INSTITUTO TECNOLÓGICO AUTÓNOMO DE MÉXICO

Volando en la esfera

Analizando patrones de variación en vuelos de Londres a Gotemburgo utilizando RFPCA.

Lars Daniel Johansson Niño

Información:
Email: LJOHANSS@ITAM.MX
Git-hub: LarsDanielJohaN

Resumen

En el mundo moderno, los datos aparecen en estructuras cada vez más diversas. Algunos de estos, debido a su naturaleza inherente, se modelan mejor en espacios geométricos distintos al euclidiano. Un caso relevante son los procesos estocásticos que ocurren en esferas o, más generalmente, en variedades de Riemann. En particular, analizamos trayectorias de vuelos entre Londres y Gotemburgo representadas en la esfera unitaria, ejemplificando la utilidad de modelos en espacios no euclidianos. Aplicamos el Análisis Funcional de Componentes Principales para variedades de Riemann, propuesto por [1], para identificar los principales patrones de variación en las trayectorias de vuelo. Además, comparamos este enfoque con su equivalente en espacios euclidianos, evaluando el rendimiento del método tanto en los datos originales como en los transformados, y destacando las ventajas de cada enfoque.

Datos

Los datos consisten en 191 trayectorias de vuelos de Londres a Gotemburgo entre el 1 de enero al 31 de marzo del 2023 operados por la aerolínea British Airways. Estos fueron recolectados de la base de datos históricos de la OpenSky Network [3] e incluyen información de latitud, longitud, fecha, hora, etc. [3] obtuvieron los registros con información reportada por las aeronaves en vivo.

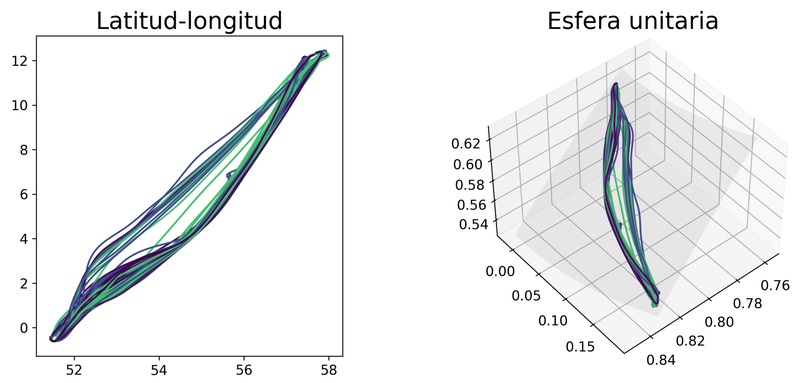

Figura 1: Izquierda: Vuelos en forma (lat, lon). Derecha: Vuelos proyectados a $\mathcal{S}^2$.

Para parametrizar cada vuelo, se tomo como tiempo inicial $t = 0$/tiempo final $t = 1$ al primer/ultimo punto (lat, lon, alt) cuya altura era mayor que cero pies. Para obtener observaciones en una maya de puntos común $(t_1, ..., t_D) \in [0, 1]^D$ $(D = 400)$ se aproximaron, por vuelo, las curvas de latitud/longitud. Las aproximaciones, realizadas por coordenada, se obtuvieron con una base de B-Splines penalizados de grado 3. Para obtener puntos en la esfera unitaria $\mathcal{S}^2 = \{\mathbf{x} \in \mathbb{R}^3 \mid \| \mathbf{x} \|_2 = 1\}$, las latitudes y longitudes se convirtieron a radianes y se tomaron sus coordenadas esféricas.

Análisis Funcional de Componentes Principales (FPCA)

Esta herramienta sirve para descomponer a las curvas $Y(t) : [0, 1] \to \mathbb{R}^3$ en patrones geométricos de variación. Con un sazón similar al análisis de componentes principales (PCA), $Y(t)$ se expresa como una desviación aleatoria de una media $\mu(t)$ determinada por una combinación lineal de funciones ortonormales $\phi_k(t)$ con coeficientes aleatorios ξ_k de media cero $E\{\xi_k\} = 0$, y no correlacionados a pares $Corr(\xi_k, \xi_{k'}) = 0$ $k \neq k'$.

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

Al igual que en el PCA, la contribución del termino $\xi_k \phi_k(t)$ a explicar la función $Y(t)$ se asocia a la varianza $Var(\xi_k) = \lambda_k$ y la proporción de variación explicada por los primeros K términos al cociente $PVE_K = \sum_{k=1}^{K} \lambda_k / \sum_{k \geq 1} \lambda_k$.

Análisis Funcional de Componentes Principales para Variedades de Riemann (RFPCA)

De forma vaga las Variedades de Riemann son espacios geométricos $\mathcal{M}$ cuya estructura local es euclidiana. Esto en el sentido de que para un $p \in \mathcal{M}$, una pieza de sus alrededores $V_p = \{q \in \mathcal{M} \mid d(p, q) < \varepsilon\}$ se puede estudiar con una porción U_p de $\mathbb{R}^n$. Dadas las circunstancias, V_p y U_p están asociadas por el mapeo logarítmico $\log_p(q) : V_p \to U_p$. Nosotros buscamos estudiar curvas aleatorias en la esfera unitaria $\mathcal{S}^2 = \{\mathbf{x} \in \mathbb{R}^3 | \| \mathbf{x} \|_2 = 1\}$. ¿Por qué no usar FPCA directamente? Este no considera la geometría particular de $\mathcal{S}^2$ por lo que podría ocurrir que $\phi_k(t), \mu(t) \notin \mathcal{S}^2$ para algún tiempo t.

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

El RFPCA es una generalización del FPCA, propuesta por [1], para Variedades de Riemann en general. Este busca describir la variación de $Y(t)$ alrededor de la media de Fréchet (2). Para ello, primero se aplica el mapeo logarítmico a $Y(t)$ para obtener las transformaciones $X(t) = \log_{\mu_{\mathcal{S}^2}(t)}(Y(t))$. Después, se encuentran los componentes principales de $X(t) = \sum_{k \geq 1} \xi_k \phi_k(t)$ con el FPCA tradicional. Finalmente, las $Y(t)$'s se aproximan al aplicar el mapeo exponencial $exp_{\mu_{\mathcal{S}^2}}(\cdot)$, el inverso de $\log_{\mu_{\mathcal{S}^2}(t)}$, a las $\xi_k \phi_k(t)$'s obteniendo (3).

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

Ya que las $\phi_k(t)$'s no viven en $\mathcal{S}^2$, las λ_k no se interpretan como el PVE_K del FPCA. Para obtener un análogo, se define la aproximación de $Y(t)$ con K componentes con (5) para obtener la varianza residual (4). Para $K = 0$ se toma $exp_{\mu_{\mathcal{S}^2}(t)}(0) = \mu_{\mathcal{S}^2}(t)$, por lo que $Y_0(t)$ es la media $\mu_{\mathcal{S}^2}(t)$. Con esto se define la Fracción de Variación Explicada $FVE_K = (U_0 - U_K)/U_0$.

Resultados

Para los tres métodos se estimaron $K = 12$ componentes principales. Con este número, el FPCA alcanzo el 100 % de la variabilidad explicada en ambos casos mientras el RFPCA alcanzó el 99.8 %. En este sentido el RFPCA tiene una leve desventaja, pues la FVE_K es mas baja que PVE_K del FPCA con los datos en $\mathcal{S}^2$ y sobre la latitud-longitud original. Aun así, los resultados entre las estrategias son relativamente similares. Por ejemplo, para los tres bastó con tomar tres componentes $K = 1, 2, 3$ para explicar el 95 % de la variabilidad.

Método / $(P/F)VE_K$	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 (datos en $\mathcal{S}^2$)	0.8244	0.9122	0.9669	0.9818	0.9884	0.9934	0.9962	0.9977	0.9987
FPCA (en $lat - lon$)	0.8395	0.9190	0.9705	0.9838	0.9896	0.9943	0.9964	0.9978	0.9988

En términos de interpretabilidad para los datos en $\mathcal{S}^2$, los primeros 3 componentes del FPCA y el RFPCA son, visualmente, similares. Sin embargo, se observa una diferencia apreciable en torno a la curvatura de los mismos. Esto podría sugerir que, incluso ante las similitud con respecto a PVE_K y FVE_K, los primeros 3 componentes del FPCA podrían no ser igual de interpretables dentro del contexto geométrico de $\mathcal{S}^2$.

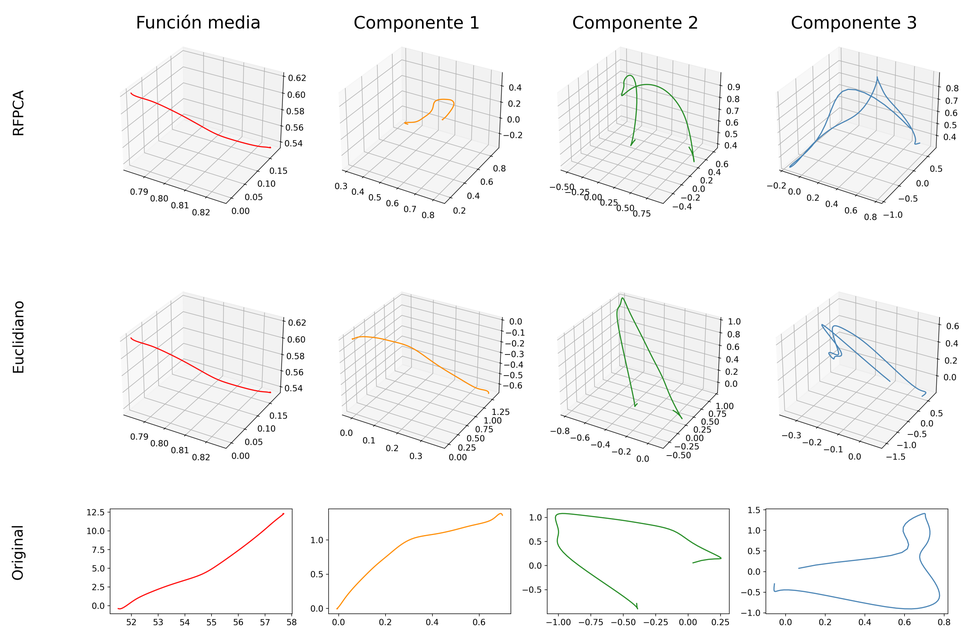

Figura 2: Función media y primeros 3 componentes principales para distintos metodos.

A pesar de la similitud visual, la diferencia más notable se observó en el grado en que los componentes del FPCA respetan la geometría de $\mathcal{S}^2$. Para cuantificar esto, se calcularon las normas $\| \hat{\phi}_k(t_d) \|_2$, $\| \hat{\mu}(t_d) \|_2$ sobre la maya de observación original y se calculo la proporción de estos que estaban en $\mathcal{S}^2$. Se consideró que $\hat{\mu}(t_d), \hat{\phi}_k(t_d) \in \mathcal{S}^2$ en caso de que $\Big| \| \hat{\mu}(t_d) \|_2 - 1 \Big| < tol$, $\Big| \| \hat{\phi}_k(t_d) \|_2 - 1 \Big| < tol$ con $tol = 0{,}0001$. Bajo este criterio, la media estimada $\hat{\mu}(t)$ tuvo el 70.25 % de sus puntos en $\mathcal{S}^2$, mientras todos los componentes $\hat{\phi}_k(t)$ tuvieron una proporción nula. Al tomar un nivel de tolerancia de $tol = 0{,}001$, todos los puntos $\hat{\mu}(t_d)$ se encontraban en $\mathcal{S}^2$, mientras los las proporciones para los componentes $K = 2, 5$ aumentaron al 0.25 % y a 0.5 % para $K = 4, 9$.

Conclusión

En este trabajo presentamos el FPCA para variedades de Riemann y lo comparamos con su versión euclidiana. Ambos métodos mostraron un comportamiento similar en variación explicada y en la interpretación visual de los componentes. La principal diferencia radica en cómo las estimaciones se ajustan a la geometría de $\mathcal{S}^2$: mientras que la media $\hat{\mu}(t)$ la respeta bajo cierta tolerancia, los componentes $\hat{\phi}_k(t)$ no. Por tanto, cuando la geometría es relevante, el RFPCA resulta más adecuado. Futuras extensiones podrían destacar aún más sus ventajas, especialmente en trayectorias que cubren regiones más amplias de la esfera (e.g. [1]) o mediante métodos de estimación bayesianos (e.g. [2] y relacionados) para explorar la incertidumbre en las ξ_k's. En conjunto, este trabajo ejemplifica cómo el uso de métodos especializados para variedades diferenciables puede marcar una diferencia crucial cuando la geometría del espacio de observación es significativa.

Por siempre responder mis dudas y empujar mis curiosidades expreso mi agradecimiento a la Dra. Laura Battagliola, al Dr. Simón Lunagómez, y al Dr. Xiongtao Dai.

Referencias

[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.