Functional Linear Mixed Models

for Sparsely or Irregularly Sampled Data

Jona Cederbaum[1] and Sonja Greven[1]

[1] *Department of Statistics, LMU Munich, Germany*

`Jona.Cederbaum@stat.uni-muenchen.de`

Abstract

We propose a non-parametric statistical method to analyze sparsely or irregularly sampled functional data that involve an additional correlation structure. Sources of correlation may be very general, such as repeated measurements, grouping in the data, or crossed designs.

We extend the Functional Linear Mixed Model for longitudinal functional data of Greven, Crainiceanu, Caffo, Reich, Electronic Journal of Statistics, 2010 to more general correlated functional data which are not sampled on a fine grid and for which only a small number of measurements may be available per curve.

Estimation is based on dimension reduction via Functional Principal Component Analysis (FPCA). Our procedure allows the decomposition of the variability in the data as well as the estimation of main effects of interest. We apply our methods in simulations (not presented here) and an application from speech production research.

The Functional Linear Mixed Model (FLMM)

The Functional Linear Mixed Model can be seen as a functional analogue of the scalar Linear Mixed Model (Laird and Ware, Biometrics, 1982) in such a way that random effects are replaced by random processes. The unit of observation is a curve.

The General Functional Linear Mixed Model (GFLMM)

$$Y_i(d) = \mu(d, \boldsymbol{x}_i) + \boldsymbol{z}_i^T B(d) + E_i(d) + \varepsilon_{id}, \quad i = 1, \ldots, n, \tag{1}$$

- $Y_i(d)$: random function observed at arguments d in some set $\mathcal{D}$
- $\mu(d, \boldsymbol{x}_i)$: fixed main effects surface dependent on known covariates $\boldsymbol{x}_i$
- $B(d)$: random functions
- $\boldsymbol{z}_i$: known covariates
- $E_i(d)$: curve-specific deviations in form of smooth residual curves
- ε_{id}: white noise measurement error with variance $\sigma^2(d)$
- n: number of observed curves

Assumption: $B(d)$, $E_i(d)$, and ε_{id} are assumed independent for all i

Estimation

Mean-, auto-covariance-, and eigenfunctions are assumed to be smooth. Dimension reduction is mandatory for estimation of functional data. We use dimension reduction via FPCA whereby the dominant modes of variation are extracted.

We face some challenges (theoretical and computational) when dealing with irregularly or even sparsely sampled data:

- PC scores cannot be estimated via numerical integration as usually done in FPCA
- smoothing of single curves may be impossible due to few measurement points
- smoothing in general is less accurate in the sparse case than for dense grid-data
- computational problems arise due to large a number of unique sampling points across curves (no Kronecker products can be used)
- implementation is more challenging with different measurement points

Yao, Müller, and Wang, JASA, 2005 propose a method to perform FPCA for sparse independent functional data. We extend this to the case of correlated functional data.

We propose an **estimation algorithm** consisting of four steps which is exemplary described for model (3):

1. Estimation of the fixed main effects function under working independence, i.e.

$$Y_{ijh}(t) = \mu(t, \boldsymbol{x}_{ij}) + \varepsilon_{ijt}$$

- subsequent centering of the data: $\tilde{Y}_{ijh}(t) = Y_{ijh}(t) - \hat{\mu}(t, \boldsymbol{x}_{ij})$

2. Estimation of the auto-covariance functions and $\sigma^2(t)$ using the variance decomposition

$$\begin{aligned} Cov\{\tilde{Y}_{ijh}(t), \tilde{Y}_{i'j'h'}(t')\} &= Cov\{B_i(t), B_i(t')\}\delta_{ii'} + Cov\{C_j(t), C_j(t')\}\delta_{jj'} \\ &+ \left[Cov\{E_{ijh}(t), E_{ijh}(t')\} + \sigma^2(t)\delta_{tt'}\right]\delta_{ii'}\delta_{jj'}\delta_{hh'} \end{aligned}$$

with $\delta_{ii'} = \begin{cases} 1, & \text{if } i = i' \\ 0, & \text{otherwise} \end{cases}$ by bivariate penalized splines.

- subsequent evaluation on a fine regular grid for eigendecomposition
- strength is borrowed across curves (important in sparse setting)

3. Expansions of $B_i(t)$, $C_j(t)$, and $E_{ijh}(t)$ in truncated bases of eigenfunctions of the auto-covariance functions

▷ bases estimated from the data
▷ results in Linear Mixed Model (random effects = PC scores)

$$Y_{ijh}(t) = \hat{\mu}(t, \boldsymbol{x}_{ij}) + \underbrace{\sum_{k=1}^{N_B} \xi_{ik}^B \phi_k^B(t)}_{B_i(t)} + \underbrace{\sum_{k=1}^{N_C} \xi_{jk}^C \phi_k^C(t)}_{C_j(t)} + \underbrace{\sum_{k=1}^{N_E} \xi_{ijhk}^E \phi_k^E(t)}_{E_{ijh}(t)} + \varepsilon_{ijht}$$

- ξ_{ik}^B, ξ_{jk}^C, and ξ_{ijhk}^E: uncorrelated random variables with zero mean and variances corresponding to the ordered eigenvalues in the decomposition
- $\phi_k^B(t)$, $\phi_k^C(t)$, and $\phi_k^E(t)$: corresponding eigenfunctions

4. Estimation of the PC scores as BLUPs of the Linear Mixed Model

▷ no need to fit Linear Mixed Model
▷ computationally highly efficient

Data Application

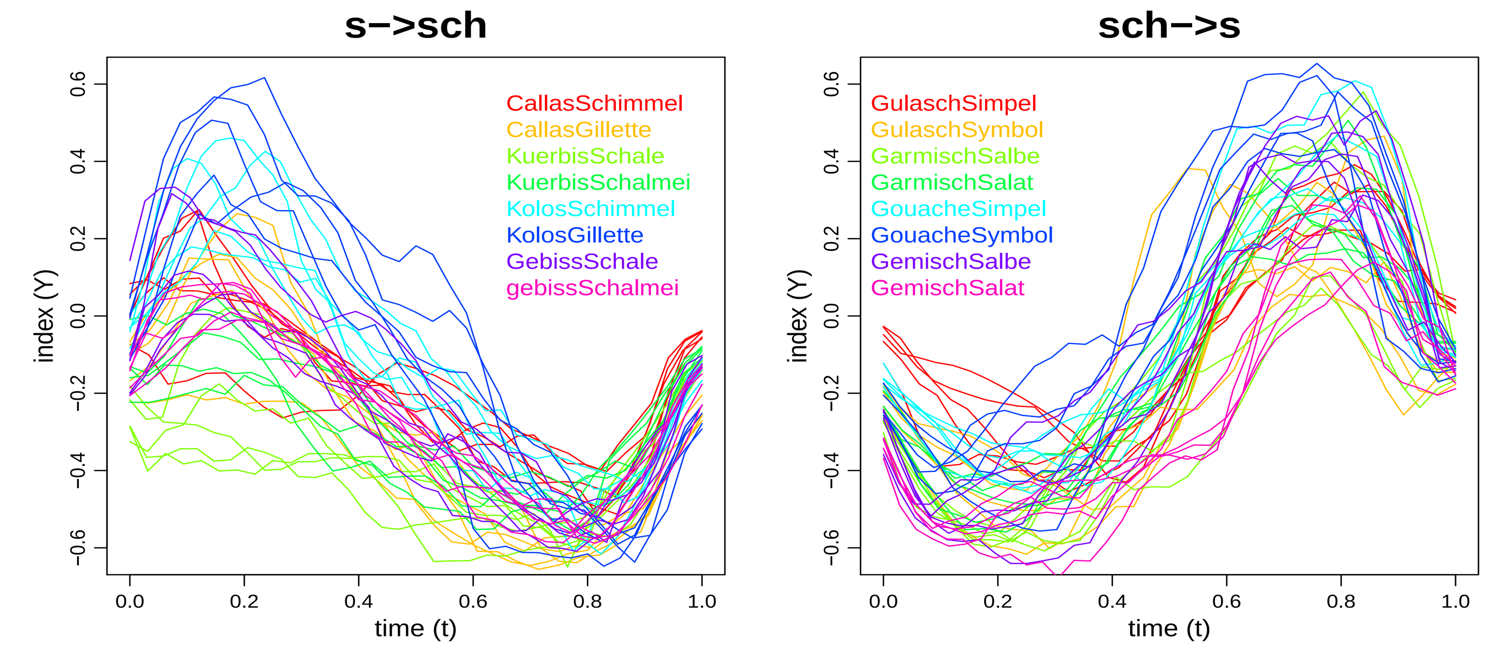

Figure 1: *Index development for one subject repeating 16 compound words. In each plot, the curves belonging to one compound word are the same color. Index values near 1 stand for sound "s" and values near -1 for sound "sch".*

Speech production researchers are interested in the change of articulation when the sounds "s" and "sch" follow each other.

- 32 different fictive compound words of the 2 word groups (s→sch, sch→s) are read out loud by 9 subjects while their tongue movement is summarized in a one-dimensional index (Y) over time
- reading durations of the sounds of interest are extracted based on acoustic signals
- remaining variability in t is inseparable from pronunciation effects of interest
- standardization of the different reading durations results in irregularly spaced measurements of the index between curves
- each compound word is read out five times by each subject (some missings)

→ correlated measurements both for each compound word and for each subject

We first consider an FLMM with a random intercept for subjects. This is a special case of model (1) in which the correlation between measurements at the same subject is taken into account. Separate models are fit for each word group.

Functional Random Intercept Model

$$Y_{ij}(t) = \mu(t) + B_i(t) + E_{ij}(t) + \varepsilon_{ijt}, \tag{2}$$

- $Y_{ij}(t)$: index over time for curve j of subject i
- $\mu(t)$: main fixed effect
- $B_i(t)$: random functional intercept for subject i
- $E_{ij}(t)$: subject- and curve-specific smooth random deviation
- ε_{ijt}: white noise measurement error

Results of Random Intercept Model: Results are exemplary shown for word group sch→s.

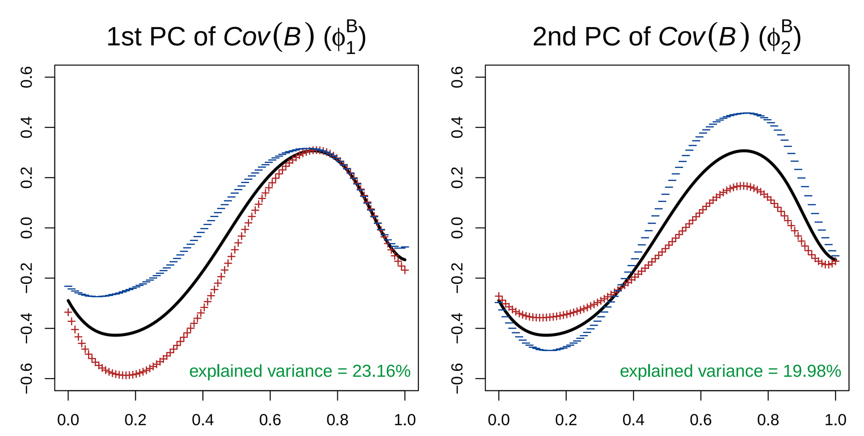

Figure 2: *First two PC curves for the random intercept. Shown: Mean curves and effect of adding (+) and of subtracting (-) a multiple of PC curve.*

Variance decomposition: For a pre-specified level L= 99% of explained average variance, we obtain

- 4 PCs for B explaining 45.85% of the variance
- 7 PCs for U explaining 49.27% of the variance
- $\hat{\sigma}^2 = 0.0027$ which is 3.89% of the variance

Interpretation: Subjects with pos. score values

- for the 1st PC pronounce sound "sch" stronger than subjects with neg. score values
- for the 2nd PC differentiate less between the two sounds and especially have a weaker pronunciation of sound "s"

We are currently extending model (2) by accounting for correlation between measurements of the same compound word leading to an FLMM with crossed design. Furthermore, we include the effect of a factor variable with 4 levels that indicates which syllables of the compound word j are stressed.

Functional Linear Mixed Model with Crossed Design

$$Y_{ijh}(t) = \mu(t, x_j) + B_i(t) + C_j(t) + E_{ijh}(t) + \varepsilon_{ijht}, \tag{3}$$

- $Y_{ijh}(t)$: index over time for h's repetition of word j by subject i
- $\mu(t, x_j)$: main fixed effect dependent on covariate "stress":

$$\mu(t, x_j) = \beta_0 + \sum_{g=1}^{4} f_g(t)\delta_{jg}, \; with \; \delta_{jg} = \begin{cases} 1, & if \; x_j = g \\ 0, & otherwise \end{cases}$$

- $B_i(t)$ and $C_j(t)$: random functional intercepts for subject i / for compound word j
- $E_{ijh}(t)$: subject-, compound word-, and repetition-specific smooth random deviation
- ε_{ijht}: white noise measurement error