Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 26 (2015) 127 – 130

Procedia Environmental Sciences

Spatial Statistics 2015: Emerging Patterns

Functional PCA for remotely sensed lake surface water temperature data

Mengyi Gong[a] *, Claire Miller[a], Marian Scott[a]

[a] *School of Mathematics and Statistics, University of Glasgow, Glasgow, G12 8QW, Scotland, U.K.*

Abstract

Functional principal component analysis is used to investigate a high-dimensional surface water temperature data set of Lake Victoria, which has been produced in the ARC-Lake project. Two different perspectives are adopted in the analysis: modelling temperature curves (univariate functions) and temperature surfaces (bivariate functions). The latter proves to be a better approach in the sense of both dimension reduction and pattern detection. Computational details and some results from an application to Lake Victoria data are presented.

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns committee

Keywords: functional PCA; dimension reduction; bivariate functions; spatial-temporal variations

1. Introduction

Remotely sensed lake surface water temperature (LSWT) data are collected in the ARC-Lake project (http://www.geos.ed.ac.uk/arclake) using (Advanced) Along Track Scanning Radiometers. These data are analyzed to assist the assessment of spatial-temporal characteristics of the ecological condition of lakes at a global scale, with the work associated to the GloboLakes project (http://www.globolakes.ac.uk). This abstract will focus on the analysis of the reconstructed LSWT data from Lake Victoria. This is a 3-dimensional data array where temperature measurements are available monthly for 2313 pixels (indexed by two geographical coordinates, longitude and

* Corresponding author.
E-mail address: m.gong.1@research.gla.ac.uk

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns Committee
doi:10.1016/j.proenv.2015.05.015

latitude) over the period of June 1995 - April 2012 (i.e. 203 months). Functional principal component analysis (FPCA) is applied to reduce the dimension of the data and to identify patterns in both time and space. Two different approaches of functional PCA are investigated. The 1-dimensional approach (1D-FPCA) treats the time series recorded in each pixel as a univariate function and performs PCA on curves; the 2-dimensional approach (2D-FPCA) views the temperature measurements at each time as a bivariate function and performs PCA on surfaces. Both methods are discussed in this abstract with proposed computational details for the 2-dimensional functional PCA explained along with the results from an application to the Lake Victoria data.

2. Methods

2.1. Functional PCA

Functional PCA refers to principal component analysis which is applied to data consisting of random functions. The idea of functional PCA is to reduce data complexity and identify dominant characteristics among functions (e.g. curve types). It is an elegant way of investigating high-dimensional data. Transforming data to functions can be regarded as the first round of dimension reduction, whereas applying a PCA on these functions serves as a second round. The dominant modes of variations can be extracted in the analysis, providing information on the spatial and temporal patterns in the data. The principal component scores (PC scores) can be used in further inference and modelling.

A functional PCA begins with a functional data representation (notation follows Ramsay and Silverman[1]). Observations from a number of N 'subjects' are treated as realizations of some smooth yet unknown functions. A basis is often used to construct these functions. With a basis of degrees of freedom K, the functional data can be approximated as $X_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t)$, where $i = 1, 2, \cdots, N$ is the index of subjects and $k = 1, 2, \cdots, K$ is the index of the basis functions. Using matrix notation, a $N \times 1$ functional data vector can be written as $X(t) = C\Phi(t)$, where C is a $N \times K$ coefficient matrix with elements c_{ik} and $\Phi(t)$ is a $K \times 1$ column vector of basis functions $\phi_k(t)$. In this setting the concept, analogous to 'variables' in a conventional PCA, becomes the realizations $X_i(t)$ evaluated at each possible value of t. Assuming zero means for all $X_i(t)$, the variance function for each pair of 'variables' $X(t_j), X(t_m)$ can be constructed as

$$V(t_j, t_m) = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{k=1}^{K} c_{ik}\phi_k(t_j) \times \sum_{k=1}^{K} c_{ik}\phi_k(t_m)\right] = \frac{1}{N}\Phi(t_j)'C'C\Phi(t_m) \quad (1)$$

The eigenproblem for functional PCA is thus

$$\int V(t_j, t)\xi(t)dt = \lambda\xi(t_j) \quad s.t. \quad \int \xi_p(t)^2 dt = 1 \quad and \int \xi_p(t)\xi_q(t)dt = 0 \quad (2)$$

where p, q are indices for eigenfunctions and $p \neq q$. Solving equation (2) requires another basis expansion, which is $\xi(t) = \sum_{k=1}^{K} b_k\phi_k(t) = b'\Phi(t)$. Define the matrix of integral $W = \int \Phi(t)\Phi(t)'dt$ and the substitution $u = W^{1/2}b$, a symmetric eigenproblem equivalent to equation (2) can be written as

$$\frac{1}{N}W^{1/2}C'CW^{1/2}u = \rho u \quad (3)$$

Note that although the argument t is usually continuous, the functions in (2) and (3) are often approximated through a fine grid, $t \in \{t_1, t_2, \cdots, t_{\mathcal{T}}\}$. Solving (3) for ρ, u and computing the reverse problem $b = W^{-1/2}u$ will give the eigenvalues and eigenfunctions (also known as principal component loadings) of the functional PCA. The principal component scores are calculated as $f_{pi} = \int \xi_p(t)X_i(t)dt$. The 1D-FPCA can be easily implemented using the R package 'fda'[6].

2.2. 2-dimensional functional PCA

Applying functional PCA to 2-dimensional data requires only a straightforward generalization. Replace the univariate basis with a bivariate basis to construct the functional data $Z_i(x,y) = \sum_{k=1}^{K} c_{ik}\phi_k(x,y)$ and update the variance function as $V(x_j, y_j; x_m, y_m) = 1/N\ \Phi(x_j, y_j)'C'C\Phi(x_m, y_m)$. The corresponding eigenproblem becomes

$$\iint V\left(x_j, y_j; x, y\right)\xi(x,y)dxdy = \lambda\xi(x_j, y_j) \tag{4}$$

Equation (4) is solved using the same approach as the one for solving equation (2). Note that this is an analysis conducted on a series of data surfaces. Hence the eigenfunctions derived from equation (4) will be bivariate functions, which summarize the common features of variations among all surfaces. Computation of 2D-FPCA makes use of the function 'pca.fd' in the package 'fda', where related code is modified to accommodate bivariate functional data. The Trapezoidal rule is proposed to approximate the double integral $W = \iint \Phi(x,y)\Phi(x,y)'dxdy$, essential to eigenproblem (4). The computation of the matrix $W^{1/2}$ follows the code in the function 'pca.fd' with the Cholesky decomposition being the main tool to obtain the matrix square root.

3. Implementation and results

An area covering the main body of Lake Victoria is investigated initially to avoid unpredictable boundary behavior induced by the irregular shape of the lake. The area is composed of 2156 lake pixels with 17 years of reconstructed monthly LSWT data available for each pixel. The strong regular seasonality displayed in the LSWT time series suggests that a Fourier basis is appropriate to model the temperature curves in 1D-FPCA. This basis has the advantage of capturing the seasonal pattern with a small degrees of freedom. However, the result is periodic eigenfunctions which cannot detect long-term temporal patterns. An alternative approach would be to use a different basis, e.g. B-spline basis of higher dimensions. However, this may not be the most practical solution, especially when the problem is extended to sparse data. Therefore, an extension to 2D-FPCA is proposed here.

A bivariate basis is used to build the temperature surfaces in the 2D-FPCA. It provides a more flexible way of extracting the spatial-temporal variations. The basis is constructed by taking the Kronecker product of two univariate B-spline bases[3], i.e. $\Phi(x,y) = \Phi_X(x)\otimes\Phi_Y(y)$. The eigenfunctions extracted from the 2D-FPCA identify the dominant spatial variations in the temperature data, whereas the PC scores can be used to examine how these spatial patterns change over time. Below are some results from a 2D-FPCA applied to Lake Victoria.

Table 1 lists the eigenvalues, along with their contributions in explaining total variance. The leading principal component explains 77.3% of the total variation, while the second PC explains 11.5%. Together they account for 88.8% of the variation, which, in general, is sufficient for most statistical analysis. Fig. 1 plots two eigenfunctions corresponding to the first two PCs and their scores. The first eigenfunction exhibits a decreasing trend from the west to the east of the lake, with the exception of the peninsula in the southeast; the second one has high loadings in the majority of the pixels, with low values appearing only in the northwest corner. The scores of the first two PCs display mainly seasonal patterns; no prominent long-term trend or change-points are evident in the plot.

Table 1. The first five eigenvalues and their corresponding variance proportions from a 2D-FPCA.

	1st	2nd	3rd	4th	5th
Eigenvalues	2.01675	0.29921	0.06885	0.05328	0.04419
Variance %	77.31%	11.47%	2.64%	2.04%	1.69%

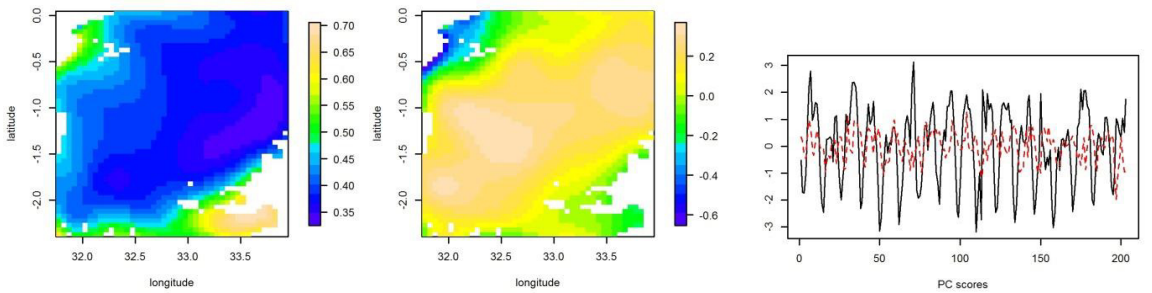

Fig. 1. (left) the first eigenfunction; (middle) the second eigenfunction; (right) scores of the first (black curve) and the second (red dashed curve) principal components from a 2D-FPCA.

4. Discussion

This abstract presents an efficient way of exploring a high-dimensional remotely sensed data set using functional PCA. Application on both univariate and bivariate functional data are discussed. Future efforts will involve developing methods to analyse real satellite measurements, where missing observations become a non-negligible problem. Conventional functional PCA may not produce robust results when it is applied to a sparse data set. According to some recent research, there are two potential solutions[2], (a) building a mixed model of random functions based on the Karhunen-Loeve expansion, then estimating the principal components using the EM algorithm based on maximum likelihood or REML[4]; (b) smoothing the raw covariance matrix using kernels and extracting principal components based on the smoothed covariance function[5]. Both methods will be investigated in the future.

Acknowledgements

Mengyi Gong is grateful to the College of Science & Engineering, University of Glasgow for PhD sponsorship. The authors gratefully acknowledge the ARC-lake project for access to the data. CM and EMS were partly funded for this work through the NERC GloboLakes project (NE/J022810/1).

References

1. Ramsay JO, Silverman BW. *Functional data analysis*. 1st ed. New York: Springer; 1997.
2. Ferraty F, Romain Y. *The oxford handbook of functional data analysis*. 1st ed. New York: Oxford University Press; 2011.
3. Ivanescu AE. A note on bivariate smoothing for two-dimensional functional data. *International Journal of Statistics and Probability* 2013;**2**:102-111.
4. James GM, Hastie TJ, Sugar CA. Principal component models for sparse functional data. *Biometrika* 2000;**87**:587-602.
5. Yao F, Muller H, Wang J. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* 2005;**100**:577-590.
6. Ramsay JO, Wickham H, Graves S, Giles H. *fda: Functional Data Analysis.* R package version 2.4.0; 2013. [Online] Available from: http://CRAN.R-project.org/package=fda