Institute of Statistics, Graz University of Technology ■ Kopernikusgasse 24/III, 8010 Graz, Austria ■

Preprocessing noisy functional data using factor models

Siegfried Hörmann

Institute of Statistics, Graz University of Technology, Graz, Austria
shoermann@tugraz.at

Fatima Jammoul

Institute of Statistics, Graz University of Technology, Graz, Austria
f.jammoul@tugraz.at

Functional Data Analysis (FDA) is concerned with the analysis of data that naturally stems from some underlying functional form. Prominent examples include temperature data, growth curves, pollution level curves or 2D images. In practice, these data are most commonly observed on a grid and are susceptible to some level of noise. We consider data $(X_t(s)\colon 0 \le s \le 1), t = 1, \dots, T$. We assume that the data has been observed at regular datapoints $0 \le s_1 < s_2 < \ldots < s_p \le 1$ with some additional error. So we actually observe

$$y_t = (X_t(s_1), \dots, X_t(s_p))' + (u_{t1}, \dots, u_{tp})' =: X_t(\boldsymbol{s}) + u_t$$

Naturally, it is of interest to recover the latent signal $X_t(\boldsymbol{s})$. There are many different approaches to obtain estimates $\hat{X}_t$ for the signal X_t, most notably spline curve fitting techniques (Ramsay and Silverman, 2005) and kernel smoothing are employed. These techniques are susceptible to a systematic bias that stems from smoothing curve by curve. A more attractive route is to include the information of the entire sample in the estimation of each curve in order to recover systematic properties. Such approaches have been considered in Staniswalis and Lee (1998) and employ functional principal components (FPCA). We consider a similar idea. It can be shown that square integrable curves X_t can be represented as a factor model

$$y_t = \mu(\boldsymbol{s}) + Bf_t + u_t,$$

where in factor model language $X_t(\boldsymbol{s}) - \mu(\boldsymbol{s}) = Bf_t$ is referred to as the *common component* and u_t is the *idiosyncratic component*. This representation follows from the Karhunen-Loève expansion.

Theoretical Results

There are numerous ways to obtain estimates for factor models and in practice, one may choose their preferred approach. In order to verify our findings theoretically, we have considered a classic principal component (PCA) approach (Bai, 2003). Here let $Y = (y_1, \dots, y_T)$ and define $U = (u_1, \dots, u_T)$ and $F' = (f_1, \dots, f_T)$. We may write $Y = BF' + U$ and let $\hat{E} = (\hat{e}_1, \dots, \hat{e}_L)$ be the eigenvectors of the $T \times T$ matrix $T^{-1}Y'Y$ associated to the L largest eigenvalues $\hat{\gamma}_1 \ge \ldots \ge \hat{\gamma}_L$. Then we obtain the estimates $\hat{F} = \sqrt{T}\hat{E}$, $\hat{B} = T^{-1}Y\hat{F}$ and in summary

$$(\hat{X}_1(\boldsymbol{s}), \dots, \hat{X}_T(\boldsymbol{s})) = \hat{B}\hat{F}' = Y\hat{E}\hat{E}'.$$

Let λ_ℓ be the non-increasing eigenvalues of the covariance operator $\Gamma^X(s, s') = \mathrm{Cov}(X_t(s), X_t(s'))$ and $\varphi_\ell(s)$ the associated eigenfunctions. Then we make the following assumptions.

Assumption 1. *The noise process (u_t) is i.i.d. zero mean and independent of the signals (X_t). The processes $(u_{ti}\colon 1 \le i \le p)$ are Gaussian with covariance function $\gamma^u(h) = \mathrm{Cov}(u_{t(i+h)}, u_{ti})$, such that $\sum_{h\in\mathbb{Z}} |\gamma^u(h)| \le C_u < \infty$.*

Assumption 2. *(a) The process $(X_t\colon t \ge 1)$ is zero mean and L^4-m-approximable. (b) The curves $X_t = (X_t(s)\colon s \in [0,1])$ define fourth order random processes with a continuous covariance kernel. (c) It holds that $E\sup_{s\in[0,1]} X_1^2(s) < \infty$. (d) Observations X_t lie in some L-dimensional function space, where $L = L(T)$ may diverge with $T \to \infty$.*

Assumption 3. *It holds that $\max_{1\le k,\ell\le L} \left|p^{-1}\sum_{i=1}^p \varphi_k(s_i)\varphi_\ell(s_i)\right| = O(1)$ as $T \to \infty$.*

Here L denotes the number of factors considered in the model. We distinguish the case of *fixed* L and *growing* $L(T)$. In classic factor analysis, only the former case is considered. We here present the case of growing L. Take note that L is assumed to be known in the following theorems. The results remain valid for consistent estimates $\hat{L}$ of L.

Theorem 1. *Let Assumptions 1–3 hold. Assume that $p = p(T)$ diverges at a subexponetial rate and that $L = L(T)$ diverges at a subpolynomial rate. Furthermore, assume that for some $\nu > 0$ and some $\rho > 0$ we have $\lambda_j \ge \rho j^{-\nu}$ and that there is some $\beta > 0$ such that $p/\hat{\gamma}_L = O_P(L^\beta)$. Then, for any $1 \le t \le T$ it holds that*

$$\max_{1\le j\le p} |\hat{X}_t(s_j) - X_t(s_j)| = O_P\left(L^{2\beta+5/2+\nu/2}\left(\frac{1}{T^{1/4}} + \frac{T^{1/4}}{\sqrt{p}}\right)\right).$$

Moreover, if $E\|X_1\|^q < \infty$ for some $q > 4$, then for all $q' < q$ we have

$$\max_{1\le t\le T}\max_{1\le j\le p} |\hat{X}_t(s_j) - X_t(s_j)| = o_P\left(L^{2\beta+5/2+\nu}T^{1/q'}\left(\frac{1}{T^{1/4}} + \frac{T^{1/4}}{\sqrt{p}}\right)\right).$$

For proofs and detailed results we refer to our preprint (Hörmann and Jammoul, 2020a).

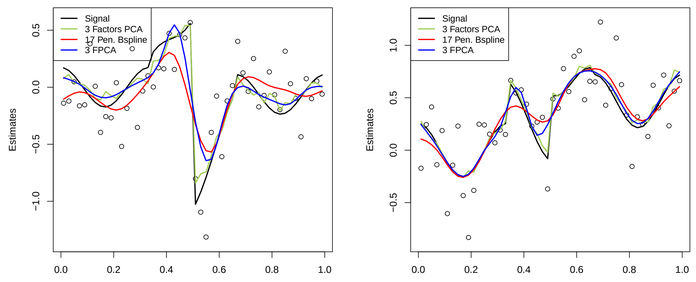

Figure 1: Estimates for the rough signal simulation for $p = 50$, $T = 200$ and $\sigma_u^2 = 0.05$. Dots represent the noisy observations.

Simulations

We have conducted several simulation studies comparing the factor model approach (PCA) to other prominent approaches, most notably Penalized Bsplines (PB) and a functional principal components (FPCA) approach as motivated by Staniswalis and Lee (1998). One can then see that especially for a growing number of curves T the factor model approach outperforms its competitors. In particular in cases with rough underlying signal $X_t(s)$ we find that the factor model manages to more accurately extract harsh features in comparison to the PB and FPCA models. We show a selection of plots from such a simulation, where the underlying signal X_t is obtained as

$$X_t(s) = \sum_{k=1}^{3} \xi_{tk}\varphi_k(s), s \in [0,1]$$

where $\varphi_1(s) = \mathbb{1}_{\{s>1/3\}}$, $\varphi_2(s) = (-1)^\kappa 4(0.2 - |s - 0.5|)\mathbb{1}_{\{s\in[1/3,2/3]\}}$ and $\varphi_3(s) = \cos 6\pi s$, with $\kappa = \mathbb{1}_{\{s\in(1/2,2/3]\}}$. Here, $\mathbb{1}_{\{s\in[a,b]\}}$ denotes the characteristic function, that is $\mathbb{1}_{\{s\in[a,b]\}} = 1$ if $s \in [a,b]$ and 0 otherwise. The associated scores are given by $\xi_{tk} \overset{\text{iid}}{\sim} N(0, 2^{-2(k-1)})$ for $k = 1, 2, 3$. The noisy observations are obtained via $y_{ti} = X_t(s_i) + u_{ti}$, where $u_{ti} \overset{iid}{\sim} N(0, \sigma_u^2)$. For further simulation studies and real life data investigations we refer to our preprint (Hörmann and Jammoul, 2020b).

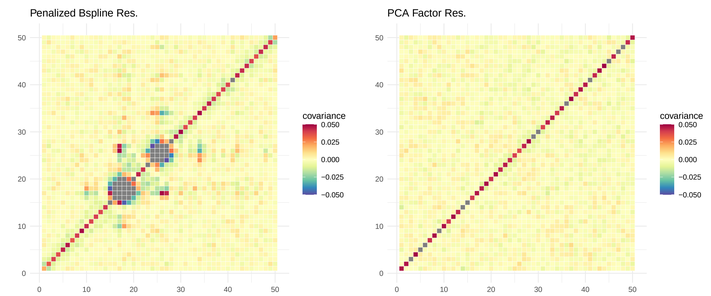

Figure 2: Heatmaps of the empirical covariance matrix of the residuals for the PB (left) and PCA approach (right) for the rough signal simulation.

References

J. Bai. Inferential theory for factor models of large dimensions. *Econometrica*, 71: 135–171, 2003.

S. Hörmann and F. Jammoul. Consistently recovering the signal from noisy functional data. *https://arxiv.org/abs/2012.05051*, 2020a.

S. Hörmann and F. Jammoul. Preprocessing noisy functional data using factor models. *https://arxiv.org/abs/2012.05824*, 2020b.

J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer, New York, 2005.

J. Staniswalis and J. Lee. Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93:1403–1418, 1998.