Maximal autocorrelation factors for function-valued spatial/temporal data

G. Hooker [a]**, S. Roberts** [b] **and H. L. Shang** [b]

[a]*Department of Biological Statistics and Computational Biology, Cornell University*
[b]*Research School of Finance, Actuarial Studies and Applied Statistics, Australian National University*
Email: hanlin.shang@anu.edu.au

Abstract: Dimension reduction techniques play a key role in analyzing functional data that possess temporal or spatial dependence. Of these dimension reduction techniques functional principal components analysis (FPCA) remains a popular approach. Functional principal components extract a set of latent components by maximizing variance in a set of dependent functional data. However, this technique may fail to adequately capture temporal or spatial autocorrelation.

Functional maximum autocorrelation factors (FMAF) are proposed as an alternative for modeling and forecasting temporally or spatially dependent functional data. FMAF find linear combinations of the original functional data that have maximum autocorrelation and that are decreasingly predictable functions of time. We show that FMAF can be obtained by searching for the rotated components that have the smallest integrated first derivatives. Through a basis function expansion, a set of scores are obtained by multiplying the extracted FMAF with the original functional data. Autocorrelation in the original functional time series is manifested in the autocorrelation of these scores derived.

Through a set of Monte Carlo simulation results, we study the finite-sample properties of the proposed FMAF. Wherever possible, we compare the performance between FMAF and FPCA. In an enhanced vegetation index data from Harvard Forest we apply FMAF to capture temporal or spatial dependency.

Keywords: Autocorrelation operator, Functional time series, Spatially dependent functional data, Linear dimension reduction technique

1 INTRODUCTION

Recent advances in computer technology facilitates the analysis of functional data, represented in the forms of curves, images and shapes. Since the pioneering work by Deville (1974), and more recently with the work by Ramsay and Silverman (2005) and Ferraty and Vieu (2006), the statistical community has shown an increasing interest in developing novel methods for analyzing functional data. A wide range of statistical tools have been extended and modified to analyze functional data, examples of such methods include: exploratory and descriptive data analysis (Ramsay and Silverman, 2005), nonparametric methods (Ferraty and Vieu, 2006), multivariate analysis (Goulard and Voltz, 1993), regression (Morris, 2015), time series analysis (Horváth and Kokoszka, 2012) and spatial statistics (Delicado et al., 2010).

It is common that data possess both spatial and functional attributes, such as in agronomy (Chan et al., 2006) and climatology (Ramsay and Silverman, 2005). Geostatistical methods can be adapted to this type of problem and the data modelled to allow for both spatial and functional attributes. This modeling approach can be useful to predict functions based on spatially observed reference curves. Goulard and Voltz (1993) were the first to apply geostatistical interpolation methods to predict functions at non-observed sites. They considered three methods: two were based on multivariate data analysis using cokriging and the other on functional kriging to predict curves directly. Using the observed data points, Goulard and Voltz (1993) fitted a presumedly known parametric model with a small number of parameters to reconstruct a set of curves.

In this paper, we overcome the parametric distributional assumption of Goulard and Voltz's (1993) method and the limitation of a small number of observed points per function by proposing a nonparametric fitting procedure. We develop a "most constant spatial direction" in which we minimize the difference between the current point and those around it. Our procedure implements a functional linear regression, where the regression coefficients can be estimated by a priori basis function or data-driven basis function expansion. The methodology is suitable for analyzing both temporally or spatially dependent functional data.

In Section 2, we present a review of functional principal component analysis and put forward the functional maximum autocorrelation factors (FMAF). Using a simulation study, we evaluate and compare the performance between the methods in Section 3. In Section 4, we apply the FMAF to a set of spatially dependent functional data involving enhanced vegetation index in Harvard Forests. Section 5 concludes.

2 METHODOLOGY

2.1 Functional principal component analysis (FPCA)

The use of FPCA dates back to the early 1940s when Karhunen (1946) and Loéve (1946) independently developed a theory on the optimal series expansion of a continuous stochastic process, and extended eigenanalysis from a symmetric matrix to integral operators with symmetric kernels. Later, Rao (1958) provides an application of the Karhunen-Loéve (KL) expansion to functional data, by applying multivariate PCA to observed functional values. For a detailed overview on FPCA consult a review article by Shang (2014).

Let $\mathcal{X}$ be a random variable $\mathcal{X} : \Omega \to L^2(\mathcal{I})$, such that $\mathcal{X} \in L^2(\Omega)$. $\mathcal{X}$ can also be seen as a stochastic process defined on a compact set $\mathcal{I}$, with finite variance $\int_{\mathcal{I}} \mathrm{E}(\mathcal{X}^2) < \infty$. Let μ be the mean function of $\mathcal{X}$, without lose of generality, let $\mathcal{X}^c = \mathcal{X} - \mu$ be a centred stochastic process. The covariance operator of $\mathcal{X}$ is defined to be the function $K : \mathcal{I} \times \mathcal{I} \to R$, such that

$$K(u,v) = \mathrm{Cov}(\mathcal{X}(\mu), \mathcal{X}(v)) = \mathrm{E}[\mathcal{X}^c(\mu))\mathcal{X}^c(v)] \qquad (1)$$

Through Mercer's lemma, the covariance operator K can be decomposed into

$$K(\mu,v) = \sum_{k=1}^{\infty} \lambda_k \phi_k(u)\phi_k(v) \qquad (2)$$

With Karhunen-Loéve expansion, a stochastic process $\mathcal{X}$ observed at t can be expressed as

$$\mathcal{X}_t^c(\mu) = \sum_{k=1}^{\infty} \sqrt{\lambda_k}\xi_{t,k}\phi_k(\mu) = \sum_{k=1}^{\infty} \beta_{t,k}\phi_k(u) \qquad (3)$$

where $\xi_k = 1/\sqrt{\lambda_k} \int_{\mathcal{I}} \mathcal{X}^c(v)\phi_k(v)dv$ is an uncorrelated random variable with zero mean and unit variance.

2.2 Functional maximum autocorrelation factors (FMAF)

In a multivariate data analysis, the method of maximum autocorrelation factors (MAF) was first introduced by Switzer and Green (1984) for the analysis of multivariate imagery data. Shapiro and Switzer (1989) adapted this method to analyze gridded multivariate data. The MAF extract a set of latent factors that are linear combinations of the data, in which autocorrelation decreases monotonically from the first factor to the last. Hence, the first factor extracts the linear combination of the variables that is the most continuous in time.

There has been a surge interest in the theoretical properties and practicality of MAF (Gallagher et al., 2014). In fisheries, Erzini (2005) investigated fishery catches using multiple time series of environmental parameters and compared MAF with dynamic factor analysis. In image interpolation, Henderson et al. (2009) compares principal component analysis (PCA) with MAF and conclude that MAF is more effective than PCA for the analysis of high signal intensity data.

We extend the MAF to the functional data context by incorporating regularisation methods to deal with infinite-dimensionality nature of functional data. We aim to find $\beta(t)$ that maximizes

$$\text{Cor}\left[\int \beta(t)\mathcal{X}_i(t)dt, \int \beta(t)\mathcal{X}_{i+1}(t)dt\right] \tag{4}$$

and we can take successive direction $\beta_1, \beta_2, \ldots$ to maximum autocorrelation subject to

$$\text{Cor}\left[\int \beta_j(t)\mathcal{X}_i(t)dt, \int \beta_k(t)\mathcal{X}_i(t)dt\right] = \delta_{jk} \tag{5}$$

where δ_{jk} denotes an indicator function; when $j = k$, $\delta_{jk} = 1$, when $j \neq k$, $\delta_{jk} = 0$. Maximizing autocorrelation in (4) is equivalent to minimize

$$\frac{\sum_{i=2}^{n}\int[\beta(t)(\mathcal{X}_i(t) - \mathcal{X}_{i+1}(t))]^2dt}{\sum_{i=1}^{n}\int[\beta(t)(\mathcal{X}_i(t) - \bar{\mathcal{X}}(t))]^2dt} = \frac{\int\int\beta(t)\Sigma_1(t,s)\beta(s)dsdt}{\int\int\beta(t)\Sigma_0(t,s)\beta(s)dsdt} \tag{6}$$

$$= \frac{\int\int\beta(t)\left\{\sum_{i=2}^{n}[\mathcal{X}_i(t) - \mathcal{X}_{i+1}(t)][\mathcal{X}_i(s) - \mathcal{X}_{i+1}(s)]\right\}\beta(s)dsdt}{\int\int\beta(t)\left\{\sum_{i=1}^{n}[\mathcal{X}_i(t) - \bar{\mathcal{X}}(t)][\mathcal{X}_i(s) - \bar{\mathcal{X}}(s)]\right\}\beta(s)dsdt} \tag{7}$$

where $\Sigma_0(t,s)$ represents the variance operator, and $\Sigma_1(t,s)$ represents the covariance operator.

Following the early work by Ramsay and Silverman (2005), we represent $\mathcal{X}_i(t) - \bar{\mathcal{X}}(t) = \sum_{k=1}^{K}B_k(t)c_{ik}$, where $B_k(t)$ is a basis system and let $\boldsymbol{C}$ be the matrix containing the c_{ik} with $S_{jk} = \int B_j(t)B_k(t)dt$. Now, we can represent the denominator of (7) as

$$\int\int\beta(t)\left\{\sum_{i=1}^{n}[\mathcal{X}_i(t) - \bar{\mathcal{X}}(t)][\mathcal{X}_i(s) - \bar{\mathcal{X}}(s)]\right\}\beta(s)dsdt = \boldsymbol{b}^{\top}\boldsymbol{S}\boldsymbol{C}\boldsymbol{C}^{\top}\boldsymbol{S}\boldsymbol{b} \tag{8}$$

where $\boldsymbol{b} = (b_1, \ldots, b_K)^{\top}$ and $\boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K)^{\top}$. Similarly, we can represent the numerator of (7) as

$$\int\int\beta(t)\sum_{i=2}^{n}[\mathcal{X}_i(t) - \mathcal{X}_{i+1}(t)][\mathcal{X}_i(s) - \mathcal{X}_{i+1}(s)]\beta(s)dsdt = \boldsymbol{b}^{\top}\boldsymbol{S}\boldsymbol{C}^{\top}\boldsymbol{E}\boldsymbol{E}^{\top}\boldsymbol{C}^{\top}\boldsymbol{S}\boldsymbol{b}, \tag{9}$$

where $\boldsymbol{E}$ is the matrix that gives the differences between rows of $\boldsymbol{S}$. We can achieve this if $\boldsymbol{E}$ is a $n \times (n-1)$ matrix with $E_{i,i} = 1$ and $E_{i+1,i} = -1$ and all other entries being 0.

Since the denominator of (7) may not be invertible, we can add regularisation by using a roughness penalty (see Silverman, 1996, for details). That is,

$$\frac{\int\int\beta(t)\Sigma_1(t,s)\beta(s)dsdt}{\int\int\beta(t)\Sigma_0(t,s)\beta(s)dsdt + \lambda\int\beta''(t)^2dt} \tag{10}$$

for which the minimum (in the basis-expansion expression above) is the largest eigenvalue of

$$(\boldsymbol{S}\boldsymbol{C}\boldsymbol{C}^{\top}\boldsymbol{S} + \lambda\boldsymbol{P})^{-1}\boldsymbol{S}\boldsymbol{C}\boldsymbol{E}\boldsymbol{E}^{\top}\boldsymbol{C}^{\top}\boldsymbol{S}, \tag{11}$$

where $P_{ij} = \int B_i''(t)B_j''(t)dt$, λ represents the roughness penalty parameter and $''$ represents the second derivative. Note that we have inverted the "natural" entries for the numerator and denominator matrices, as we want to regularize the eigenfunction associated with the smallest eigenvalue of $\Sigma_1(t, s)$.

To select optimal λ, we consider minimizing predicted squared error $\mathcal{X}_{i+1}$ based on $\mathcal{X}_i$, given by

$$G(\lambda) = \sum_{i=2}^{n} \int \left(\mathcal{X}_i(t) - \sum_{k=1}^{K} \left[\int \beta_k^{-i}(s)\mathcal{X}_{i-1}(s)ds \right] \beta_k^{-i}(t) \right)^2 dt \tag{12}$$

where $\beta_k^{-i}(t)$ is estimated without the ith curve, and K denotes the number of retained components.

From equations (4) to (12), we express FMAF for one-dimensional functional time series. However, FMAF can also be extended for two-dimensional functional images. FMAF aims to minimize the difference between current spatial point and those around it, this is given by

$$\int \left[\beta(t) \left(\frac{\mathcal{X}_{i-1,j}(t) + \mathcal{X}_{i+1,j}(t) + \mathcal{X}_{i,j-1}(t) + \mathcal{X}_{i,j+1}(t)}{4} - \mathcal{X}_{i,j}(t) \right) \right]^2 dt, \tag{13}$$

where $\mathcal{X}_{i,j}(t)$ represents a two-dimensional functional image.

3 A SIMULATION STUDY

To examine the effects of using FMAF on functional auto-regressive time series, we set up the following simulation. Functional time series are generating according to the process

$$\mathcal{X}_{t+1}(t) = \int_0^1 B(s,t)\mathcal{X}_i(s)ds + \epsilon_i(t).$$

We represented each of the $\epsilon_i(t)$, $B(s,t)$ and $\mathcal{X}_i(t)$ by a linear combination of 25 Fourier basis functions using the system $\phi_1(t) = 1$, $\phi_{2j}(t) = \sin(2^j\pi t)$, $\phi_{2j+1} = \cos(2^j\pi t)$, $j = 1, 2, \ldots, 12$. Here, the $\epsilon_i(t)$ are generated as $\epsilon_i(t) = \sum_{i=1}^{25} c_{ij}\phi_j(t)$, where c_{ij} are independently generated as $N\left(0, 0.2e^{-(i-1)/4}\right)$. This has the effect of emphasizing low-period functions with decaying variance associated with higher periods (see also Hall and Hooker, 2015).

We specified $B(s,t)$ in terms of

$$B(s,t) = \sum_{j=1}^{25} \sum_{k=1}^{25} b_{ij}\phi_j(t)\phi_k(t)$$

where $b_{ij} = (2\pi)^{-1/2} \exp(-(i+j-2)/4) \exp(-(i-j)^2/2)$. That is lower-frequency components are more persistant and there is some "bleed" between neighbouring frequencies. With this specification, the $\mathcal{X}_i(t)$ can be generated explicitly. A plot of $B(s,t)$ and a plot of simulated functional time series are given in Figure 1.
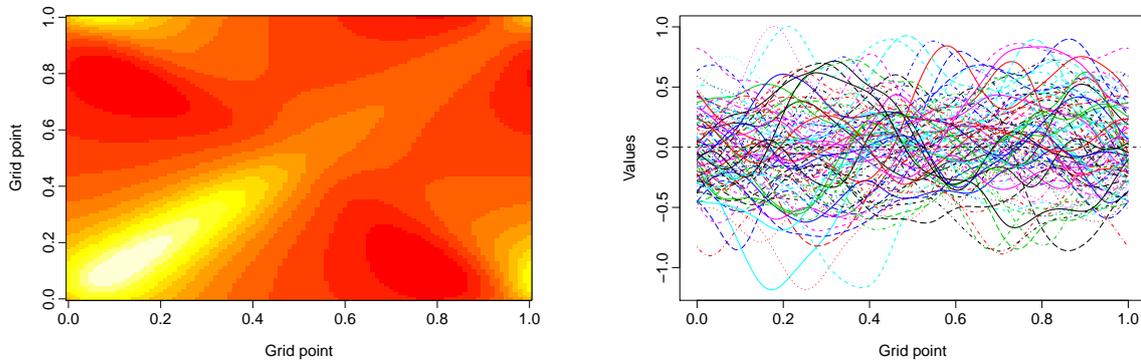
Figure 1. Plots of simulated regression coefficient function $B(s,t)$ and simulated functional time series. The regions in red and yellow show a strong and a weak spatial autocorrelation, respectively

In the top panel of Figure 2, we plot the first five latent components extracted by FPCA and FMAF, respectively. These five functional principal components explain around 93% of total variation in the original functional time series. By multiplying these latent components with their corresponding scores, a reconstruction of the original functional time series is obtained for both FPCA and FMAF (see the bottom panel of Figure 2).
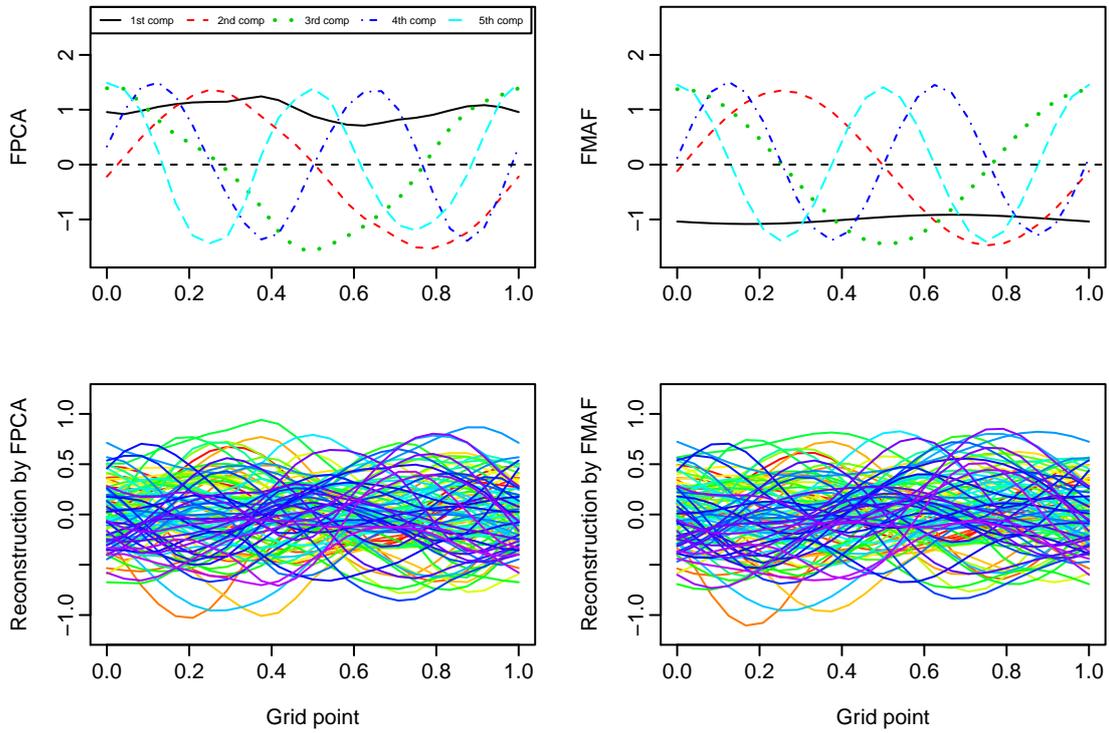
Figure 2. The first five latent components extracted by FPCA and FMAF, and their reconstructions of the original functional time series

While the first five retained components explain a large amount of variations in the original functional time series, we further investigate the reconstruction errors for FPCA and FMAF under different numbers of components. We found that the FMAF produce comparably smaller reconstruction errors than did FPCA.

Table 1. A comparison of reconstruction errors, as measured by root mean squared error (RMSE) and mean absolute error (MAE), between the FPCA and FMAF for different numbers of retained components. The smaller forecast errors are highlighted in bold

Number of retained components	Percentage of variance explained	FPCA		FMAF	
		RMSE	MAE	RMSE	MAE
1	0.3984	**0.2616**	**0.2086**	0.2623	0.2095
5	0.9225	0.0990	0.0790	**0.0972**	**0.0770**
10	0.9938	0.0427	0.0354	**0.0376**	**0.0295**
15	0.9996	0.0339	0.0304	**0.0270**	**0.0208**
20	1.0000	0.0333	0.0300	**0.0263**	**0.0202**
25	1.0000	0.0333	0.0299	**0.0262**	**0.0201**

4   REMOTE SENSING DATA ANALYSIS

The FMAF are also motivated by the spatial autocorrelation observed in the Harvard Forest vegetation index data, previously studied by Liu et al. (2012) and Liu et al. (2014). This data set consists of time series of remotely sensed images acquired over a site in central Massachusetts for studying the enhanced vegetation index (EVI) at Harvard Forest. EVI is constructed from surface reflectance measurements obtained from Moderate Resolution Imaging Spectroradiometer onboard NASA's Tarra and Aqua satellites. Gridded data were extracted

for a 25 by 25 pixel window which covers approximately 134km^2 area centred over the Harvard Forest Long Term Experimental Research site in Petershan, MA. Data are provided at 8-day interval (46 data points per year) from the period from January 1, 2001 to December 31, 2006. Among 625 pixels, there are 23 missing observations. Averaged over six years, we obtain a (46×602) smoothed discretized functional data and their centered values shown in Figure 3.
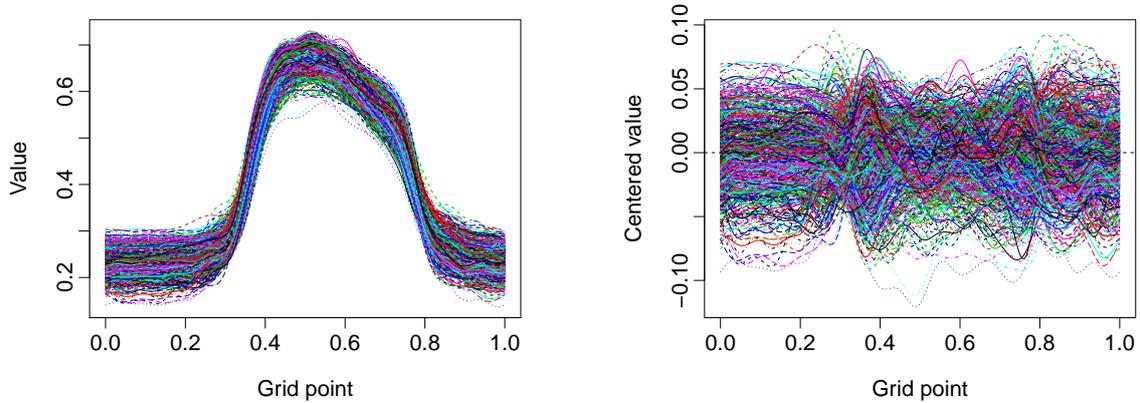
Figure 3. EVI raw and centred functional data

In Figure 4, we plot the first five extracted FMAF and their spatial autocorrelation captured by the FMAF, in the first and last directions. Spatial autocorrelation is well captured in the first spatial direction, whereas it seems to be independent and identically distributed random noise in the 25th spatial direction. By contrast, the latent components extracted by FPCA seem to be less satisfactory.
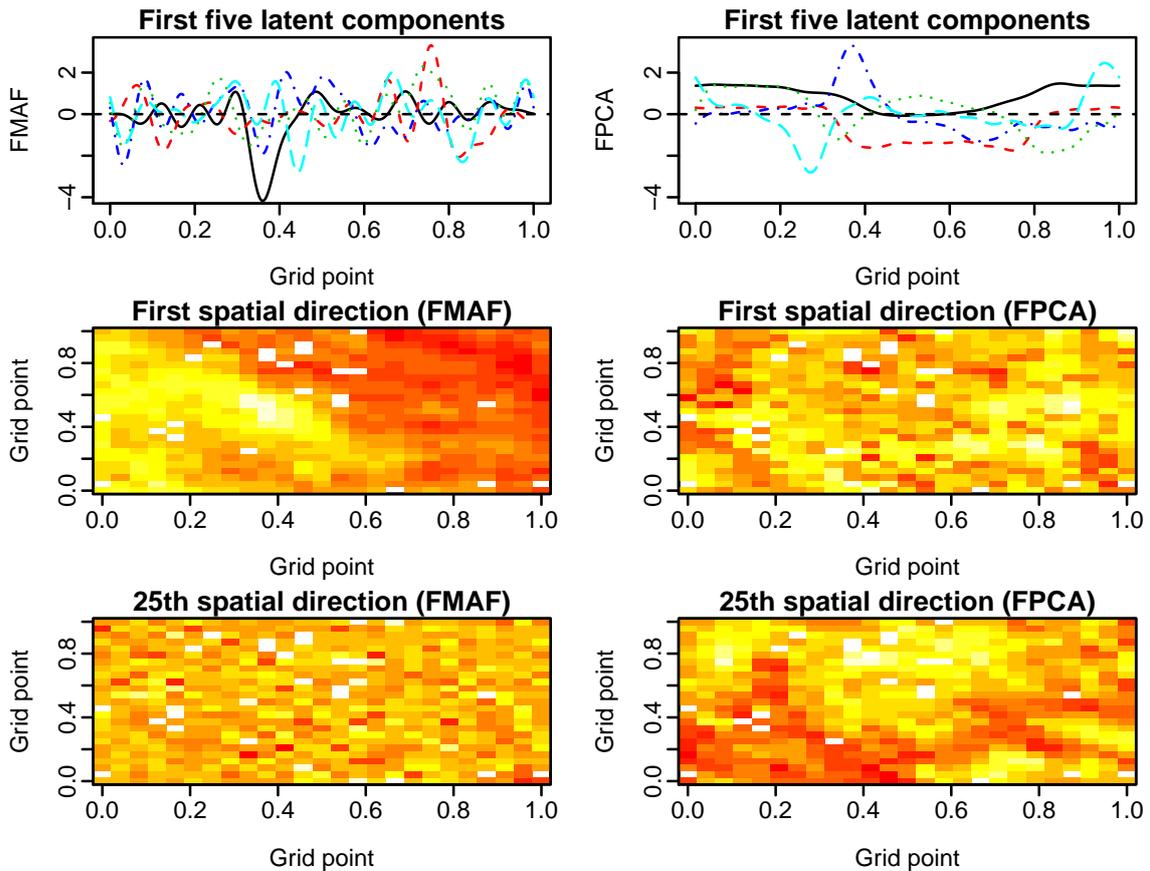
Figure 4. The first five extracted FMAF and their spatial autocorrelation captured by the FMAF in the first and last directions. Note that the 23 missing observations are manifested by the white square regions in the plot

5 CONCLUSIONS

We introduce functional maximum autocorrelation factors, which extract latent components by maximizing autocorrelation for a set of dependent functional data that possess temporal or spatial dependency. Through a simulation study, we show the functional maximum autocorrelation factors produce smaller reconstruction errors than the ones by functional principal component analysis. Using the enhanced vegetation index data set from Harvard Forest, we demonstrate how functional maximum autocorrelation factors are capable of capturing two-dimensional spatial autocorrelation. Thus, it should be considered as a part of toolbox for analyzing ever-increasingly high-dimensional data.

REFERENCES

Chan, K., A. Oates, A. Swan, R. Hayes, B. Dear, and M. Peoples (2006). Agronomic consequences of tractor wheel compaction on a clay soil. *Soil & Tillage Research 89*, 13–21.

Delicado, P., R. Giraldo, C. Comas, and J. Mateu (2010). Statistics for spatial functional data: some recent contribution. *Environmetrics 21*, 224–239.

Deville, J.-C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'inséé 15*, 3–101.

Erzini, K. (2005). Trends in ne atlantic landings (southern portugal): Identifying the relative importance of fisheries and environmental variables. *Fisheries Oceanography 14*(3), 195–209.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis*. New York: Springer Science+Business Media, Inc.

Gallagher, N. B., J. M. Shaver, R. Bishop, R. T. Roginski, and B. M. Wise (2014). Decompositions using maximum signal factors. *Journal of Chemometrics 28*, 663–671.

Goulard, M. and M. Voltz (1993). Geostatistical interpolation of curves: a case study in soil science. In *Geostatistics Tróia*, pp. 805–816. Dordrecht: Kluwer.

Hall, P. and G. Hooker (2015). Truncated linear models for functional data. *Journal of the Royal Statistical Society, Series B in press*.

Henderson, A., J. S. Fletcher, and J. C. Vickerman (2009). A comparison of PCA and MAF for ToF-SIMS image interpolation. *Surface and Interface Analysis 41*(8), 666–674.

Horváth, L. and P. Kokoszka (2012). *Inference for Functional Data with Applications*. New York: Springer.

Karhunen, K. (1946). Zur spektraltheorie stochastischer prozesse. *Annales Academiae Scientiarum Fennicae 37*, 1–37.

Liu, C., S. Ray, and G. Hooker (2014). Functional principal components analysis of spatially correlated data. Technical report, Cornell University.

Liu, C., S. Ray, G. Hooker, and M. Friedl (2012). Functional factor analysis for periodic remote sensing data. *The Annals of Applied Statistics 6*(2), 601–624.

Loéve, M. (1946). Fonctions aléatoires a decomposition orthogonale exponentielle. *La Revue Scientifique 84*, 159–162.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and its Applications 2*, 321–359.

Ramsay, J. and B. Silverman (2005). *Functional Data Analysis* (2nd ed.). New York: Springer Series in Statistics.

Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics 14*(1), 1–17.

Shang, H. L. (2014). A survey of functional principal component analysis. *AStA Advances in Statistical Analysis 98*(2), 121–142.

Shapiro, D. E. and P. Switzer (1989). Extracting time trends from multiple monitoring sites. Technical report 132, Stanford University.

Silverman, B. (1996). Smoothed functional principal components analysis by choice of norms. *The Annals of Statistics 24*, 1–24.

Switzer, P. and A. A. Green (1984). Min/max autocorrelation factors for multivariate spatial imaging. Technical report 6, Stanford University.