite-sample data, get smoothest $\hat{\beta}(s, t)$ that yields a good fit. Most problematic settings can be identified and circumvented.” (Scheipl and Greven, 2012)

But:
Can we really interpret $\hat{\beta}(s, t)$ meaningfully?
Identifiability of function-on-function terms

In theory, coefficient surfaces $\beta(s, t)$ are not identifiable except in the span of the covariate’s eigenfunctions — for $X_i(s) = \sum_{m=1}^{M} \phi_m(s)\xi_{m,i}$:

$$\int X_i(s)\beta(s, t)ds = \sum_{m=1}^{M} \xi_{im} \int \phi_m(s)\beta(s, t)ds$$

$$= \sum_{m=1}^{M} \xi_{im} \int \phi_m(s)(\beta(s, t) + o(s))ds$$

$\forall o(s) \notin \text{span}(\{\phi_m(s) : m = 1, \ldots, M\})$ as $\int \phi_m(s)o(s)ds = 0$.

$\Rightarrow$ adding any $o(s)$ from kernel of $X$’s covariance does not change fit

(Cardot et al., 1999; He et al., 2003)
Identifiability of function-on-function terms

In theory, coefficient surfaces $\beta(s, t)$ are not identifiable except in the span of the covariate’s eigenfunctions — for $X_i(s) = \sum_{m=1}^{M} \phi_m(s)\xi_{m,i}$:

$$
\int X_i(s)\beta(s, t)ds = \sum_{m=1}^{M} \xi_{im} \int \phi_m(s)\beta(s, t)ds
$$

$$
= \sum_{m=1}^{M} \xi_{im} \int \phi_m(s)(\beta(s, t) + o(s))ds
$$

$\forall$ $o(s) \not\in \text{span}(\{\phi_m(s) : m = 1, \ldots, M\})$ as $\int \phi_m(s)o(s)ds = 0$.

$\Rightarrow$ adding any $o(s)$ from kernel of $X$’s covariance does not change fit

(Cardot et al., 1999; He et al., 2003)

$\Rightarrow$ in practice: if

$\Rightarrow$ $K_s > M$ and

$\Rightarrow$ kernel of $X(s)$-covariance overlaps penalty nullspace

$\Rightarrow$ adding $o(s)$ does not change penalty term either $\Rightarrow$ **not identifiable**

(Scheipl and Greven, 2012)
Identifiability: Example I

\[ E(Y(t)) \]

\[ \hat{Y}(t) \]

\[ \hat{Y}(t) \]

\[ \tilde{Y}(t) \]

\[ \tilde{Y}(t) \]
Identifiability: Example II

\( \hat{\beta}(s, t) \) for Canadian Weather:

\[(K_s, K_t) = (3, 3) \quad (K_s, K_t) = (8, 8) \quad (K_s, K_t) = (11, 11) \]

Fitted values & remaining term estimates are (almost) exactly the same!

Scheipl, Staicu, Greven (2012) Function-on-Function AMMs