INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

Flying on the sphere

Analizying geometric patterns of behaviour of flights from London to Gothemburg using RFPCA.

Lars Daniel Johansson Niño

Other information:
Email: LJOHANSS@ITAM.MX
Git-hub: LarsDanielJohaN

Abstract

In the modern world, data appear in increasingly diverse structures. Some of these, due to their inherent nature, are better modeled in geometric spaces other than the Euclidean one. A relevant case involves stochastic processes that occur on spheres or, more generally, on Riemannian manifolds. In particular, we analyze flight trajectories between London and Gothenburg represented on the unit sphere, illustrating the usefulness of models in non-Euclidean spaces. We apply the Riemannian Functional Principal Component Analysis (RFPCA) proposed by [1] to identify the main patterns of variation in the flight trajectories. Furthermore, we compare this approach with its Euclidean counterpart, evaluating the method's performance on both the original and transformed data, and highlighting the advantages of each approach.

Data

The data consist of 191 flight trajectories from London to Gothenburg between January 1 and March 31, 2023, operated by the airline British Airways. These were collected from the OpenSky Network historical database [3] and include information such as latitude, longitude, date, time, etc. [3] obtained the records from live aircraft-reported data.

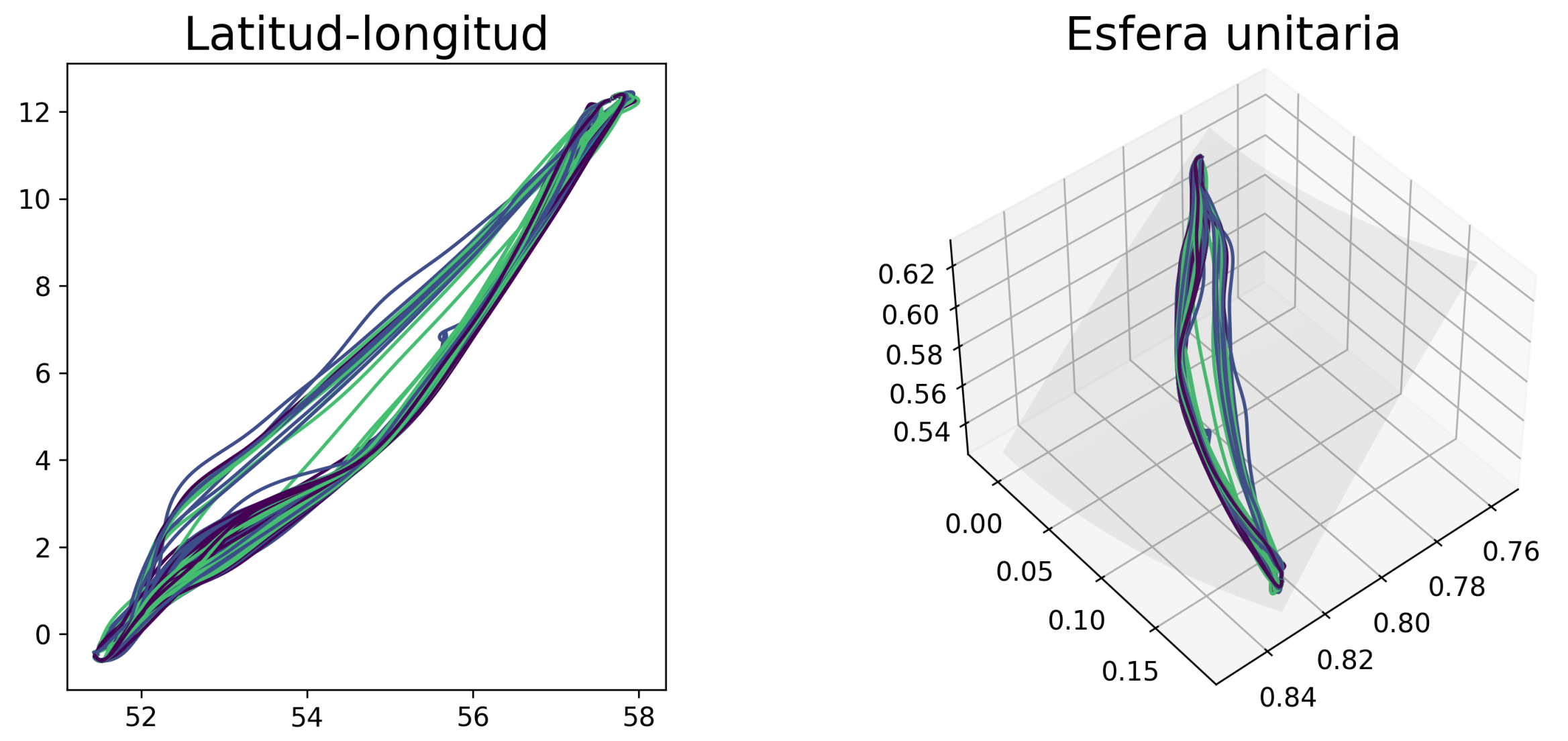

Figure 1: Left: Flights on a (lat, lon) form. Right: Projected data on $\mathcal{S}^2$.

To parameterize each flight, we took the initial time $t = 0$ / final time $t = 1$ as the first / last point (lat, lon, alt) with altitude greater than zero feet. To obtain observations on a common grid of points $(t_1, ..., t_D) \in [0, 1]^D$ ($D = 400$), the latitude and longitude curves were approximated for each flight. The approximations, done separately for each coordinate, were obtained using a penalized B-spline basis of degree 3. To obtain points on the unit sphere $\mathcal{S}^2 = \{\mathbf{x} \in \mathbb{R}^3 | \parallel \mathbf{x} \parallel_2 = 1\}$, the latitudes and longitudes were converted to radians and transformed into spherical coordinates.

Functional Principal Component Analysis (FPCA)

This tool is used to decompose the curves $Y(t) : [0, 1] \to \mathbb{R}^3$ into geometric patterns of variation. In a manner similar to principal component analysis (PCA), $Y(t)$ is expressed as a random deviation from a mean $\mu(t)$, determined by a linear combination of orthonormal functions $\phi_k(t)$ with random coefficients ξ_k that have zero mean $E\xi_k = 0$, and are pairwise uncorrelated $Corr(\xi_k, \xi_{k'}) = 0 \; k \neq k'$.

$$Y(t) = \mu(t) + \sum_{k \geq 1} \xi_k \phi_k(t) \qquad (1)$$

As in PCA, the contribution of the term $\xi_k \phi_k(t)$ to explaining the function $Y(t)$ is associated with the variance $Var(\xi_k) = \lambda_k$, and the proportion of variation explained by the first K terms corresponds to the ratio $PVE_K = \sum_{k=1}^{K} \lambda_k / \sum_{k \geq 1} \lambda_k$.

Functional Principal Component Analysis for Riemmanian Manifolds (RFPCA)

Loosely speaking, Riemannian manifolds are geometric spaces $\mathcal{M}$ whose local structure is Euclidean. This means that for a point $p \in \mathcal{M}$, a neighborhood around it $V_p = \{q \in \mathcal{M}; |d(p, q) < \varepsilon\}$ can be studied using a region U_p of $\mathbb{R}^n$. Under certain circumstances, V_p and U_p are related through the logarithmic map $\log_p(q) : V_p \to U_p$.We aim to study random curves on the unit sphere $\mathcal{S}^2 = \{\mathbf{x} \in \mathbb{R}^3 | \parallel \mathbf{x} \parallel_2 = 1\}$. Why not use FPCA directly? It does not take into account the specific geometry of $\mathcal{S}^2$, which means it is possible that $\phi_k(t), \mu(t) \notin \mathcal{S}^2$ for some time t.

$$\mu_{\mathcal{S}^2}(t) = \arg\min_{p \in \mathcal{S}^2} E\Big\{d^2(p, Y(t))\Big\} \quad (2) \qquad Y(t) \approx exp_{\mu_{\mathcal{S}^2}(t)}\Big(\sum_{k \geq 1} \xi_k \phi_k(t)\Big) \quad (3)$$

RFPCA is a generalization of FPCA, proposed by [1], for Riemannian manifolds in general. It aims to describe the variation of $Y(t)$ around the Fréchet mean (2). To do this, the logarithmic map is first applied to $Y(t)$ to obtain the transformations $X(t) = \log_{\mu_{\mathcal{S}^2}(t)}(Y(t))$. Then, the principal components of $X(t) = \sum_{k \geq 1} \xi_k \phi_k(t)$ are obtained using traditional FPCA. Finally, the $Y(t)$'s are approximated by applying the exponential map $exp_{\mu_{\mathcal{S}^2}}(\cdot)$, the inverse of $\log_{\mu_{\mathcal{S}^2}(t)}$, to the $\xi_k \phi_k(t)$'s, yielding (3).

$$U_K = E\Big\{\int_{\mathcal{T}} d^2(Y(t), Y_K(t))\, dt\Big\} \quad (4) \qquad Y_K(t) = exp_{\mu_{\mathcal{S}^2}(t)}\Big(\sum_{k=1}^{K} \xi_k \phi_k(t)\Big) \quad (5)$$

Since the $\phi_k(t)$'s do not lie in $\mathcal{S}^2$, the λ_k's cannot be interpreted as the PVE_K in FPCA. To obtain an analogue, the approximation of $Y(t)$ with K components is defined using (5) to compute the residual variance (4). For $K = 0$, we take $exp_{\mu_{\mathcal{S}^2}(t)}(0) = \mu_{\mathcal{S}^2}(t)$, so $Y_0(t)$ corresponds to the mean $\mu_{\mathcal{S}^2}(t)$. With this, the Fraction of Variation Explained is defined as $FVE_K = (U_0 - U_K)/U_0$.

Results

For all three methods, $K = 12$ principal components were estimated. With this number, FPCA reached 100% of the explained variability in both cases, while RFPCA reached 99.8%. In this sense, RFPCA has a slight disadvantage, as the FVE_K is lower than the PVE_K from FPCA on both the data in $\mathcal{S}^2$ and the original latitude-longitude representation. Nevertheless, the results across the strategies are relatively similar. For example, in all three cases, using just three components $K = 1, 2, 3$ was sufficient to explain 95% of the variability.

Method / $(P/F)VE_K$	1	2	3	4	5	6	7	8	9
RFPCA	0.8224	0.9107	0.9657	0.9806	0.9873	0.9921	0.9949	0.9964	0.9971
FPCA (data on $\mathcal{S}^2$)	0.8244	0.9122	0.9669	0.9818	0.9884	0.9934	0.9962	0.9977	0.9987
FPCA $(lat - lon)$	0.8395	0.9190	0.9705	0.9838	0.9896	0.9943	0.9964	0.9978	0.9988

In terms of interpretability for the data in $\mathcal{S}^2$, the first 3 components of FPCA and RFPCA are visually similar. However, a noticeable difference is observed regarding their curvature. This could suggest that, even though the PVE_K and FVE_K are similar, the first 3 components of FPCA might not be equally interpretable within the geometric context of $\mathcal{S}^2$.

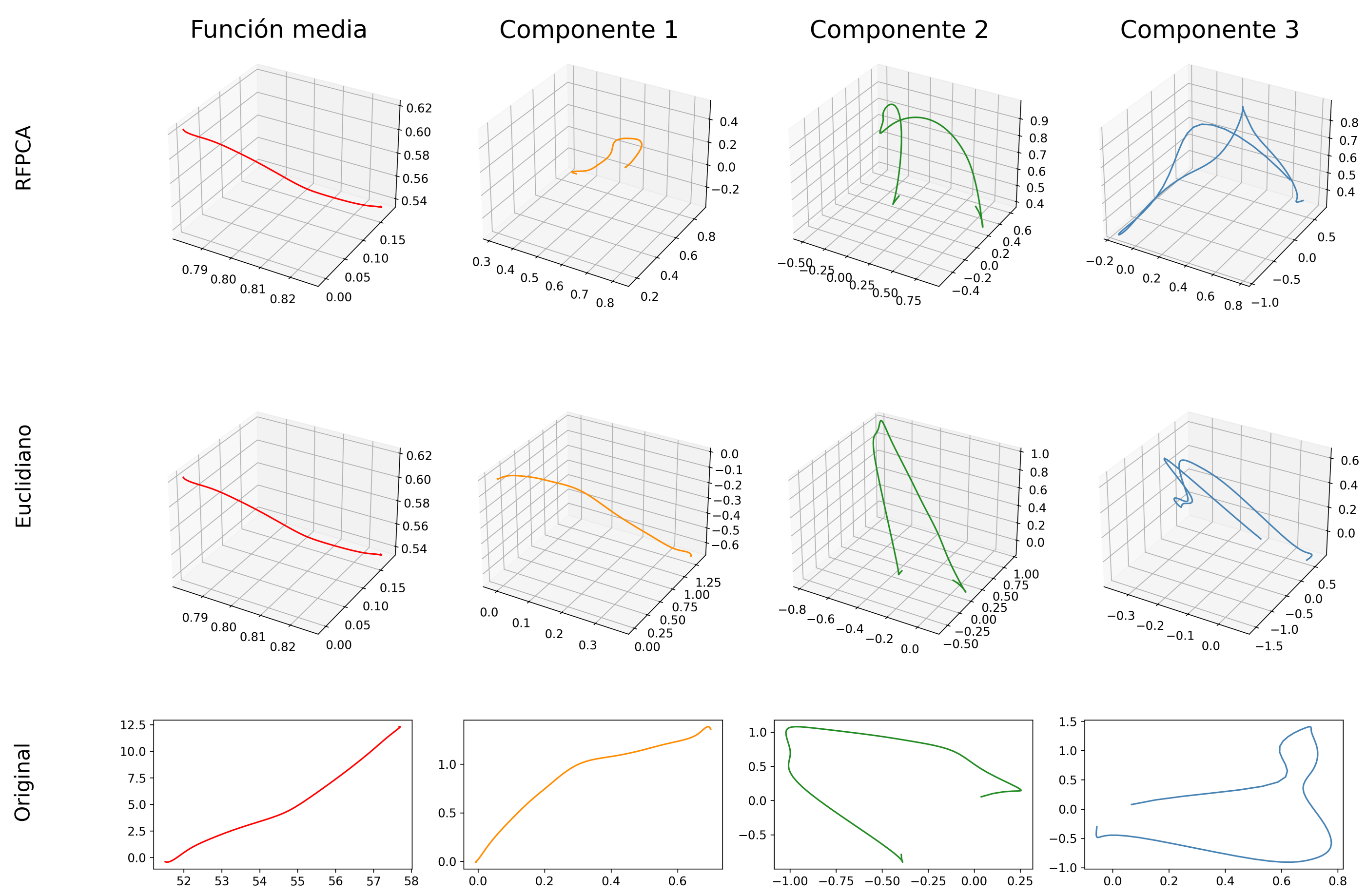

Figure 2: Mean function and first three principal components for various methods. .

Despite the visual similarity, the most notable difference was observed in the extent to which the FPCA components respect the geometry of $\mathcal{S}^2$. To quantify this, the norms $\parallel \hat{\phi}_k(t_d) \parallel_2$, $\parallel \hat{\mu}(t_d) \parallel_2$ were calculated over the original observation grid, and the proportion of these that were on $\mathcal{S}^2$ was computed. It was considered that $\hat{\mu}(t_d), \hat{\phi}_k(t_d) \in \mathcal{S}^2$ if $\Big| \parallel \hat{\mu}(t_d) \parallel_2 -1 \Big| <$ tol, $\Big| \parallel \hat{\phi}_k(t_d) \parallel_2 -1 \Big| <$ tol with tol $= 0.0001$. Under this criterion, the estimated mean $\hat{\mu}(t)$ had 70.25% of its points on $\mathcal{S}^2$, while all the components $\hat{\phi}_k(t)$ had a null proportion. With a tolerance level of tol $= 0.001$, all the points $\hat{\mu}(t_d)$ were on $\mathcal{S}^2$, while the proportions for the components $K = 2, 5$ increased to 0.25%, and to 0.5% for $K = 4, 9$.

Conclusion

In this work, we present FPCA for Riemannian manifolds and compare it with its Euclidean version. Both methods showed similar behavior in terms of explained variation and visual interpretation of the components. The main difference lies in how the estimates fit the geometry of $\mathcal{S}^2$: while the mean $\hat{\mu}(t)$ respects it within a certain tolerance, the components $\hat{\phi}_k(t)$ do not. Therefore, when geometry is relevant, RFPCA is more suitable. Future extensions could further highlight its advantages, especially in trajectories covering larger regions of the sphere (e.g. [1]) or through Bayesian estimation methods (e.g. [2] and related works) to explore uncertainty in the ξ_k's. Overall, this work exemplifies how using specialized methods for differentiable manifolds can make a crucial difference when the geometry of the observation space is significant.

I would like to express my gratitude to Dr. Laura Battagliola, Dr. Simón Lunagómez, and Dr. Xiongtao Dai for always answering my questions and stimulating my curiosity.

References

[1] Xiongtao Dai and Hans-Georg Müller. Principal component analysis for functional data on Riemannian manifolds and spheres. *The Annals of Statistics*, 46(6B):3334 – 3361, 2018.

[2] Tui H. Nolan, Jeff Goldsmith, and David Ruppert. Bayesian Functional Principal Components Analysis via Variational Message Passing with Multilevel Extensions. *Bayesian Analysis*, 20(1):157 – 183, 2025.

[3] Matthias Schäfer, Martin Strohmeier, Vincent Lenders, Ivan Martinovic, and Matthias Wilhelm. Bringing up opensky: A large-scale ads-b sensor network for research. In *Proceedings of the 13th IEEE/ACM International Symposium on Information Processing in Sensor Networks (IPSN)*, pages 83–94. IEEE, April 2014.