Functional Principal Component in Two-Dimensions

Spencer Matthews[a], **Esteban Escobar** [b]
[a,b]The University of California, Irvine, Department of Statistics

1 Introduction to Functional Principal Component Analysis (FPCA)

1.1 Functional Data

The first step in understanding FPCA is to introduce functional data. Suppose we have data on temperature changes in California each day for multiple years. Or perhaps we have the annual number of bird sighting for different species or daily closing prices on different houses in a city. As shown in Figure 1 left, we can graph our observation as a discrete time series. In functional data analysis, we would instead like to interpret our measured data as a single function or curve, as shown in Figure 1 right. There are multiple advantages to this method. For example, we can analyze the derivatives of functions, and measurements do not have to be taken at the same point in time to be comparable [5]. We do this by estimating the underlying function of our data, which induces smoothing and reduces noise. A common approach is to use basis expansion. The most common basis function is B-splines which our paper [4] also uses in the two-dimensional case.

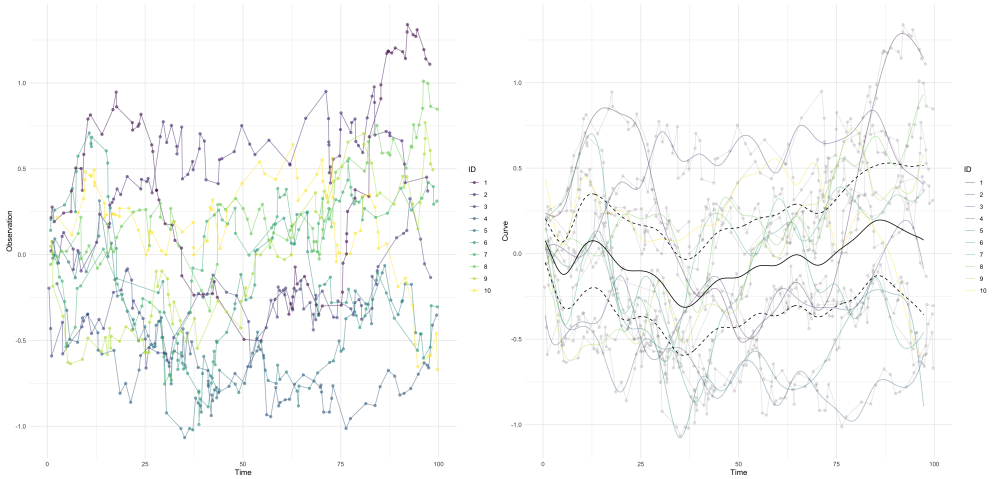

Fig 1: Example of Functional Data

1.2 Functional Principal Component

Now that we have an introduction to functional data, if we use the ideas behind principal component analysis (PCA), we have FPCA. Recall that PCA is a dimension reduction algorithm. It can reduce the dimensionality of our data by using a subset of PC's to represent our data points. The same idea applies to FPCA, reducing our data's dimensionality while keeping vital information. However, we deal with vectors in PCA, while in FPCA, we deal with functions. To compute these fpc's suppose we have n number of curves. We find the first fpc by maximizing the variance over the first fpc-score. We constrain the first fpc to one, or we could endlessly maximize by increasing the first fpc. Hence, the first fpc is the major source of variation in our data. We can find the second fpc similarly, but instead, apply an additional constraint where the second fpc is orthogonal to the first. This results in all of the fpc having to be orthogonal to one another. Let's consider a simple example of the log mortality rate of french males from 1816 to 2020 [3]. We see from Figure 2 the log mortality rate and then use FDA as explained in Section 1.1 to estimate the underlying function of our data. Plotting the first two components, we see a big increase in the mortality rate for french males between

20 to 40 years old. This hits that the mortality rate of this group varies greatly from another age group. One possibility is that the two great wars caused many young men's deaths. Now that we understand FPCA in the one-dimensional case, we can extend it to the two-dimensional as the paper has done.

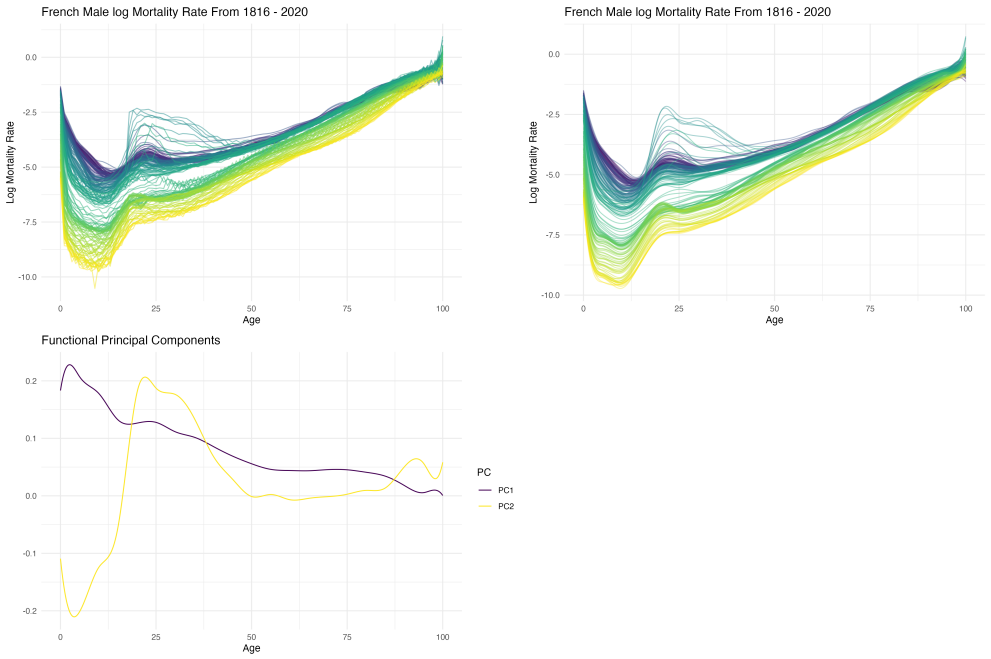

Fig 2: Example of Functional principal component analysis.

2 Summary of Paper and Methodology

The paper extends the traditional one-dimensional FPCA to a two-dimensional principal component analysis for functional data in a surface setting, where n independent realizations of a stochastic process correspond to noisy samples from the surface. For the two-dimensional functional principal component analysis, the authors use a Mercer expansion of a series of orthonormal basis functions, and the algorithm seeks to estimate the first M functional principal components by locating the minimizer of the objective function given in Equation 1.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{n_i}\sum_{j=1}^{n_i}\left\{y_{ij}^{*}-\sum_{m=1}^{M}\alpha_{im}\psi_m(s_{ij},t_{ij})\right\}^2 \tag{1}$$

This equation clearly has the form of a generic least squares estimator, where we have y_{ij}^{*} as our mean-centered noisy observation, and $\sum_{m=1}^{M}\alpha_{im}\psi_m(s_{ij},t_{ij})$ as our estimator of the two-dimensional surface, with ψ_m being the m-th basis function.

Estimation of the M principal components is undertaken in a sequential manner, conditional on the estimated values of the first m – 1 FPCs. The estimation procedure seeks to minimize the least-square error. In general, the algorithm for the first two-dimensional functional principal component has the form:

1. Specify the spline bases, including the number of bases in each direction
2. Initialize parameter values
3. Obtain the least squares estimate of α conditional on the current spline parameters. This is computed as $(\psi_{i1}^{T}\psi_{i1})^{-1}\psi_{i1}^{T}y_i^{*}$

4. Update the internal spline parameters based on the current value of α, subject to the constraint of having a norm of 1 (or rather, being a basis function)

5. Repeat Steps 3-4 Until convergence

The algorithm for estimating the second and later principal components is similar, with the caveat that previously estimated principal components are taken into account.

In order to demonstrate the efficacy of this method, the researchers applied this algorithm to estimate the two-dimensional functional principal components of a brain imaging dataset as well as the MNIST data set. Both exercises yielded interesting results by way of extracting important sources of variation from these images. In addition, the researchers used their functional principal component scores as inputs to a digit classifier and ran predictions on a subset of the MNIST dataset. When compared to a traditional classifier the accuracy was almost identical, but the computational time was reduced by 75%, even after accounting for the time taken to compute the FPC for each classified image.

3 Numerical Experimentations

In the Github repository that accompanies the paper, the author documents the code used in the examples that are presented in the paper. Taking this code an implementing it ourselves was fairly straightforward for the given use cases, so we attempted to extend it by applying the method to a new dataset as well as testing the 2DFPCA method against other more traditional methods.

3.1 Extension to New Data

When looking for a dataset on which to apply 2DFPCA, we immediately thought of some sort of spatial dataset, as it is a common use case (in addition to the image use case shown in the paper). The National Aeronautics and Space Administration (NASA) makes satellite data available dealing with many of earth's systems. We decided to focus on average July temperature. This variable could be of interest, for instance, if you were trying to model the effect of temperature on agricultural outcomes in the United States. July is directly in the middle of the growing season, and temperature can have an impact on the harvest in the fall. However, using each data point on the surface would be very difficult, since it is potentially infinite-dimensional. Using 2DFPCA, we can condense information about the July temperature across the US into coefficients on basis functions, which will be easier to estimate given a finite data set.

Data was obtained from NASA's website for the average July Temperature from July of 1980 to July of 2022 [2]. This data was given on a grid spacing of 0.625 degrees longitude and 0.5 degrees latitude. When filtered to our selected area, this resulted in over 6,000 data points for each year. After modifying the code given by the author to allow for non-square matrices to be passed to the 2DFPCA, we were able to apply 2DFPCA to the resulting temperature dataset. We used a B-Spline basis with 12 basis functions in each dimension, with the splines having order 4. The results can be seen in Figure 3.

The effects of the 2DFPCA are apparent. The top left map is granular, and shows large fluctuations in small areas. These fluctuations are smoothed out across the first three principal components as the 2DFPCA finds those bases that account for most of the variation at each step. The benefit of doing this is that now the majority of the information from the fluctuating map in the top left ins contained in the resulting 144 parameters for each principal component. Thus, if we were to use all three principal components in order to predict some outcome with this temperature data, we would only be working with 432 parameters instead of over 6,000.

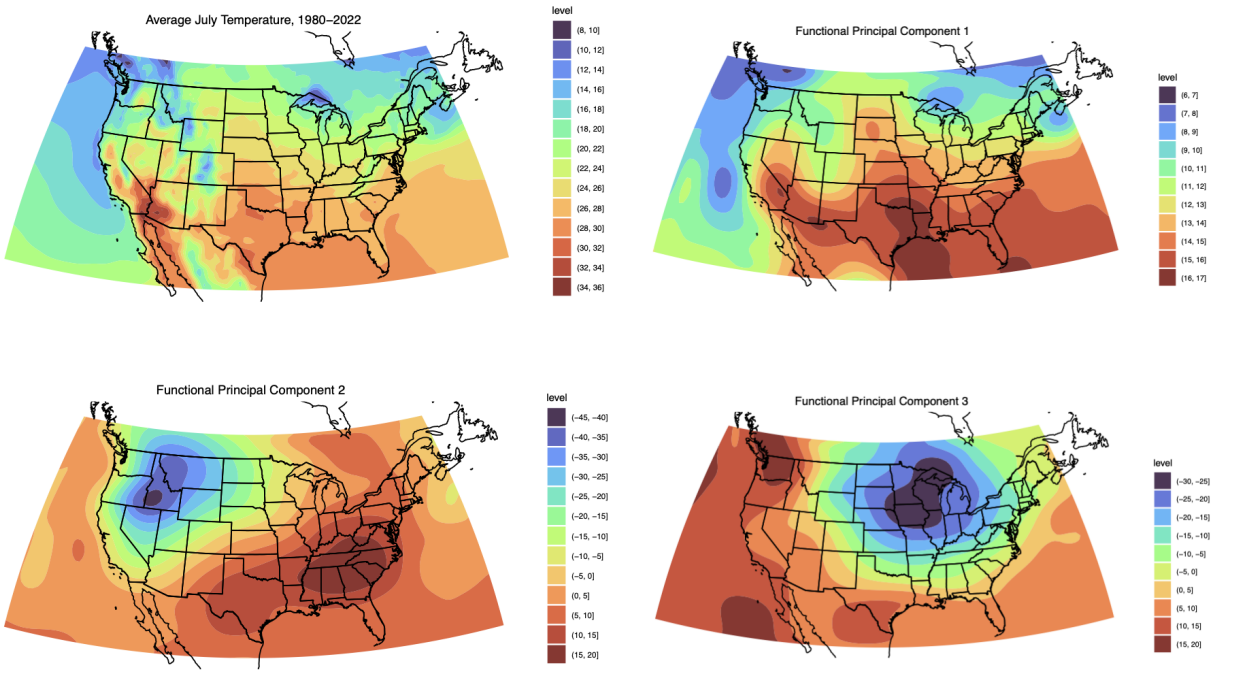

Fig 3: Application of 2DFPCA to the July Temperature Dataset. Top right shows the average over all years (no FPCA), Top left, bottom right, and bottom left show 2D functional principal components one, two, and three respectively.

3.2 Benchmarking of Computational Cost

In addition to applying this method to new data, we also wanted to test the researchers' claim that this method improved upon the existing methods in terms of computational efficiency. Repeatedly in the paper the researchers mention that their method is "faster than the conventional method" but fail to cite any references when referring to said conventional method. This may be due to the fact that the conventional method is so slow as to be infeasible. Searching online repositories for code to compute 2DFPCA did not yield anything other than the multivariate FPCA, which the authors mention only to note that it is computing a different quantity, and that the two methods are inherently different [1, 4]. We take this to mean that before this paper there was not an implemented method to compute 2DFPCA, and thus nothing to benchmark the current method against.

4 Conclusion

Imaging data is a very rich data source with numerous applications. However, the large number of pixels in an image or two-dimensional surface can result in computational challenges when attempting to use an image as a predictor of some outcome. By using two-dimensional Functional Principal Component Analysis, the high-dimensional image data can be reduced while still maintaining information about most of the variation in the image. The methodology presented in the paper and reviewed in this report is an important advancement in the field of image analysis and computation for statistical modeling.

References

[1] Happ, C., Greven, S., 2018. Multivariate functional principal component analysis for data observed on different (dimensional) domains. Journal of the American Statistical Association 113, 649–659.

[2] Office, G.M.A.A., Pawson, S., 2015. MERRA-2 tavg1_2d_slv_nx: 2d,1-Hourly,Time-Averaged,Single-Level,Assimilation,Single-Level Diagnostics V5.12.4. URL: https://disc.gsfc.nasa.gov/datacollection/M2T1NXSLV_5.12.4.html, doi:10.5067/VJAFPLI1CSIV. type: dataset.

[3] Shang, H.L., 2014. A survey of functional principal component analysis. AStA Advances in Statistical Analysis 98, 121–142. URL: http://link.springer.com/10.1007/s10182-013-0213-1, doi:10.1007/s10182-013-0213-1.

[4] Shi, H., Yang, Y., Wang, L., Ma, D., Beg, M.F., Pei, J., Cao, J., 2022. Two-dimensional functional principal component analysis for image feature extraction. Journal of Computational and Graphical Statistics 31, 1127–1140.

[5] Wang, J.L., Chiou, J.M., Müller, H.G., 2016. Functional Data Analysis. Annual Review of Statistics and Its Application 3, 257–295. URL: https://www.annualreviews.org/doi/10.1146/annurev-statistics-041715-033624, doi:10.1146/annurev-statistics-041715-033624.