Supplementary material for "On the use of the Gram matrix for multivariate functional principal components analysis"

Steven Golovkine[*] Edward Gunning[†] Andrew J. Simpkin[‡] Norma Bargary[§]

October 21, 2024

In this Supplementary Material, we provide insights for when the data are already decomposed in a basis, e.g. Fourier or polynomials. In particular, we explain how to perform MFPCA as described in Section 4 in the main text.

1 Basis decomposition

In many practical situations, functional data are noisy and only observed at specific time points. To extract the underlying functional features of the data, smoothing and interpolation techniques are commonly employed. These techniques involve approximating the true underlying function generating the data by a finite-dimensional set of basis functions. Assume that for each feature $p = 1, \dots, P$, there exists a set of basis of functions $\Psi^{(p)} = \{\psi_k^{(p)}\}_{1 \leq k \leq K_p}$ such that each feature of each curve $n = 1, \dots, N$ can be expanded using the basis:

$$X_n^{(p)}(t_p) = \sum_{k=1}^{K_p} c_{nk}^{(p)} \psi_k^{(p)}(t_p), \quad t_p \in \mathcal{T}_p,$$

where $\{c_{nk}^{(p)}\}_{1 \leq k \leq K_p}$ is a set of coefficients for feature p of observation n. We denote by $\overline{c}_k^{(p)} = \sum_{n=1}^N \pi_n c_{nk}^{(p)}$ the mean coefficient of feature p corresponding to the kth basis function. The pth feature of the mean function can be then expanded in the same basis as:

$$\widehat{\mu}^{(p)}(t_p) = \sum_{k=1}^{K_p} \overline{c}_k^{(p)} \psi_k^{(p)}(t_p), \quad t_p \in \mathcal{T}_p.$$

Similarly, the covariance function of the pth and qth features is given by:

$$\widehat{C}_{p,q}(s_p, t_q) = \sum_{k=1}^{K_p} \sum_{l=1}^{K_q} \left(\sum_{n=1}^N \pi_n c_{nk}^{(p)} c_{nl}^{(q)} - \overline{c}_k^{(p)} \overline{c}_l^{(q)} \right) \psi_k^{(p)}(s_p) \psi_l^{(q)}(t_q), \quad s_p \in \mathcal{T}_p, \quad t_q \in \mathcal{T}_q.$$

[*] MACSI, Department of Mathematics and Statistics, University of Limerick, Ireland steven.golovkine@ul.ie

[†] Department of Biostatistics and Epidemiology, University of Pennsylvania, USA edward.gunning@pennmedicine.upenn.edu

[‡] School of Mathematical and Statistical Sciences, University of Galway, Ireland andrew.simpkin@nuigalway.ie

[§] MACSI, Department of Mathematics and Statistics, University of Limerick, Ireland norma.bargary@ul.ie

These formulas can be written in matrix form as follows. For $\mathbf{t} \in \mathcal{T}$, we have that $X(\mathbf{t}) = \mathbf{C}\Psi(\mathbf{t})$ where $X(\mathbf{t})$ is a $N \times P$ matrix with entries $X_n^{(p)}(t_p),\ t_p \in \mathcal{T}_p,\ 1 \leq p \leq P,\ 1 \leq n \leq N,$

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}^{(1)} & \cdots & \mathbf{C}^{(P)} \end{pmatrix}, \quad \text{and} \quad \Psi(\mathbf{t}) = \text{diag}\{\Psi^{(1)}(t_1), \dots, \Psi^{(P)}(t_P)\},$$

where

$$\mathbf{C}^{(p)} = \begin{pmatrix} c_{11}^{(p)} & \cdots & c_{1K_p}^{(p)} \\ \vdots & \ddots & \vdots \\ c_{N1}^{(p)} & \cdots & c_{NK_p}^{(p)} \end{pmatrix} \quad \text{and} \quad \Psi^{(p)}(t_p) = \begin{pmatrix} \psi_1^{(p)}(t_p) \\ \vdots \\ \psi_{K_p}^{(p)}(t_p) \end{pmatrix}.$$

Using the basis expansion and denoting $\Pi^\top = (\pi_1, \dots, \pi_N)$, the mean and covariance functions are given by

$$\widehat{\mu}(\mathbf{t}) = \Psi(\mathbf{t})^\top \mathbf{C}^\top \Pi \quad \text{and} \quad \widehat{C}(\mathbf{s}, \mathbf{t}) = \Psi(\mathbf{s})^\top \mathbf{C}^\top \left(\text{diag}\{\pi_1, \dots, \pi_N\} - \Pi\Pi^\top \right) \mathbf{C}\Psi(\mathbf{t}).$$

Finally, we denote by $\mathbf{W}$ the matrix of inner products of the functions in the basis Ψ. The matrix $\mathbf{W}$ is a block-diagonal matrix such that $\mathbf{W} = \text{blockdiag}\{\mathbf{W}^{(1)}, \dots, \mathbf{W}^{(P)}\}$ where each entry is given by

$$\mathbf{W}_{k,l}^{(p)} = \left\langle \psi_k^{(p)}, \psi_l^{(p)} \right\rangle, \quad 1 \leq k, l \leq K_p, \quad 1 \leq p \leq P.$$

We remark that, if the basis Ψ is an orthonormal basis, the matrix $\mathbf{W}$ is equal to the identity matrix of size $\sum_{p=1}^P K_p$. Using the expansion of the data into the basis of functions Ψ, the inner-product matrix $\mathbf{M}$ is written

$$\mathbf{M} = \text{diag}\{\sqrt{\pi_1}, \dots, \sqrt{\pi_N}\} \left(\mathrm{I}_N - \mathbf{1}_N \Pi^\top \right) \mathbf{CWC}^\top \left(\mathrm{I}_N - \Pi \mathbf{1}_N^\top \right) \text{diag}\{\sqrt{\pi_1}, \dots, \sqrt{\pi_N}\} \tag{SM.1}$$

where I_N is the identity matrix of size N and $\mathbf{1}_N$ is a vector of 1 of length N.

2 MFPCA with a basis expansion

In this section, we assume that the observations are expanded into a basis of functions, as explained in Section 1. Using the expansion of the data into the basis of function Ψ and $\mathbf{W}$, the matrix of inner products of the functions in the basis Ψ, we write (SM.1) as

$$\mathbf{M} = \left(\text{diag}\{\sqrt{\pi_1}, \dots, \sqrt{\pi_N}\} \left(\mathrm{I}_N - \mathbf{1}_N \Pi^\top \right) \mathbf{CW}^{1/2} \right) \left(\text{diag}\{\sqrt{\pi_1}, \dots, \sqrt{\pi_N}\} \left(\mathrm{I}_N - \mathbf{1}_N \Pi^\top \right) \mathbf{CW}^{1/2} \right)^\top.$$

We note

$$\mathbf{A} = \text{diag}\{\sqrt{\pi_1}, \dots, \sqrt{\pi_N}\} \left(\mathrm{I}_N - \mathbf{1}_N \Pi^\top \right) \mathbf{CW}^{1/2},$$

such that $\mathbf{M} = \mathbf{AA}^\top$. We also assume that $\phi_1, \phi_2, \dots$ the eigenfunctions of the covariance operator Γ have a decomposition into the basis Ψ

$$\phi_k(\cdot) = \begin{pmatrix} \phi_k^{(1)}(\cdot) \\ \vdots \\ \phi_k^{(P)}(\cdot) \end{pmatrix} = \begin{pmatrix} \psi^{(1)\top}(\cdot) b_{1k} \\ \vdots \\ \psi^{(P)\top}(\cdot) b_{Pk} \end{pmatrix}, \quad \text{where} \quad b_{pk} = \left(b_{pk1}, \dots, b_{pkK_p} \right)^\top.$$

We have, for $p = 1, \dots, P$,

$$\begin{aligned}
\left(\Gamma\phi_k\right)^{(p)}(\cdot) &= \sum_{q=1}^{P} \int_{\mathcal{T}_q} C_{p,q}(\cdot, s_q)\phi_k^{(q)}(s_q)\mathrm{d}s_q \\
&= \sum_{q=1}^{P} \int_{\mathcal{T}_q} \Psi(\cdot)^{(p)\top}\mathbf{C}^{(p)\top}\left(\operatorname{diag}\{\pi_1, \dots, \pi_N\} - \Pi\Pi^\top\right)\mathbf{C}^{(q)}\Psi^{(q)}(s_q)\Psi^{(q)}(s_q)^\top b_{qk}\mathrm{d}s_q \\
&= \Psi(\cdot)^{(p)\top}\mathbf{C}^{(p)\top}\left(\operatorname{diag}\{\pi_1, \dots, \pi_N\} - \Pi\Pi^\top\right)\sum_{q=1}^{P}\mathbf{C}^{(q)}\int_{\mathcal{T}_q}\Psi^{(q)}(s_q)\Psi(s_q)^{(q)\top}\mathrm{d}s_q b_{qk} \\
&= \Psi(\cdot)^{(p)\top}\mathbf{C}^{(p)\top}\left(\operatorname{diag}\{\pi_1, \dots, \pi_N\} - \Pi\Pi^\top\right)\sum_{q=1}^{P}\mathbf{C}^{(q)}\mathbf{W}^{(q)}b_{qk}.
\end{aligned}$$

This equation is true for all $p = 1, \cdots, P$, this can be rewritten with matrices as

$$\Gamma\phi_k(\cdot) = \Psi(\cdot)^\top\mathbf{C}^\top\left(\operatorname{diag}\{\pi_1, \dots, \pi_N\} - \Pi\Pi^\top\right)\mathbf{C}\mathbf{W}b_k.$$

From the eigenequation, we have that

$$\Gamma\phi_k(\cdot) = \lambda_k\phi_k(\cdot) \Longleftrightarrow \Psi(\cdot)^\top\mathbf{C}^\top\left(\operatorname{diag}\{\pi_1, \dots, \pi_N\} - \Pi\Pi^\top\right)\mathbf{C}\mathbf{W}b_k = \lambda_k\Psi(\cdot)^\top b_k.$$

Since this equation must be true for all $t_p \in \mathcal{T}_p$, this imply the equation

$$\mathbf{C}^\top\left(\operatorname{diag}\{\pi_1, \dots, \pi_N\} - \Pi\Pi^\top\right)\mathbf{C}\mathbf{W}b_k = \lambda_k b_k. \tag{SM.2}$$

As the eigenfunctions are assumed to be normalized, $\left\|\phi_k\right\|^2 = 1$. And so, $b_k^\top\mathbf{W}b_k = 1$. Let $u_k = \mathbf{W}^{1/2}b_k$. Then, from (SM.2), we obtain

$$\mathbf{W}^{1/2}\mathbf{C}^\top\left(\operatorname{diag}\{\pi_1, \dots, \pi_N\} - \Pi\Pi^\top\right)\mathbf{C}\mathbf{W}^{1/2}u_k = \lambda_k u_k \Longleftrightarrow \mathbf{A}^\top\mathbf{A}u_k = \lambda_k u_k. \tag{SM.3}$$

From the eigendecomposition of the matrix M, we get

$$\mathbf{M}\boldsymbol{u}_k = l_k\boldsymbol{u}_k \Longleftrightarrow \mathbf{A}\mathbf{A}^\top\boldsymbol{u}_k = l_k\boldsymbol{u}_k. \tag{SM.4}$$

The equations (SM.3) and (SM.4) are eigenequations in the classical PCA case, with the duality $X^\top X$ and $XX^\top$. Following Pagès (2014); Härdle and Simar (2019), we find that, for $1 \leq k \leq K$,

$$\lambda_k = l_k, \quad \boldsymbol{u}_k = \frac{1}{\sqrt{l_k}}\mathbf{A}u_k \quad \text{and} \quad u_k = \frac{1}{\sqrt{l_k}}\mathbf{A}^\top\boldsymbol{u}_k.$$

And finally, to get the coefficient of the eigenfunctions, for $1 \leq k \leq K$,

$$b_k = \mathbf{W}^{-1/2}u_k = \frac{1}{\sqrt{l_k}}\mathbf{C}^\top\left(\mathrm{I}_N - \Pi\mathbf{1}_N^\top\right)\operatorname{diag}\{\sqrt{\pi_1}, \dots, \sqrt{\pi_N}\}\boldsymbol{u}_k.$$

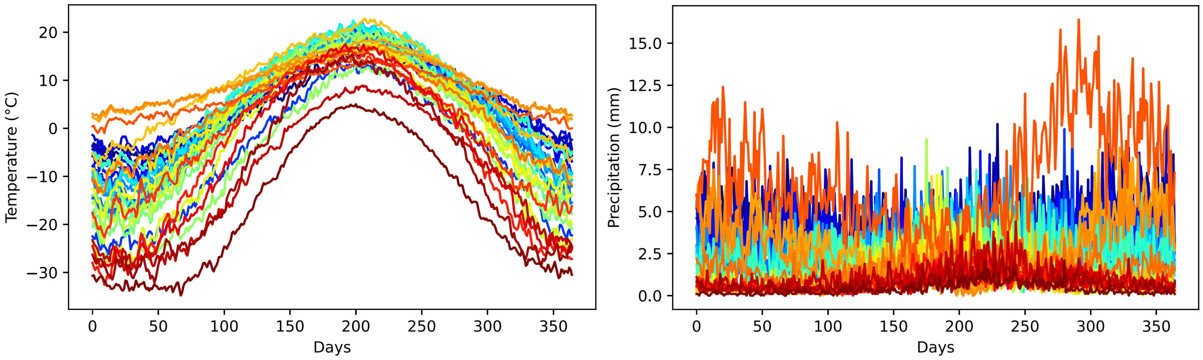


Figure 1: Canadian weather dataset.

3 Canadian weather

We apply the different methods to the Canadian weather dataset. This Canadian weather dataset, available in the R package `fda` Ramsay et al. (2023), provides daily temperature (°C) and precipitation (mm, rounded to 0.1mm) recordings for 35 Canadian cities. Originally presented in Ramsay and Silverman (2005), the dataset spans all days of the year, averaging data from 1960 to 1994. This is an example of multivariate functional data with two variables ($P = 2$), incorporating measurement errors. Figure 1 presents the data, showcasing temperature and precipitation trends across different cities.

We estimate the functional principal components using the `Gram`, `(Tensor) PCA` and `2D/1D B-Splines` methods from the Canadian weather dataset. Prior to applying each method, the data was smoothed using P-splines with a fixed penalty. For the `(Tensor) PCA` method, the estimation of the multivariate eigenfunctions is based on the univariate estimation of 5 univariate eigenfunctions. For the `2D/1D B-splines` method, the two components are expanded in 13 B-splines. Figure 2 presents the results of MFPCA retaining the top three principal components. Recalling that eigenfunctions are defined up to a sign, the results across all three methods are similar. Our analysis focuses on the Gram method's results. The first principal component (red) exhibits positive values for both temperature and precipitation, indicating that weather stations with positive scores will experience above-average temperatures and precipitation. This effect is more pronounced during winter than summer, as the eigenfunctions approach zero during the summer months. Similar interpretations can be applied to the remaining eigenfunctions.

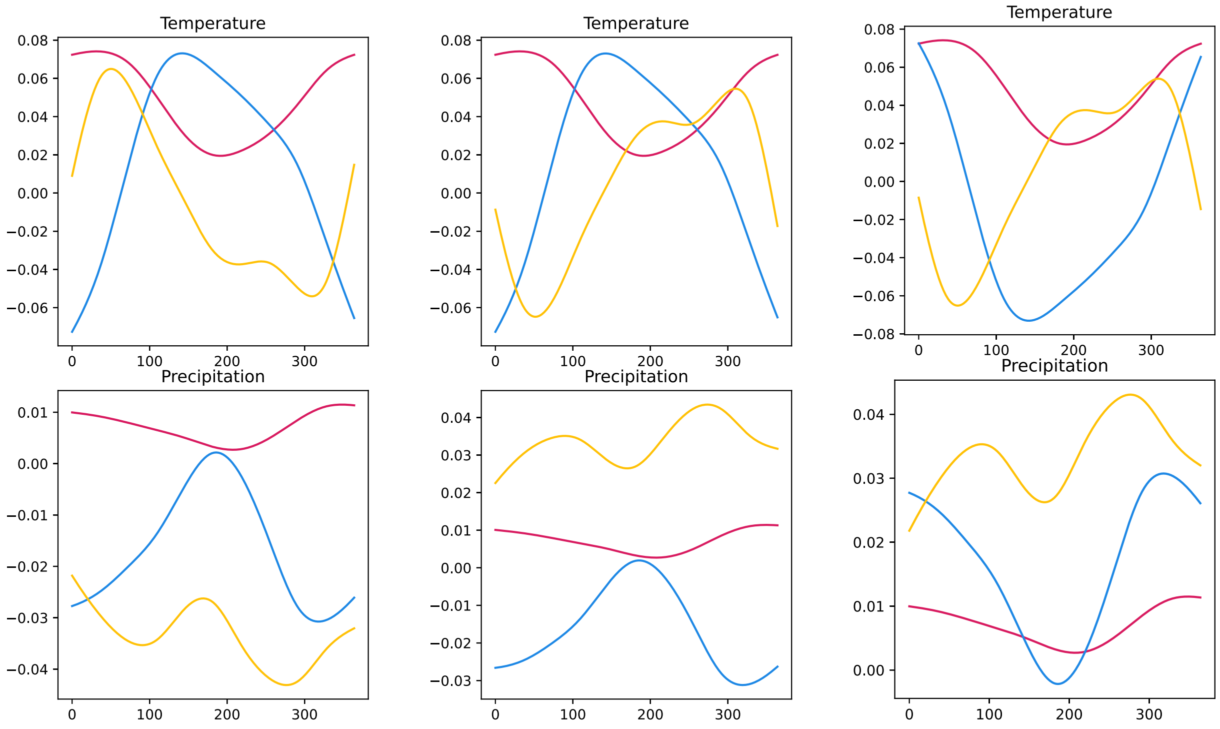
References

Härdle, W. K. and Simar, L. (2019). *Applied Multivariate Statistical Analysis*. Springer Nature.

Pagès, J. (2014). *Multiple Factor Analysis by Example Using R*. CRC Press.

Ramsay, J., Hooker, G., and Graves, S. (2023). Fda: Functional Data Analysis.

Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Science & Business Media.



(a) `Gram` method.

(b) `(Tensor) PCA` method.

(c) `2D/1D B-Splines` method.

Figure 2: The estimated eigenfunctions for the Canadian weather dataset using the different methods.