Multilevel Functional Principal Component Analysis for Unbalanced Data

Zuzana Rošťáková*[1] and Roman Rosipal[1]

[1]*Institute of Measurement Science, Slovak Academy of Sciences, Slovakia*

Functional principal component analysis (FPCA) is the key technique for dimensionality reduction and detection of main directions of variability present in functional data. However, it is not the most suitable tool for the situation when analysed dataset contains repeated or multiple observations, because information about repeatability of measurements is not taken into account. Multilevel functional principal component analysis (MFPCA) is the modified version of FPCA developed for data observed at multiple visits. The original MFPCA method was designed for balanced data only, where for each subject the same number of measurements is available. In this article we propose the modified MFPCA algorithm which can be applied for unbalanced functional data. The modified algorithm is validated and tested on real–world sleep data.

Keywords: multilevel functional principal component analysis, functional data with multiple observations, sleep probabilistic curves

Introduction

Functional principal component analysis (FPCA) is an appropriate tool for detecting main directions of variability and dimensionality reduction of functional data [1]. On the other hand, FPCA considers each curve as a single observation and therefore it is not appropriate for detecting sources of variability in datasets with multiple observations. These multiple observations can be represented by repeated collection of data at multiple visits.

To address this repeated observations data design, the multilevel functional principal component analysis (MFPCA) method was developed [1]. MFPCA decomposes observed functional data into three parts i) the overall mean, common for all subjects, ii) the subject–specific deviation from the overall mean, and iii) the remaining deviation from a subject–specific profile. Moreover, the method is able to transform high dimensional functional data (possibly

*Corresponding author: zuzana.rostakova@gmail.com

infinite) into finite dimensional vector spaces of principal components at two levels.

The original MFPCA method was proposed and validated only for data with the same number of observations per subject. In this article we demonstrate that in its original form the method is not able to properly detect subject–specific profiles when the number of observations among subjects is different. Therefore we propose the modification of the original MFPCA method which can better deal with the unbalanced data situation.

The article is organised in the following way. The general description of MFPCA is given in the first section. The modified MFPCA method for unbalanced data is proposed in Section 2. In Section 3 the method is validated on real–world sleep data. Finally, Section 4 provides discussion and a few concluding remarks.

1 Multilevel functional principal component analysis

MFPCA deals with functional data with repeated observations in order to detect sources of variability at two levels; the between– and within–subject variability [1].

Let consider I subjects with J observations $X_{ij}, i = 1, \ldots, I; j = 1, \ldots, J$. For simplicity we assume that observed functional data are defined at the same time grid within a closed interval T and are sufficiently smooth. Moreover the observations or visits within each subject should have natural ordering. In [1], the authors used a two–way functional ANOVA model in order to decompose X_{ij} into a fixed and random part

$$X_{ij}(t) = \mu(t) + \eta_j(t) + Z_i(t) + W_{ij}(t), \qquad t \in T. \tag{1}$$

The overall mean μ and the visit–specific deviation from the overall mean $\eta_j, j = 1, \ldots, J$ are fixed effects. For identifiability we assume $\sum_{j=1}^{J} \eta_j(t) = 0$, $t \in T$. The subject–specific deviation from the visit–specific mean Z_i and the remaining deviation from the subject– and visit–specific profiles W_{ij} are uncorrelated stochastic processes with mean 0 and covariance functions $S_1 : T \times T \to \mathbb{R}$ and $S_2 : T \times T \to \mathbb{R}$.

According to the Karhunen-Loewe expansion the stochastic processes Z_i and W_{ij} can be decomposed in the following way

$$Z_i(t) = \sum_{k=1}^{\infty} \alpha_{ik} \phi_k^{(1)}(t) \qquad\qquad W_{ij}(t) = \sum_{l=1}^{\infty} \beta_{ijl} \phi_l^{(2)}(t)$$

where α_{ik} and β_{ijl} are random variables with mean 0 and

$$\mathrm{E}(\alpha_{ik}\alpha_{il}) = \begin{cases} 0, & \text{if } k \neq l, \\ \lambda_k^{(1)}, & \text{if } k = l, \end{cases} \qquad \mathrm{E}(\beta_{ijk}\beta_{ijl}) = \begin{cases} 0, & \text{if } k \neq l, \\ \lambda_k^{(2)}, & \text{if } k = l. \end{cases}$$

Moreover, $\{\alpha_{ik}, k = 1, 2, \dots\}$ are uncorrelated with $\{\beta_{ijl}, l = 1, 2, \dots\}$. We call them the level 1 and level 2 principal component scores. Two sets of orthonormal functional bases of the L^2 space

$$\{\phi_k^{(1)}, k = 1, 2, \dots\} \qquad \text{and} \qquad \{\phi_l^{(2)}, l = 1, 2, \dots\}$$

which represents the functional principal components (FPCs) at level 1 and level 2 are not necessarily mutually orthogonal.

In [1], the following three covariance functions are considered in order to estimate functional principal components at both levels

$$\begin{aligned} K_T(s,t) &= Cov\left(X_{ij}(s), X_{ij}(t)\right) = S_1(s,t) + S_2(s,t), \\ K_B(s,t) &= Cov\left(X_{ij}(s), X_{ik}(t)\right) = S_1(s,t), \\ K_W(s,t) &= K_T(s,t) - K_B(s,t) = \frac{1}{2} Cov\left(X_{ij}(s) - X_{ik}(s), X_{ij}(t) - X_{ik}(t)\right) = S_2(s,t). \end{aligned}$$

In other words, FPCs at level 1 are eigenfunctions of K_B and FPCs at level 2 are eigenfunctions of K_W.

Using the method of moments, the following estimators of unknown quantities are proposed in [1]

$$\widehat{\mu}(t) = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}(t), \qquad \widehat{\eta_j}(t) = \frac{1}{I}\sum_{i=1}^{I} X_{ij}(t) - \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ij}(t), \qquad t \in T$$

$$\widehat{K_T}(s,t) = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} \left(X_{ij}(s) - \widehat{\mu}(s) - \widehat{\eta_j}(s)\right)\left(X_{ij}(t) - \widehat{\mu}(t) - \widehat{\eta_j}(t)\right), \tag{2}$$

$$\widehat{K_B}(s,t) = \frac{1}{IJ(J-1)}\sum_{i=1}^{I}\sum_{j\neq l}^{J} \left(X_{ij}(s) - \widehat{\mu}(s) - \widehat{\eta_j}(s)\right)\left(X_{il}(t) - \widehat{\mu}(t) - \widehat{\eta_l}(t)\right), \tag{3}$$

$$\widehat{K_W}(s,t) = \widehat{K_T}(s,t) - \widehat{K_B}(s,t), \tag{4}$$

where $\widehat{\mu}$ and $\widehat{\eta_j}$ are estimated similarly as in the standard ANOVA model [1].

The way of selecting the number of functional principal components at each level separately, as well as the procedure for computing principal component scores at both levels are described in details in [1].

2 MFPCA for unbalanced data design

The original MFPCA algorithm was designed for balanced data with ordered visits. However, the authors state that this assumption is not restrictive and the method is able to deal with unbalanced data as well.

Let consider I subjects with $J_i, i = 1, \ldots, I$ observations. In this case, the number of observations may differ among subjects and we assume that the order of observations within each subject is exchangeable. Therefore the visit–specific deviations η_j from the overall mean are set to zero. The model (1) changes into one–way functional ANOVA

$$X_{ij}(t) = \mu(t) + Z_i(t) + W_{ij}(t), \quad t \in T, \quad j = 1, \ldots, J_i, \quad i = 1, \ldots I. \tag{5}$$

By computing the expected values of the covariance functions estimators (2), (3) and (4) for data with unbalanced design and $\widehat{\eta}_j \equiv 0$ we obtain

$$\begin{aligned}
\mathrm{E}\left(\widehat{K_T}(s,t)\right) &= \left(1 - \frac{A_2}{A_1^2}\right) S_1(s,t) + \left(1 - \frac{1}{A_1}\right) S_2(s,t), \\
\mathrm{E}\left(\widehat{K_B}(s,t)\right) &= \left(1 - \frac{2}{A_1}\frac{A_3 - A_2}{A_2 - A_1} + \frac{A_2}{A_1^2}\right) S_1(s,t) - \frac{1}{A_1} S_2(s,t), \\
\mathrm{E}\left(\widehat{K_W}(s,t)\right) &= \left(\frac{2}{A_1}\frac{A_3 - A_2}{A_2 - A_1} - 2\frac{A_2}{A_1^2}\right) S_1(s,t) + S_2(s,t),
\end{aligned} \tag{6}$$

where $A_k = \sum_{i=1}^{I} J_i^k, k = 1, 2, 3$. It means, that for $I \to \infty$ and a bounded number of observations for each subject $1 \leq J_i \leq M, M \in \mathbb{N}$, the matrices $\widehat{K_B}$ and $\widehat{K_W}$ are only asymptotically unbiased estimators of S_1 and S_2.

Therefore, when data are unbalanced, we propose the following modification of the covariance functions estimators. First, let define

$$\begin{aligned}
\widehat{K_W}^{UU}(s,t) &= \frac{1}{\sum_{i=1}^{I} J_i} \sum_{i=1}^{I} \sum_{j:J_i>1}^{J_i} \left(X_{ij}(s) - \widehat{\mu}(s)\right)\left(X_{ij}(t) - \widehat{\nu}_i^{(-j)}(t)\right), \\
\widehat{\nu}_i^{(-j)}(t) &= \frac{1}{J_i - 1} \sum_{l \neq j}^{J_i} X_{il}(t), \quad t \in T.
\end{aligned}$$

While $\mathrm{E}\left(\widehat{K_W}^{UU}(s,t)\right) = S_2(s,t)$ which holds also for unbalanced data, we can estimate FPCs at level 2 directly from $\widehat{K_W}^{UU}$. The estimator (2) for K_T remains the same with expected value (6). Therefore FPCs at level 1 can be estimated as eigenfunctions of the following function

$$\widehat{K_B}^{UU} = \frac{A_1^2}{A_1^2 - A_2}\left(\widehat{K_T} - \frac{A_1 - 1}{A_1}\widehat{K_W}^{UU}\right), \qquad \mathrm{E}\left(\widehat{K_B}^{UU}(s,t)\right) = S_1(s,t).$$

3 Application to sleep data

Sleep is a continuous process which can be described by a finite number of sleep stages. Probabilistic sleep model (PSM) characterises sleep with probability values of 20 sleep microstates [3]. Considering the probability values as a function of time we obtain a curve.

In the first step we took 292 probabilistic sleep curves of the PSM applied to sleep recordings from the SIESTA database [2]. These curves represent the sleep microstate similar to REM (or rapid eye movement sleep stage). Using the two–step clustering approach [4], the curves were divided into 12 clusters depicted in Figure 1. Objective of this study is to identify cluster representatives, which can be used for the further analysis of the sleep process. With this aim in mind, we applied model (5) to the clustered curves. Effectively this means that we have 12 clusters (or 'subjects') with a different number of observations, in this case the number of curves in each cluster. The number of curves varied from 4 (cluster 9) to 117 (cluster 12).

Using the original and modified MFPCA algorithms the cluster–specific profiles $P_i(t) = \widehat{\mu}(t) + \widehat{Z_i}(t),\ \ t \in T$ were computed for each cluster. The superior performance of the modified MFPCA algorithm is visible especially for clusters 2, 5 or 9 consisting of a smaller number of curves. Taking into account that the original sleep probabilistic curves are strictly positive, the cluster–specific profiles estimated by the original MFPCA method reached for short time subintervals unexpected negative values.

4 Conclusion

In this article we described modified version of the multilevel functional principal component analysis method [1]. MFPCA is an appropriate tool for detection of main direction of variability for functional data with repeated observations. Original MFPCA was developed only for balanced data where each subject has the same number of observations and the observations have natural order.

However, we found and demonstrated on real sleep data, that in its original form the algorithm applied to unbalanced data leads to inferior results because the estimators of covariance functions described in [1] are biased.This is especially true for datasets with a small sample size.

In this article we proposed the modified estimators of covariance functions for unbalanced data. These leads to the unbiased estimation of functional principal components at level 1 and 2. We proved good performance of the proposed modified version of MFPCA on the analysed sleep data.

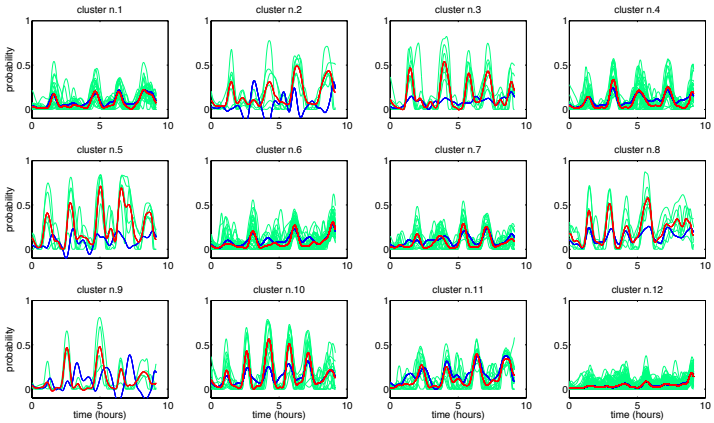

Figure 1: Cluster analysis of the sleep microstate similar to the REM sleep stage with 292 sleep probabilistic curves (light green) divided into 12 clusters. Cluster–specific profiles estimated by the modified MFPCA algorithm (red) form better cluster representatives than their counterparts estimated by the original MFPCA algorithm (blue).

Acknowledgements: This work has been supported by the Slovak Research and Development Agency (grant APVV-0668-12) and by the VEGA grant 2/0011/16.

References

[1] C.-Z. Di, C. M. Crainiceanu, B. S. Caffo, and N. M. Punjabi. Multilevel functional principal component analysis. *The Annals of Applied Statistics*, 3(1):458–488, 2009.

[2] G. Klösch, B. Kemp, T. Penzel, A. Schlögl, P. Rappelsberger, E. Trenker, G. Gruber, J. Zeitlhofer, B. Saletu, W. Herrmann, S. Himanen, D. Kunz, M. Barbanoj, J. Röschke, A. Varri, and G. Dorffner. The SIESTA project polygraphic and clinical database. *Medicine and Biology Magazine*, 20(3):51–57, 2001.

[3] A. Lewandowski, R. Rosipal, and G. Dorffner. Extracting more information from EEG recordings for a better description of sleep. *Computer Methods and Programs in Biomedicine*, 108(3):961 – 972, 2012.

[4] Z. Rošťáková and R. Rosipal. A novel two-step iterative approach for clustering functional data. In *22nd International Conference on Computational Statistics (COMPSTAT 2016): Book of Abstracts*, page 61, 2016.