UNIVERSITY OF GLASGOW

Statistical Methods for Sparse Image Times Series of Remote-sensing Lake Environmental Measurements

Mengyi Gong

A thesis submitted to the University of Glasgow
for the degree of Doctor of Philosophy

Supervised by
Dr. Claire Miller and Prof. Marian Scott

in the
School of Mathematics and Statistics
College of Science and Engineering

November 2017

Declaration of Authorship

I, Mengyi Gong, declare that this thesis titled, 'Statistical Methods for Sparse Image Times Series of Remote-sensing Lake Environmental Measurements' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Part of the work in Chapter 2 has been presented as a poster in the Spatial Statistics 2015 conference under the theme 'Emerging Patterns' in Avignon, France. A long abstract, titled 'Functional PCA for remotely sensed lake surface water temperature data', was published in the conference proceeding Procedia Environmental Sciences (Volume 26). The work in Chapter 3 has been presented in the 26th Annual Conference of the International Environmetrics Society in 2016 in Edinburgh, U.K. Part of the work in Chapters 4 and 5 has been presented in the Spatial Statistics 2017 conference, under the theme 'One World: One Health' in Lancaster, U.K.

Signed: _____

Date: _____

'Houston, we've had a problem.'

Apollo 13 Mission

Abstract

Remote-sensing technology is widely used in Earth observation, from everyday weather forecasting to long-term monitoring of the air, sea and land. The remarkable coverage and resolution of remote sensing data are extremely beneficial to the investigation of environmental problems, such as the state and function of lakes under climate change. However, the attractive features of remote-sensing data bring new challenges to statistical analysis. The wide coverage and high resolution means that data are usually of large volume. The orbit track of the satellite and the occasional obscuring of the instruments due to atmospheric factors could result in substantial missing observations. Applying conventional statistical methods to this type of data can be ineffective and computationally intensive due to its volume and dimensionality. Modifications to existing methods are often required in order to incorporate the missingness. There is a great need of novel statistical approaches to tackle these challenges.

This thesis aims to investigate and develop statistical approaches that can be used in the analysis of the sparse remote-sensing image time series of environmental data. Specifically, three aspects of the data are considered, (a) the high dimensionality, which is associated with the volume and the dimension of data, (b) the sparsity, in the sense of high missing percentages and (c) the spatial/temporal structures, including the patterns and the correlations.

Initially, methods for temporal and spatial modelling are explored and implemented with care, e.g. harmonic regression and bivariate spline regression with residual correlation structures. In recognizing the drawbacks of these methods, functional data analysis is employed as a general approach in this thesis. Specifically, functional principal component analysis (FPCA) is used to achieve the goal of dimension reduction. Bivariate basis functions are proposed to transform the satellite image data, which typically consists of thousands/millions of pixels, into functional data with low dimensional representations. This approach has the advantage of identifying spatial variation patterns through the principal component (PC) loadings, i.e. eigenfunctions. To overcome the high missing percentages that might invalidate the standard implementation of the FPCA, the mixed model FPCA (MM-FPCA) was investigated in Chapter 3. Through estimating the PCs using a mixed effect model, the influence of sparsity could be accounted for appropriately. Data imputation can be obtained from the

fitted model using the (truncated) Karhunen-Loéve expansion. The method's applicability to sparse image series is examined through a simulation study.

To incorporate the temporal dependence into the MM-FPCA, a novel spatio-temporal model consisting of a state space component and a FPCA component is proposed in Chapter 4. The model, referred to as SS-FPCA in the thesis, is developed based on the dynamic spatio-temporal model framework. The SS-FPCA exploits a flexible hierarchical design with (a) a data model consisting of a time varying mean function and random component for the common spatial variation patterns formulated as the FPCA, (b) a process model specifying the type of temporal dynamic of the mean function and (c) a parameter model ensuring the identifiability of the model components. A 2-cycle alternating expectation - conditional maximization (AECM) algorithm is proposed to estimate the SS-FPCA model. The AECM algorithm allows different data augmentations and parameter combinations in various cycles within an iteration, which in this case results in analytical solutions for all the MLEs of model parameters. The algorithm uses the Kalman filter/smoother to update the system states according to the data model and the process model. Model investigations are carried out in Chapter 5, including a simulation study on a 1-dimensional space to assess the performance of the model and the algorithm. This is accompanied by a brief summary of the asymptotic results of the EM-type algorithm, some of which can be used to approximate the standard errors of model estimates.

Applications of the MM-FPCA and SS-FPCA to the remote-sensing lake surface water temperature and Chlorophyll data of Lake Victoria (obtained from the European Space Agency's Envisat mission) are presented at the end of Chapter 3 and 5. Remarks on the implications and limitations of these two methods are provided in Chapter 6, along with the potential future extensions of both methods. The Appendices provide some additional theorems, computation and derivation details of the methods investigated in the thesis.

Acknowledgements

First of all, I would like to express my gratitude to my supervisors, Dr. Claire Miller and Prof. Marian Scott. Thank you for your guidance throughout the past four years. Thank you for encouraging me to explore the wonderful world of statistics and for leading me back to the pathway every time I got too obsessed with the flowers at the roadside (my pony and flower problem). You are the best I can ever have for my PhD!

Special acknowledgment is given to the College of Science and Engineering, University of Glasgow, for sponsoring my PhD. It was my honour to be awarded the College Scholarship, without which this PhD would not be possible.

Acknowledgment is also given to the GloboLakes project, the ARC-Lake project, Plymouth Marine Laboratory and Biological and Environmental Sciences, University of Stirling for providing the remote-sensing lake data used in this thesis.

A big 'thank you' goes to all staffs and colleagues in the old and the new Maths buildings. In particular, to my lovely office mates in library 321a, Ameneh, Linda, Reem, Jorge; in the basement office 120, Kelly, Amira (I was lucky to have you around all the way through my PhD), Guowen (It was fun talking about stats with you), Craig, Cunyi, Aisyah, George, Daniel; Francesca in office 228 in the new building and Qingying (my 'quasi' office mate, though never be in the same office). Thank you for all your company! To Duncan and Adrian, for the helpful discussions and advice during the annual reviews; to Ludger, for giving me the chance to try stats tutoring; to Ruth, for all the helps on those amazing yet sometimes annoying lakes.

Also to Matthew and Calum, who taught me a little something other than statistics; to Qi in the Chemistry building next door, who is also my caring neighbour in Anniesland.

Finally, to my family and friends in Shanghai, especially to my parents for the constant support over the years!

Contents

List of Figures

List of Tables

Abbreviations

LSWT	**L**ake **S**urface **W**ater **T**emperature
Chl	**Chl**orophyll a
AATSR	**A**dvanced **A**long-**T**rack **S**canning **R**adiometer
MERIS	**ME**dium-spectral **R**esolution **I**maging **S**pectrometer
ESA	**E**uropean **S**pace **A**gency
GAMM	**G**eneralized **A**dditive **M**ixed **M**odel
FDA	**F**unctional **D**ata **A**nalysis
PC	**P**rincipal **C**omponent
FPCA	**F**unctional **P**rincipal **C**omponent **A**nalysis
GRF	**G**aussian **R**andom **F**ield
DSTM	**D**ynamic **S**patio-**T**emporal **M**odel
STRE	**S**patio-**T**emporal **R**andom **E**ffect model
KF	**K**alman **F**ilter
KS	**K**alman **S**moother
FRF	**F**ixed **R**ank **F**iltering
MLE	**M**aximum **L**ikelihood **E**stimator
EM	**E**xpectation - **M**aximization
ECM	**E**xpectation - **C**onditional **M**aximization
AECM	**A**lternating **E**xpectation - **C**onditional **M**aximization
IMSE	**I**ntegrated **M**ean **S**quared **E**rror
RSS	**R**esidual **S**um of **S**quares
i.i.d.	**i**ndependent and **i**dentically **d**istributed
MM-FPCA	**M**ixed **M**odel - **F**unctional **P**rincipal **C**omponent **A**nalysis
SS-FPCA	**S**tate **S**pace - **F**unctional **P**rincipal **C**omponent **A**nalysis

Chapter 1

Introduction

Remote-sensing technology is widely used in Earth observation, from everyday weather forecasting to long-term monitoring of the air, sea and land. '*The objective and continuous views of our planet supplied by satellite images and data provide scientists and decision makers with the information they need to understand and protect our environment*' (European Space Agency (ESA) Earth Observation Mission, https://earth.esa.int/web/guest/missions). The remarkable coverage and resolution of remote sensing data are extremely beneficial in the investigation of the impacts of environmental change, especially for those inaccessible remote areas on Earth.

In 2002, ESA launched its Earth observation mission, Envisat. It was ESA's successor to the European Remote Sensing satellite, which was retired in 2001. With 10 instruments aboard and at eight tons, Envisat was the largest civilian Earth observation mission. The advanced radio/spectrometers on board were designed to measure the ocean surface temperature, atmospheric ozone, wind fields, land features, etc (https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/envisat). Unfortunately, the satellite lost contact with the Earth in May 2012, thus ending the mission. However, during its ten-years' mission, Envisat has provided scientists with some of the most valuable observations and a novel source of information for understanding environmental change. Its successors, Sentinel-1, 2 and 3, were launched between 2014 and 2017, continuing the mission of Earth observation.

The appealing features of remote-sensing data bring new problems to the processing and modelling of data. The wide coverage and high resolution means the data are usually of large volume. The occasional obscuring of the Earth due to cloud cover means that data

can be missing from time to time. Therefore, conventional statistical methods may not be appropriate for this new source of data and there is a great demand for novel approaches to the analysis of the remote-sensing data. This thesis develops statistical methods to address these challenges. The research is motivated by remote-sensing image time series data of lakes across the world obtained by two of the radio/spectrometers on board the Envisat, Advanced Along-Track Scanning Radiometer (AATSR) and the Medium-spectral Resolution Imaging Spectrometer (MERIS).

1.1 Remote-sensing measurements of lakes

As described in the overview of the Globolakes project (http://www.globolakes.ac.uk), '*the Earth's freshwater ecosystems are vital components of the global biosphere, yet they are vulnerable to the forces of climate and human induced change*'. So far, peoples' understanding of lakes' response to these changes and their impacts on the status of lakes are still limited. Recent developments in remote-sensing and data retrieval technology provide an opportunity to study the ecological condition of lakes from a brand-new perspective. Scientists are interested in the study of various remote-sensing measurements of lake ecology, such as lake surface water temperature (LSWT) and Chlorophyll a (Chl). LSWT reflects the physical dynamics of lakes. The data are retrieved from the measurements of AATSR for 2002 onward and ATSR (the predecessor of AATSR) prior to 2002 (MacCallum & Merchant, 2013). Both ATSR and AATSR are imaging multi-spectral radiometers, primarily designed to measure sea surface temperature (SST) and the spatial resolution of the infra red ocean channels is 1km $\times$ 1km (Hout *et al.*, 2001) (here 'km' stands for kilometer). Chlorophyll a is an indicator of lake ecosystem condition and change. The data are retrieved from the measurements of MERIS (Doerffer & Schiller, 2008), a programmable, medium-spectral resolution, imaging spectrometer operating in the solar reflective spectral range. The spatial resolution of the ocean channels is 1040m $\times$ 1200m (here 'm' stands for meter); that of the land and coast channels is 260m $\times$ 300m (Hout *et al.*, 2001). The next two subsections provide a detailed description of the LSWT and Chl data.

1.1.1 Lake surface water temperature and Chlorophyll data

First note that the phrase 'remote-sensing data' in this thesis refers to the satellite processed data, such as the LSWT and Chl data. They are different from the satellite raw

measurements, which are often recorded as intensity of the radiance per unit area. These raw measurements are transformed into 'remote-sensing data' using advanced retrieval algorithms, which associates the radiation measurements with the reflectance characteristics of different objects on Earth. For example, the SST retrieval algorithm requires a radiative transfer model, accompanied by observed radiance and other calibration data, to define the optimal retrieval coefficients (Merchant & Le Borgne, 2004). During the process of LSWT retrieval, there is also the need for cloud detection based on a Bayesian approach (MacCallum & Merchant, 2012). There are various types of uncertainty associated to this process (Rodgers, 1990). Some of them have been quantified, but the rest are still unknown. These uncertainties are not considered in this research, i.e. the analyses in this thesis do not account for the measurement errors of the retrieved data due to data availability.

The LSWT data were derived from the (A)ATSR observations. The ARC-Lake project processed the (A)ATSR data to obtain the LSWT for more than 900 lakes across the world, from June 1995 to April 2012. The spatial resolution of the retrieved LSWT data is $0.05° \times 0.05°$ (here '$\circ$' stands for degree in the geographical coordinates). Data sets typically consist of monthly aggregated measurements as spatial images, spatially aggregated lake mean products, etc (MacCallum & Merchant, 2013). They are available from the ARC-Lake v3.0 database (`http://www.geos.ed.ac.uk/arclake/data.html`). Reconstructed LSWT using geographical empirical orthogonal functions (EOFs) (Alvera-Azárate *et al.*, 2005) are also provided by the ARC-Lake project. The LSWT data was originally recorded in Kelvin and can be converted to Celsius by adding 273.15. The monthly aggregated LSWT data of Lake Victoria are used throughout the thesis. The lake, named by explorer John Hanning Speke after Queen Victoria, is the second largest fresh water lake on Earth. It is located between $31°39'\text{E} - 34°53'\text{E}$ and $03°00'\text{S} - 00°20'\text{N}$, covering an area of of 68,800 km^2. The ARC-Lake retrieved LSWT of Lake Victoria is defined on a grid of $65 \times 66 = 4290$ pixels, among which 2313 are identified as lake pixels. For monthly aggregated LSWT, this gives a data set of dimension 2313×203, or effectively an array of $65 \times 66 \times 203$, if the entire grid is considered. The Lake Victoria LSWT data show strong seasonality in individual pixels. In the meantime, there is large variation across the pixels, displaying interesting spatial/temporal patterns.

The Chlorophyll data, recorded in mg/m^3, were processed by the Diversity II project (`http://www.diversity2.info/products/`) using the MERIS measurements. The Diversity II demonstration sites include 340 large perennial inland waters distributed around the world. The spatial resolution of the monthly Chl data is 300m $\times$ 300m and the temporal coverage is from 2002 to 2012. Monthly, yearly and decadal aggregates are available from the database

(Brokeman Consult GmbH, 2015). The Globolakes project (http://www.globolakes.ac.uk/) covers a wider range of lakes globally, of more than 1000 lakes over 20 years. However, data are not fully accessible to the public currently. As the research in this thesis is associated with the Globolakes project, permission is given to use the Chl data of Lake Victoria as illustrations in this thesis. There are 732,585 pixels in the Lake Victoria Chl data set. The time coverage is from July 2002 to May 2012, giving 119 months in total. While sharing some common physical features as the LSWT data, the spatial/temporal dynamics of the Chl data behave in a slightly different way than the LSWT data. This difference helps to highlight some properties of the statistical methods investigated in this thesis.

1.1.2 Features of data and their influence on statistical analysis

One distinctive feature of the remote-sensing data is its dimensionality and large volume. The data are usually recorded as 3-dimensional arrays, defined by three coordinates, longitude, latitude and time. Observations may be densely recorded for either coordinate. The number of observations along each coordinate, when multiplied together, could result in thousands or millions of observations, presenting challenges to data analysis. This problem is referred to as 'high-dimensionality' in this thesis, although it is actually a combination of dimension and volume, not necessarily corresponding to data in a high dimensional space. Typically, there are two perspectives to investigate this type of data, (i) as a collection of time series, observed over a vast number of spatial locations, (ii) as a time series of spatial images, each consisting of a large number of pixels. Each has its own advantages according to the purposes of the analysis. However, neither perspective is straightforward to reveal the spatio-temporal features of the data due to the dimensionality. It would be attractive to develop a modelling framework to carry out the investigations of a large number of time series/images.

The second feature is the high percentage of missing observations per image/time series, which is referred as 'sparsity' in this thesis. It is a result of, e.g. cloud cover and the satellite orbit, and is common to the majority of remote-sensing data (Brokeman Consult GmbH, 2015, MacCallum & Merchant, 2013). For example, there are 7 months without a single observation in the Lake Victoria LSWT data set. For the rest of the months, the average missing percentage reaches almost 50%. Table 1.1 summarises the percentage of data available for the monthly images in the data set. 47 images show substantive missing, where less than 30% of the data are observed. To fully illustrate the sparsity in the LSWT data, plots using two perspectives (i) and (ii) described above, were produced. Figure 1.1

provides examples of the time series in 4 different pixels; Figure 1.2 presents images recorded at 8 different time points. The colours reflect the values of the LSWT, with the green end of the palette indicating low values and the blue end indicating high values[i]. Figure 1.1 suggests that there can be long periods of no observation in certain pixel locations; whereas Figure 1.2 suggests that the missing in space often appears as missing regions. Conventional statistical methods may not be applicable due to the missing data. There is often a need to modify the specification or the algorithm in order to accommodate the sparsity.

TABLE 1.1: A summary of the percentage of data available for 203 LSWT images of Lake Victoria.

% data available	$\leq 30\%$	$30\% - 50\%$	$50\% - 80\%$	$\geq 80\%$
image counts	47 (7 blank)	50	68	38

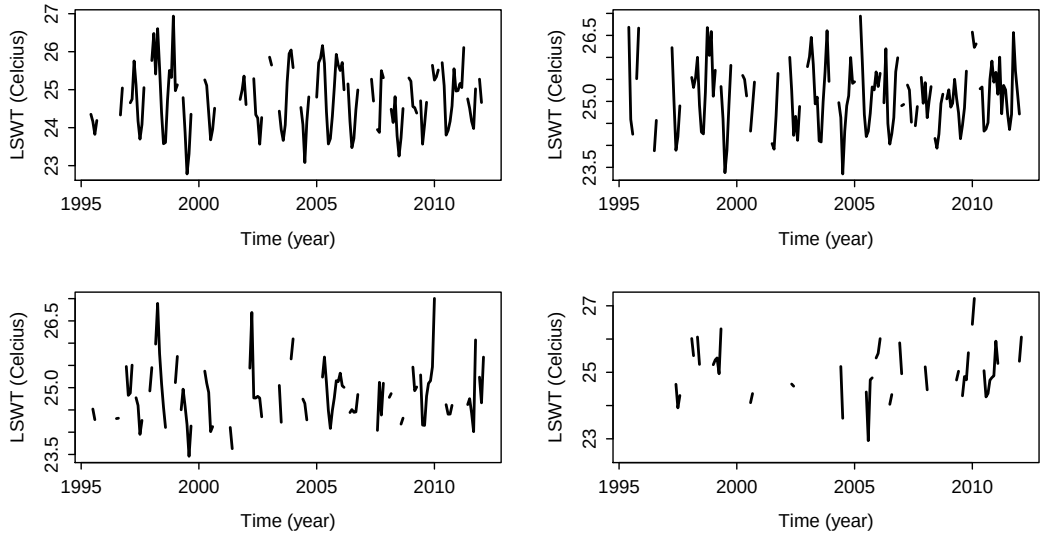

FIGURE 1.1: Examples of the sparse LSWT time series of Lake Victoria from 1995 to 2012 recorded at four different pixel locations.

Finally, the spatial/temporal dependence of the remote-sensing data is worth mentioning. This is not a feature unique to remote-sensing data, but is common to all spatio-temporal data. However, the dimensionality and sparsity of remote-sensing data make the spatial/temporal dependence especially interesting. On the one hand, these features complicate the modelling of the spatial/temporal correlation, as a result of the computational intensity, the adaptability of model specification and the estimation algorithm. On the other hand, the

[i]The same colour scheme is used throughout the thesis for displaying the image data (LSWT and Chl). The ranges of the values varies from figure to figure, but the green end of the palette is always for low values and the blue end for high values.

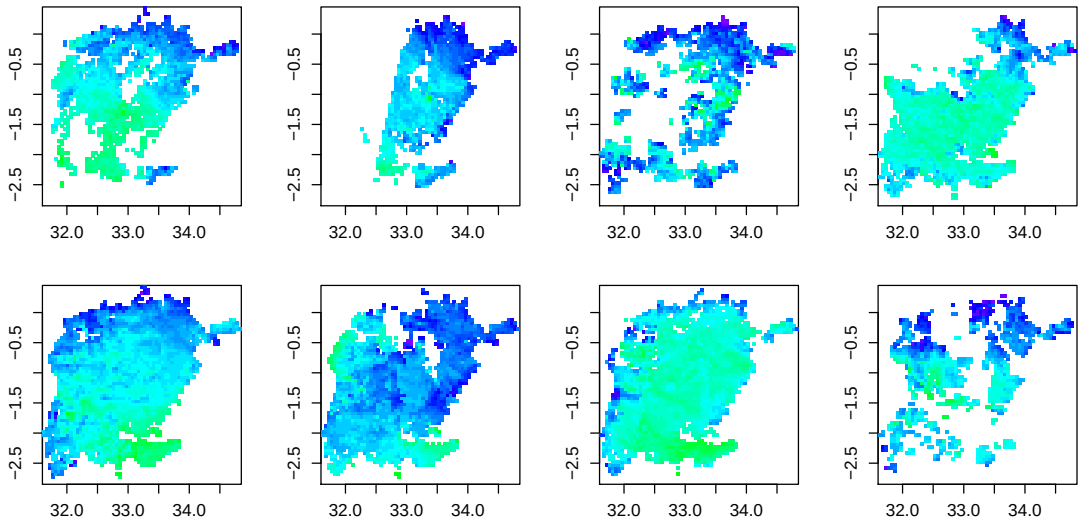

FIGURE 1.2: Examples of the sparse LSWT images of Lake Victoria from eight different time points. The green end of the palette indicates low values and the blue end indicates high values. The horizontal and the vertical axes represent longitude and latitude respectively.

process of dimension reduction and missing data imputation may benefit immensely from such a dependence structure.

1.2 Aims and objectives

The aim of this thesis is to provide novel statistical approaches to the analysis of the remote-sensing lake environmental data, so that the results may be used by ecologists to study the functions of lakes under climate change. It is of special interest to identify the general spatial/temporal patterns in the remote-sensing data for individual lakes. Specifically, there are three main objectives of this research.

(a) ***Dimension reduction***. The aim is to reduce the complexity in the data whilst identifying the main spatial/temporal features in the data. To achieve this, smoothing and functional data analysis techniques, using both univariate and bivariate functions, are investigated and developed.

(b) ***Missing data imputation***. Reliable imputations can improve the analyses of the data. To provide better data imputations, statistical methods based on mixed effect models are investigated. In particular, methods that combine the mixed effect model

and the functional data representations are developed to impute the missing values through a lower dimensional model with higher computational efficiency.

(c) ***Spatial-temporal modelling***. To model the spatial/temporal structures in the remote-sensing image time series, a classic spatio-temporal modelling framework using hierarchical design is investigated and a novel spatio-temporal model is proposed by extending existing models for sparse, high-dimensional data. The new model improves the data imputation and the extraction of the spatial/temporal patterns.

In the next section, statistical methods that are fundamental to the objectives of this research are introduced briefly.

1.3 Preliminary methodologies

1.3.1 Dimension reduction, smoothing and functional data analysis

The approaches to dimension reduction in this thesis are smoothing and functional data representation. Smoothing is a non-parametric technique for flexible modelling of non-linearity in curves, images, etc. Ruppert *et al.* (2003) described it as a method of '*freeing oneself of the restriction of parametric regression models*'. Without loss of generality, consider a model involving one univariate smooth function $f(x)$, expressed through a collection of K basis functions $\phi_k(x)$ and basis coefficients β_k,

$$Z_i = f(x_i) + \epsilon_i = \sum_{k=1}^{K} \phi_k(x_i)\beta_k + \epsilon_i \ , \tag{1.1}$$

where Z_i, $i = 1, \cdots, n$, are the observed data and x_i is the function argument associated with Z_i. Various data features can be modelled using appropriately chosen basis functions, e.g. Fourier basis for periodical patterns, natural cubic spline and B-spline bases for curvature. The basis coefficients are often estimated using ordinary least squares. A penalty is sometimes added to the estimation for more flexibility on the smoothness of function $f(x)$. In these situations, the estimation criterion can be written as (Wood, 2006)

$$\| \boldsymbol{Z} - \boldsymbol{\Phi\beta} \|^2 + \omega \boldsymbol{\beta}^\top \mathcal{S} \boldsymbol{\beta} \tag{1.2}$$

where $\boldsymbol{Z}$ is the vector of data Z_i, $i = 1, \cdots, n$; $\boldsymbol{\Phi}$ is the basis matrix, whose columns are basis functions $\phi_k(x)$, $k = 1, \cdots, K$, evaluated at $x = x_i$, $i = 1, \cdots, n$; $\boldsymbol{\beta}$ is the vector of

basis coefficients β_k, $k = 1, \cdots, K$; $\mathcal{S}$ is a penalty matrix, such as the second derivatives of $f(x)$, and ω is a smoothing parameter controlling the smoothness of the fit. Typically, $\omega \to 0$ indicates no penalty, resulting in a wiggly fit; whereas $\omega \to \infty$ would force $\boldsymbol{\beta}^\top \mathcal{S} \boldsymbol{\beta}$ to 0 so that criterion (1.2) can be minimized (as anything else would make it ∞), producing a smooth fit. The general estimation equation for $\boldsymbol{\beta}$ can be written as

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \omega \mathcal{S} \right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{Z} \,. \tag{1.3}$$

Methods for selecting smoothing parameter ω include (generalized) cross validation, information criteria, restricted maximum likelihood, etc (Reiss & Ogden, 2009, Wood, 2006).

Smoothing itself is not intended for dimension reduction. However, when it is paired with functional data analysis (FDA), the effect of dimension reduction becomes almost instant. FDA views the observations of individual objects in the data set as realizations of certain smooth functions, e.g. univariate functions for curves, bivariate functions for images. In other words, the 'observation' in FDA is a function and statistical analysis is carried out at the function level. Continuing the above example, consider now that the data collection process is carried out T times and at each time n observations are obtained, giving data Z_{ti}, $t = 1, \cdots, T$, $i = 1, \cdots, n$. Treat the T repeated measures as the 'individual objects' and assume that data are smooth by nature. Functions, $f_t(x)$, $t = 1, \cdots, T$, can be obtained by smoothing the data Z_{ti}, $i = 1, \cdots, n$, at time t respectively using model (1.1). Applying FDA on these functions means that a high-dimensional problem of $T \times n$ observations is transformed into a low-dimensional problem of T smooth function. This is very appealing for remote-sensing data, which often have much higher dimension in space than in time ($n \gg T$). Some frequently used FDA methods include, functional regression, functional clustering, functional PCA, etc (Ramsay & Silverman, 1997). The technique used in this thesis is the functional principal component analysis (FPCA). Details on the estimation and interpretation of the FPCA are provided in Chapter 2.

1.3.2 Missing data imputation, mixed effect model and EM algorithm

There are typically regarded to be three categories of missing data, missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) (Little & Rubin, 2002). The first category assumes that the probability of an observation being missing is independent of the observed and missing values of data. The second category

describes a situation where the probability an observation is missing is independent of the values that are missing, but may depend on the values of the data that are observed. In the third category, there is often a missing data mechanism associated with the missingness. In practice, data which are categorized as MCAR or MAR are often modelled with the missing data mechanism ignored. Discussion on the ignorability of the missing data mechanism and its modelling strategies can be found in Lu & Copas (2004), Seaman *et al.* (2013), etc.

Due to the complexity of the satellite measurements and retrieval algorithm, there is no universal agreement on whether the missingness should be treated as random or systematic. The missingness in the remote-sensing data considered in this thesis (LSWT and Chl) is associated with cloud cover and satellite orbit tracks (Brokeman Consult GmbH, 2015, MacCallum & Merchant, 2013), two factors that are independent of the unobserved values of the variable. There are situations where the missingness is a result of the data retrieval algorithm. As some algorithms perform better in certain spectral range than others, the value of an observation (a realization of the observed spectrum) may actually affect the probability it is missing. However, as the retrieval algorithms are often complicated and the data product may even be a blend of several algorithms, it is impractical to form a missing data mechanism based on these and incorporate it into the modelling. Therefore, the missing data mechanism is not considered and the missing data are treated as MAR in this thesis[ii].

Under the scenario of MAR, the missing data mechanism may be ignored in the modelling process (for likelihood inference and Bayesian inference alike), if the parameters governing the missing data mechanism are distinct from the parameters in the model (Heitjan & Rubin, 1991, Lu & Copas, 2004). In this thesis, the distinctness of parameters is assumed. Statistical methods based on the mixed effect modelling framework using likelihood inference are adopted to impute data that are MAR. This approach offers the possibility of utilizing the entire data set to improve data imputation. A general linear mixed effect model can be written (using matrix notation) as

$$\boldsymbol{Z} = \boldsymbol{X}_f \boldsymbol{b} + \boldsymbol{X}_r \boldsymbol{\eta} + \boldsymbol{\epsilon} \,, \tag{1.4}$$

where $\boldsymbol{Z}$ is a vector of observations, $\boldsymbol{X}_f$ is matrix of the fixed effect covariates and $\boldsymbol{X}_r$ is the design matrix of the random effect. $\boldsymbol{X}_r$ is usually specified based on the type of random effect, such as individual effect and group effect. Distributional assumptions are often assigned

[ii] For data as combinations of different algorithms, the blending process would reduce the chance of an observation being missing due to algorithm failure. This reduces the influence of the missing data mechanism.

to both the random effect coefficient $\boldsymbol{\eta}$ and the model residual $\boldsymbol{\epsilon}$ as, $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R})$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{V})$. This gives the covariance matrix of the model

$$\boldsymbol{\Sigma} = \mathbf{Cov}[\boldsymbol{X}_r\boldsymbol{\eta} + \boldsymbol{\epsilon}] = \boldsymbol{X}_r\boldsymbol{R}\boldsymbol{X}_r^\top + \boldsymbol{V}\,. \tag{1.5}$$

According to Ruppert *et al.* (2003), given $\boldsymbol{\Sigma}$, the fixed effect coefficient can be estimated using generalized least squares; given $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, the random effect $\boldsymbol{\eta}$ can be obtained as the best linear predictor based on conditional distribution of $\boldsymbol{\eta}|\boldsymbol{Z}$. That is

$$\hat{\boldsymbol{b}} = \left(\boldsymbol{X}_f^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_f\right)^{-1} \boldsymbol{X}_f^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{Z}\,, \tag{1.6}$$

$$\hat{\boldsymbol{\eta}} = \boldsymbol{R}\boldsymbol{X}_r^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{Z} - \boldsymbol{X}_f \boldsymbol{b})\,. \tag{1.7}$$

Model parameters $\boldsymbol{R}$ and $\boldsymbol{V}$ can be estimated using maximum likelihood (ML) or restricted maximum likelihood (REML), which is an averaged version of ML over all possible values of $\boldsymbol{b}$. The corresponding log-likelihood based on the observed data are

$$\mathcal{L}(\Psi; \boldsymbol{Z}) = -\frac{1}{2}\left\{\ln(|\boldsymbol{\Sigma}|) + (\boldsymbol{Z} - \boldsymbol{X}_f\boldsymbol{b})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{Z} - \boldsymbol{X}_f\boldsymbol{b})\right\} + \text{constant}$$

for ML and

$$\mathcal{L}_{re}(\Psi; \boldsymbol{Z}) = -\frac{1}{2}\left\{\ln(|\boldsymbol{\Sigma}|) + (\boldsymbol{Z} - \boldsymbol{X}_f\boldsymbol{b})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{Z} - \boldsymbol{X}_f\boldsymbol{b}) + \boldsymbol{X}_f^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_f\right\} + \text{constant}\,,$$

for REML, where $\Psi = \{\boldsymbol{R}, \boldsymbol{V}\}$ is the parameter collection. On substituting equations (1.5) and (1.6) into the log-likelihood functions, the maximum likelihood estimates (MLEs) of $\boldsymbol{R}$ and $\boldsymbol{V}$ can be obtained (Ruppert *et al.*, 2003).

In some situations, it is easier to maximize the joint log-likelihood of the observed data and the random effect component $\mathcal{L}(\Psi; \boldsymbol{Z}, \boldsymbol{\eta})$ based on $f(\boldsymbol{Z}, \boldsymbol{\eta}) = f(\boldsymbol{Z}|\boldsymbol{\eta})f(\boldsymbol{\eta})$ than the log-likelihood of the observed data alone $\mathcal{L}(\Psi; \boldsymbol{Z})$. This is due to the complexity in evaluating the derivatives of the observed log-likelihood $\mathcal{L}(\Psi; \boldsymbol{Z})$ to obtain the MLEs. One way to implement the estimation using $\mathcal{L}(\Psi; \boldsymbol{Z}, \boldsymbol{\eta})$ is the expectation-maximization (EM) algorithm. It is a general method for obtaining MLEs in incomplete data problems (Little & Rubin, 2002). The algorithm, first formalized in statistical literature by Dempster *et al.* (1977), consists of two iterative steps, (a) an expectation step (E-step), where the missing information is estimated based on a conditional distribution evaluated at the current parameter estimates and the expectation of the complete data log-likelihood is computed accordingly, (b) a maximization

step (M-step), where the parameters are updated through maximizing the expectation of the complete data log-likelihood.

In terms of a mixed effect model, the complete data are often defined to be the joint of the observations and the random effect $\{\boldsymbol{Z}, \boldsymbol{\eta}\}$, with the random effect $\boldsymbol{\eta}$ treated as the missing information. This gives an incomplete data problem which can be estimated using the EM algorithm. Specifically, in the it-th iteration, the E-step calculates the expectation of the joint log-likelihood

$$\mathcal{Q}\left(\Psi; \Psi^{(it-1)}\right) = \mathbf{E}\left[\mathcal{L}(\Psi; \boldsymbol{Z}, \boldsymbol{\eta}) \,\middle|\, \boldsymbol{Z}, \Psi^{(it-1)}\right] .$$

The M-step then updates the parameter estimation to $\Psi^{(it)}$, such that the condition

$$\mathcal{Q}\left(\Psi^{(it)}; \Psi^{(it-1)}\right) \geq \mathcal{Q}\left(\Psi; \Psi^{(it-1)}\right), \quad \forall\, \Psi \in \mathcal{W}$$

is satisfied, where $\mathcal{W}$ is the parameter space. The iteration terminates when certain convergence criterion is met. Various extensions have been developed based on this general framework. This is discussed in detail in later chapters.

1.3.3 Spatial/temporal dependence and dynamic models

The spatial/temporal patterns in environmental data are usually of great interest, as they help to answer questions about long term change, spatial clustering of environmental variables, coherent evolution under climate change, etc. In order to model these patterns, the spatial/temporal dependence needs to be assessed appropriately.

One way of describing the spatial/temporal dependence is through some descriptive functions of correlation/covariance. The autocorrelation function (ACF) is one of the most important measures of the temporal dependence. Typically, for a temporal process $\{Z_t\}$, $t \in \mathcal{T}$, a lag-τ ACF measures the linear dependence of the series at time t on the observation at time $t-\tau$. For a second-order stationary process, the ACF is determined through the time lag τ only, denoted as $\rho(\tau)$ (Shumway & Stoffer, 2006). A frequently used measure to quantify spatial dependence is a (semi-)variogram. For a spatial process $\{Z_{\boldsymbol{s}}\}$, $\boldsymbol{s} \in \mathcal{D}$, the semi-variogram is defined as (Cressie, 1993)

$$\begin{aligned} \gamma(\boldsymbol{s}, \boldsymbol{r}) &= \frac{1}{2}\mathbf{Var}[Z_{\boldsymbol{s}} - Z_{\boldsymbol{r}}] \\ &= \frac{1}{2}\left(\mathbf{Var}[Z_{\boldsymbol{s}}] + \mathbf{Var}[Z_{\boldsymbol{r}}] - 2\mathbf{Cov}[Z_{\boldsymbol{s}}, Z_{\boldsymbol{r}}]\right) , \quad \boldsymbol{s}, \boldsymbol{r} \in \mathcal{D} . \end{aligned} \tag{1.8}$$

For a second order stationary spatial process, $\gamma(\boldsymbol{s}, \boldsymbol{r})$ is determined only through the spatial difference $\boldsymbol{h} = \boldsymbol{s} - \boldsymbol{r}$, giving

$$\gamma(\boldsymbol{h}) = \mathbf{Var}[Z_{\boldsymbol{s}}] - \mathbf{Cov}[Z_{\boldsymbol{s}}, Z_{\boldsymbol{s}+\boldsymbol{h}}]\,.$$

Furthermore, for an isotropic spatial process where the spatial correlation being the same whichever direction it takes, the spatial lag can be replaced by the Euclidean distance $\|\boldsymbol{h}\| = \|\boldsymbol{s} - \boldsymbol{r}\|$. The semi-variogram can be linked to a correlation function $\rho(\boldsymbol{h})$, which describe the type of the spatial dependence, e.g. Gaussian, exponential and Matérn.

For a spatio-temporal process, $\{Z_{(\boldsymbol{s},t)}\}$ defined on $\boldsymbol{s} \in \mathcal{D}$, $t \in \mathcal{T}$, the covariance function is often written as $\mathbf{Cov}[Z_{(\boldsymbol{s};t)}, Z_{(\boldsymbol{r};u)}] = C((\boldsymbol{s};t),(\boldsymbol{r};u))$ for some positive-definite function $C((\boldsymbol{s};t),(\boldsymbol{r};u))$ on $\mathbb{R}^d \times \mathbb{R}$, for d-dimensional spatial domain $\mathcal{D}$ and 1-dimensional temporal domain $\mathcal{T}$ (Cressie & Wikle, 2011). Depending on different assumptions, various types of covariance models can be constructed, such as

$$\mathbf{Cov}[Z_{(\boldsymbol{s};t)}, Z_{(\boldsymbol{r};u)}] = C^{(st)}(\boldsymbol{s} - \boldsymbol{r}; t - u)$$

for a second-order stationary spatio-temporal process and

$$\mathbf{Cov}[Z_{(\boldsymbol{s};t)}, Z_{(\boldsymbol{r};u)}] = C^{(s)}(\boldsymbol{s}, \boldsymbol{r})C^{(t)}(t, u)$$

for a space-time separable covariance structure, where $C^{(s)}(\boldsymbol{.},\boldsymbol{.})$ and $C^{(t)}(\boldsymbol{.},\boldsymbol{.})$ are valid spatial and temporal covariance functions respectively. A separable covariance function is perhaps the easiest to implement, hence received extensive study over the years. However, such a setting is not always realistic in practice. Readers are referred to Cressie & Wikle (2011) for a review. For non-separable covariance structures, methodologies such as the spatio-temporal variogram, spectral representation (Cressie & Huang, 1999) and Taylor's hypothesis in fluid dynamic (Gneiting, 2006), have been developed to model the covariance functions. However, as pointed out in Cressie & Wikle (2011), these models usually play '*a descriptive role in representing the spatio-temporal dependence in the process... That is, it is very difficult to look at a covariance function and determine the etiology of the spatio-temporal process under study*'. An alternative method to construct a valid covariance function directly is to model the spatial/temporal covariance structure based on a specific type of stochastic partial differential equations (SPDE), which can be linked to the Gaussian Markov random fields (Lindgren *et al.*, 2011). The authors established the connections between the SPDEs

and the precision matrices of a wide variety of spatial/temporal processes, including non-stationary, non-separable, anisotropic processes, etc. This is a flexible approach, though its interpretation is again non-trivial.

A slightly different way of describing the spatio-temporal dependence is through a dynamic model, where the dependence is motivated by the evolution of, e.g. physical, chemical and economic processes. Such models are usually built on the conditional distributions describing how the current state behaves given the 'nearby' current and past values (Cressie & Wikle, 2011). For example, a general model for the dependence of a spatial process at time t on that of time $t-\tau$, for a positive real value τ, can be written as

$$Z_{(\boldsymbol{s};t)} = \mathcal{M}_t\left(\boldsymbol{s}, \boldsymbol{Z}_{(\cdot;t-\tau)}\right) + \epsilon_{(\boldsymbol{s};t)}\,, \tag{1.9}$$

where function $\mathcal{M}_t(\cdot,\cdot)$ depends on both the spatial location $\boldsymbol{s}$ and the observations at time $t-\tau$, $Z_{(\cdot;t-\tau)}$. The function $\mathcal{M}_t(\cdot,\cdot)$ is possibly non-linear and can be either time dependent or invariant, providing enough flexibility to generate the (non)stationary spatio-temporal processes. Some examples of the discrete time $\mathcal{M}_t(\cdot,\cdot)$ function are first-order vector autoregressive model, or VAR(1) model

$$\boldsymbol{Z}_{(\cdot;t)} = \boldsymbol{M}\boldsymbol{Z}_{(\cdot;t-1)} + \boldsymbol{\epsilon}_{(\cdot;t)}\,,$$

and stochastic integro-difference equation (IDE)

$$Z_{(\boldsymbol{s};t)} = \int_{\mathcal{D}} m(\boldsymbol{s},\boldsymbol{x}) Z_{(\boldsymbol{x};t-1)} d\boldsymbol{x} + \epsilon_{(\boldsymbol{s};t)}\,.$$

Note that both examples use $\tau = 1$, which is the unit of the discretization of time. Cressie & Wikle (2011) encourage the use of scientific knowledge to motivate the design of the $\mathcal{M}_t(\cdot,\cdot)$, in the sense of a 'physical statistical' model. A review on this topic can be found in Wikle & Hooten (2010).

In this thesis, a particular type of spatio-temporal process, referred to as the 'time series of spatial process' in Cressie & Wikle (2011), receives in-depth investigation. The process can be written as

$$\boldsymbol{Z}_t(\cdot) \equiv \left\{Z_{(\boldsymbol{s},t)} \,:\, \boldsymbol{s} \in \mathcal{D}_t\right\}\,, \quad t = 1, 2, \cdots$$

with the index t in $\mathcal{D}_t$ emphasising that the spatial index set is allowed to change with time. The dynamic model (1.9) for this type of process thus becomes

$$Z_t(\boldsymbol{s}) = \mathcal{M}_t\left(\boldsymbol{s}, \boldsymbol{Z}_{t-1}(\centerdot)\right) + \epsilon_t(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}_t \tag{1.10}$$

It is straightforward to see that the remote-sensing image time series can be viewed as a time series of spatial process. Therefore, model (1.10) can be used to describe the spatio-temporal dependence of the data studied in this thesis. Perhaps an even more attractive feature is the model's potential to achieve dimension reduction through appropriate design of the system dynamics (Wikle & Cressie, 1999). Investigation with respect to this route is carried out in Chapter 4.

1.4 Thesis structure

This remainder of the thesis is made up of five chapters. Chapter 2 presents the exploratory analysis of the spatial and temporal features of the remote-sensing data, using the Lake Victoria LSWT data as an example. Studies from both the temporal curves and spatial images perspectives were carried out. Classic techniques, such as harmonic regression and autoregressive models, spatial smoothing and covariogram models are used to investigate the spatial and temporal properties of the data. The chapter also presents an initial investigation of the data using FPCA. Based on the exploratory analysis, a baseline model for analysing the sparse image time series is introduced in Chapter 3. The model inherits the specifications of FPCA, but is parameterized as a mixed effect model. Model estimation exploits the EM algorithm, so that the missing data problem can be overcome. However, this model assumes that there is no temporal dependence between images, which could be problematic in some situations. Therefore, methodologies for incorporating the temporal correlations between the images are explored in Chapter 4. In particular, a dynamic spatio-temporal modelling framework is investigated, with special attention paid to model specification and computation details. Based on these studies, a spatio-temporal model that updates the mixed model FPCA is proposed to analyse the sparse image time series, along with an estimation method making use of an extension of the EM algorithm. Chapter 5 is dedicated to the investigation of the proposed model, using both simulated and real remote-sensing data. A study on the asymptotic behaviors of the estimation algorithm is presented, followed by an application of the proposed model to the Lake Victoria LSWT and Chl data. Final remarks on these methodologies and potential future works are provided in Chapter 6.

Before leaving this chapter, a list of subsets of the Lake Victoria LSWT and Chlorophyll data used in the thesis for illustration purposes is presented here.

(a) The 'Re LSWT' data set. This is a subset of the ARC-Lake reconstructed LSWT of Lake Victoria. It is defined on a grid of size 26×27 and consists of 203 monthly images with no missing observations. The data set is used where complete data are required in order to implement the method.

(b) The 'LSWT section' data set. This is extracted from the sparse LSWT data of Lake Victoria. It is defined on a grid of size 34×24 and consists of 202 monthly images with missing observations. This data set is used in the thesis for general illustrations.

(c) The 'Artificial section ' data set. This is constructed using the reconstructed LSWT data of Lake Victoria. It is defined on the same grid as the 'LSWT section' data set, with sparsity imposed using the missing patterns of the 'LSWT section' data set. This data set is used in model investigation because it provides 'true values' for the missing observations, which is helpful in assess the quality of data imputation.

(d) The 'Chl section' data set. This is a subset of the 3×3 spatially aggregated Lake Victoria Chlorophyll data, defined on a 36×36 grid, including 119 monthly images. The spatial aggregation is carried out by taking the average of the values from 9 pixels in a 3×3 grid and then using this averaged value as the observation of the larger pixel which covers the 3×3 grid[iii]. This data set is used in model investigations because the Chl data display different spatio-temporal feature as compared to the LSWT data and can thus highlight model properties that cannot be discovered using the LSWT data.

(e) The applications of the main statistical methods in this thesis are carried out on larger date sets, for both the LSWT (size: $47 \times 57 \times 202$) and Chlorophyll (size: $72 \times 72 \times 119$) data of Lake Victoria. Details of the two application data sets are provided in the corresponding sections of Chapter 3 and 5.

[iii] This can be done using the `R` package `raster`.

Chapter 2

Exploratory analysis

Nothing puzzles me more than space and time.

Charles Lamb (1810)

This chapter presents the exploratory analysis of the remote-sensing image time series. The data used as illustrations are the LSWT data of Lake Victoria. Standard time series and spatial analysis are carried out to investigate the data from two perspectives (a) temporal curves recorded for 2313 pixels in a 65×66 grid, (b) spatial images recorded monthly from May 1995 to April 2012. Functional principal component analysis is applied to explore the general spatial and temporal patterns in the data set. Drawbacks of these methods on applying to remote-sensing data are discussed and potential solutions to these problems are reviewed at the end of the chapter.

2.1 Investigating temporal patterns

Exploratory analysis was first carried out from the temporal perspective, that is, modelling the time series of LSWT in individual pixels. The aim was to investigate the long-term temporal patterns in the time series other than the obvious seasonal patterns. The main approach used here was harmonic regression with residual autocorrelation structure incorporated as an auto-regressive (AR) model.

2.1.1 Harmonic regression

Harmonic regression is frequently used to model periodic data (Pigorsch & Bailer, 2005). The model considered in this analysis consists of a harmonic component and a general temporal trend component, to capture the strong seasonal patterns and the potential long-term trend in the data. A general harmonic regressors can be written as a sinusoid signal

$$A\cos(2\pi\nu t + \varphi)\,,$$

where t is the time covariate, A represents the amplitude, ν is a parameter associated with the frequency and φ is the phase parameter. Since the majority of the LSWT time series have one peak and one trough within a 12-month cycle, $\nu = \frac{1}{12}$ was used in this problem. Expanding the above sinusoid and re-parameterizing the coefficients gives the specific harmonic regressors,

$$A_1\cos\left(\frac{2\pi t}{12}\right) + A_2\sin\left(\frac{2\pi t}{12}\right)\,.$$

The general temporal trend component can be formulated using polynomials of covariate t. To allow more flexibility, a smoothed function of t was considered here, giving

$$Z_t = A_1\cos\left(\frac{2\pi t}{12}\right) + A_2\sin\left(\frac{2\pi t}{12}\right) + f(t) + \epsilon_t\,. \tag{2.1}$$

In the above model, Z_t is the LSWT at time t. Smooth function $f(t)$ is constructed using a cubic spline basis as $f(t) = \Phi(t)\boldsymbol{\beta}$, with basis $\Phi(t) = (\phi_1(t), \cdots, \phi_K(t))$ and basis coefficient vector $\boldsymbol{\beta}$. Note that the intercept term of the model is incorporated in the basis, corresponding to $\phi_1(t) = 1$. The smoothness of function $f(t)$ is penalized by the integrated squared second derivative (Wood, 2006)

$$\mathcal{P}(f) = \int_{\mathcal{T}} \left[f''(t)\right]^2 dt\,.$$

The above penalty can be written in the form of $\boldsymbol{\beta}^\top\mathcal{S}\boldsymbol{\beta}$ as described in section 1.3.2. In this case, the matrix $\mathcal{S}$ is determined by the second derivative of the basis matrix $\boldsymbol{\Phi}$. The minimization criterion of this problem thus becomes

$$\sum_t \left[Z_t - A_1\cos\left(\frac{2\pi t}{12}\right) - A_2\sin\left(\frac{2\pi t}{12}\right) - \Phi(t)\boldsymbol{\beta}\right]^2 + \omega\boldsymbol{\beta}^\top\mathcal{S}\boldsymbol{\beta}\,.$$

Additionally, the model residuals are assumed to be independently and identically distributed (i.i.d) with $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$. The time point $t = 0$ is taken to be the January 1995, which is

the first month of the year when the observing began.

The effective degrees of freedom (EDF) was used to measure the smoothness of function $f(t)$ in model (2.1). In a simple regression model, the degrees of freedom are determined by the dimension of the design matrix. Whereas in a regression model using smooth functions, such as $\boldsymbol{Z} = \boldsymbol{\Phi}\boldsymbol{\beta}+\boldsymbol{\epsilon}$, the dimension of the basis matrix $\boldsymbol{\Phi}$ usually does not reflect the actual degrees of freedom of the model. According to Wood (2006), *'the basis dimension is only setting an upper bound on the flexibility of a term: it is the smoothing parameter that controls the actual effective degrees of freedom.'*. For a smooth term $\boldsymbol{\Phi}\boldsymbol{\beta}$, with associated penalty $\mathcal{P}(f) = \boldsymbol{\beta}^\top \mathcal{S}\boldsymbol{\beta}$ and smoothing parameter ω, the EDF can be computed as

$$\text{EDF} = \text{tr}\left\{ \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \omega\mathcal{S}\right)^{-1}\boldsymbol{\Phi}^\top \right\}, \tag{2.2}$$

where tr$\{\boldsymbol{.}\}$ denotes the trace of the matrix. Due to the existence of the penalty, the EDF is always smaller or equal to the dimension of the basis, with equality holding when $\omega = 0$. In addition, the above trace does not need to be an integer, neither does the EDF. In fact, it can take any real value between 1 and the number of parameters. For example, in terms of model (2.1), EDF $= 1$ would suggest a constant (or intercept) term and EDF $= 2$ corresponds to a linear function of t. The EDF is sometimes used in model selection for the optimal smoothness of the fitted function.

As mentioned in section 1.3.1, the smoothing parameter ω also needs to be selected. Some frequently used methods include cross validation,

$$\text{CV}(\omega) = \frac{1}{T}\sum_{t=1}^{T}\left[Z_t - \hat{f}^{[-t]}(t)\right]^2,$$

and generalized cross validation,

$$\text{GCV}(\omega) = T\,\frac{\sum_{t=1}^{T}[Z_t - \hat{f}(t)]^2}{[\text{tr}\{\boldsymbol{I} - \widetilde{\boldsymbol{P}}\}]^2} = T\,\frac{\parallel (\boldsymbol{I} - \widetilde{\boldsymbol{P}})\boldsymbol{Z} \parallel^2}{[\text{tr}\{\boldsymbol{I} - \widetilde{\boldsymbol{P}}\}]^2},$$

where $\hat{f}^{[-t]}(t)$ is the smooth function, with smoothing parameter ω, fitted to all the data except Z_t, $\hat{f}(t)$ is the smooth function, with smoothing parameter ω, fitted to all the data, $\widetilde{\boldsymbol{P}} = \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \omega\mathcal{S}\right)^{-1}\boldsymbol{\Phi}^\top$ is the influence (or projection) matrix and $\boldsymbol{Z}$ is the vector of all Z_t, $t = 1, \cdots, T$. Alternatively, various information criteria can be used in the selection. These criteria are often formulated as twice the negative log-likelihood of the model (a measure of distance between the candidate model and the 'true' model) plus certain forms of penalty on

the degrees of freedom. Two of the most frequently used information criteria are

$$\text{AIC} = -2\mathcal{L}(\widehat{\Psi}, \omega; \boldsymbol{Z}) + 2q\,, \tag{2.3}$$

$$\text{BIC} = -2\mathcal{L}(\widehat{\Psi}, \omega; \boldsymbol{Z}) + \log(n)q\,, \tag{2.4}$$

where $\mathcal{L}(\widehat{\Psi}; \boldsymbol{Z})$ is the log-likelihood evaluated at $\widehat{\Psi}$ with smoothing parameter ω, q is the dimension of the parameter collection Ψ and n is the number of observations[i]. The ω value minimizing equation (2.3) and (2.4) is considered as the solution. Although sometimes, the AIC/BIC values may only be used as a guide, as the results can be misleading when the number of observations is not large enough compared to the dimension of the model (Hurvich *et al.*, 1998), or when the data are highly correlated.

In this analysis, all the LSWT time series of Lake Victoria of pixels with over 50% observations available were investigated using model (2.1). Function `gam` in `R` package `mgcv` (Wood, 2011) was used to fit the models. The model was estimated using REML, where the smoothing parameter ω was selected by re-parameterizing the higher order smooth component as random effect and incorporating the smoothing parameter into the model covariance structure. In the package `mgcv`, the influence of the intercept is not counted in the output of the EDF. It gives the value of (2.2) minus 1. That is, EDF = 1 from the `gam` output corresponds to a linear function of t. It is found that 94.8% of the fitted smooth functions have EDF between 1 and 2, the majority of which have EDF just slightly larger than 1. This means that most of the estimated $\hat{f}(t)$ are nothing more than a linear function of t. Additionally, the p-values based on the pseudo-inverse of the covariance matrix of the estimated basis coefficient $\hat{\boldsymbol{\beta}}$ (i.e. the approximated significance of the fitted smooth function) are large in most of the cases, suggesting that $\hat{f}(t)$ have very limited influence on these models.

Figure 2.1 is a map showing whether the time series in a pixel is considered to have a temporal trend or not. The dark grey dots represent pixels with EDF < 2; the red dots represent pixels with EDF ≥ 2. The majority of the dark grey pixels have approximated p-values greater than 0.05, suggesting that no distinctive linear temporal trend is found in the majority of the pixels. The red pixels may be considered as to exhibit certain non-linear temporal trend, although some of them still have relatively large approximated p-values. Two examples of the fitted harmonic regression models (2.1) are given in Table 2.1. The models were applied to two dark grey pixels in the map, each with more than 80% of the data observed. Both

[i]Note that for a model involves smoothing, the dimension of the parameters associated with the smooth term is determined by the effective degrees of freedom, not the number of elements in the parameter collection.

models have EDF slightly over 1 and approximated p-values greater than 0.05. Based on these results, it can be concluded that there is no clear long-term trend in the LSWT time series for the majority of the pixels of Lake Victoria. Therefore, in the rest of the analysis, the harmonic regression model (2.1) is replaced by a model without trend component,

$$Z_t = A_0 + A_1 \cos\left(\frac{2\pi t}{12}\right) + A_2 \sin\left(\frac{2\pi t}{12}\right) + \epsilon_t \,, \tag{2.5}$$

for simplicity and ease of comparison.

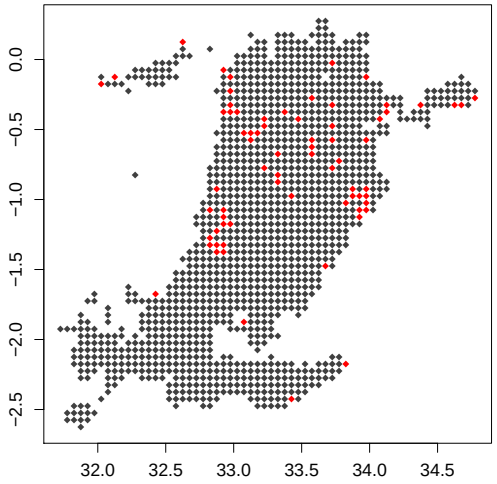

FIGURE 2.1: Map of pixels (with $\geq 50\%$ data available) investigated using model (2.1). The dark grey dots represent pixels without temporal trend and the red dots represent pixels with a temporal trend. The horizontal and vertical axes are longitude and latitude respectively.

TABLE 2.1: Results from the harmonic regression model (2.1) fitted to the LSWT time series in two pixels, located at (33.275E°, −2.375N°) and (33.925E°, −0.125N°) respectively.

Location	Intercept	$\hat{A}_1$	$\hat{A}_2$	$\hat{\sigma}^2$	EDF of $\hat{f}(t)$	p-value of $\hat{f}(t)$
33.275E°, −2.375N°	24.81	0.29	1.02	0.42	1.001	0.362
33.925E°, −0.125N°	25.58	0.34	0.49	0.32	1.317	0.793

2.1.2 Temporal autocorrelation

In the previous analysis, the model residuals were assumed to be independent and identically distributed (i.i.d.). This is an assumption made to simplify the initial investigation and could be inappropriate to model time series data. Therefore, the autocorrelations in the residuals from fitting model (2.5) are examined here using empirical (or sample) autocorrelation functions (ACF) and variograms.

For fully observed time series, under the second-order stationary assumption, the empirical autocorrelation function can be computed as

$$\hat{\rho}(\tau) = \frac{\sum_{t=1}^{T-\tau} \left(Z_t - \bar{Z}\right)\left(Z_{t+\tau} - \bar{Z}\right)}{\sum_{t=1}^{T} \left(Z_t - \bar{Z}\right)^2}, \tag{2.6}$$

where τ is the time difference (or time lag) and $\bar{Z} = \frac{1}{T}\sum_{t=1}^{T} Z_t$. However, for sparse or irregularly observed time series, the computation of the empirical ACF can be difficult. Alternatively, a method based on the idea of the variagrom in spatial statistics may be used (Haslett, 1997). Variograms are often constructed as a function of the distance in space (Cressie, 1993, Pigorsch & Bailer, 2005). Recall equation (1.8) from section 1.3.3 about the variogram of a spatial process. The same formula can be adopted to examine the correlation between two residuals with certain time difference apart. Rescaling the variogram formula (1.8) by the variance of the residuals, a measure of the autocorrelation between the pair of residuals, ϵ_t and $\epsilon_{t+\tau}$, can be constructed as

$$\rho(\tau) = \frac{\gamma(\tau)}{\mathbf{Var}[\epsilon_t]} = \frac{\mathbf{Var}[\epsilon_t] + \mathbf{Var}[\epsilon_{t+\tau}] - 2\mathbf{Cov}[\epsilon_t, \epsilon_{t+\tau}]}{\mathbf{Var}[\epsilon_t]} \tag{2.7}$$

The component $\gamma(\tau)$ in equation (2.7) is referred to as the 'temporal variogram' in the rest of the thesis. A general procedure to compute the empirical 'temporal variogram', $\hat{\gamma}(\tau)$, consists of the following steps.

(a) Calculate the time difference τ_{ij} between each pairs of residuals i and j; group the time differences into M intervals, denoted as L_m, $m = 1, \cdots, M$.

(b) Calculate the empirical variance within each interval L_m.

$$\hat{\gamma}_m = \frac{1}{n_m} \sum_{\tau_{ij} \in L_m} (\epsilon_i - \epsilon_j)^2,$$

where n_m is the number of residual pairs in interval L_m

(c) Plot $\hat{\gamma}_m$ against the median of each interval L_m.

The resulting plot can be used to investigate how the temporal correlation changes with the increasing time lag. For example, if a process displays a first order auto-regressive (or AR(1)) structure, $\epsilon_t = \psi\epsilon_{t-1} + v_t$, then the empirical temporal variogram should be able to match the theoretical ACF of an AR(1) process, $\rho(\tau) = 1 - \psi^\tau$, rescaled by a factor of $\sigma^2 = \mathbf{Var}[\epsilon_t]$ or its sample version (Haslett, 1997).

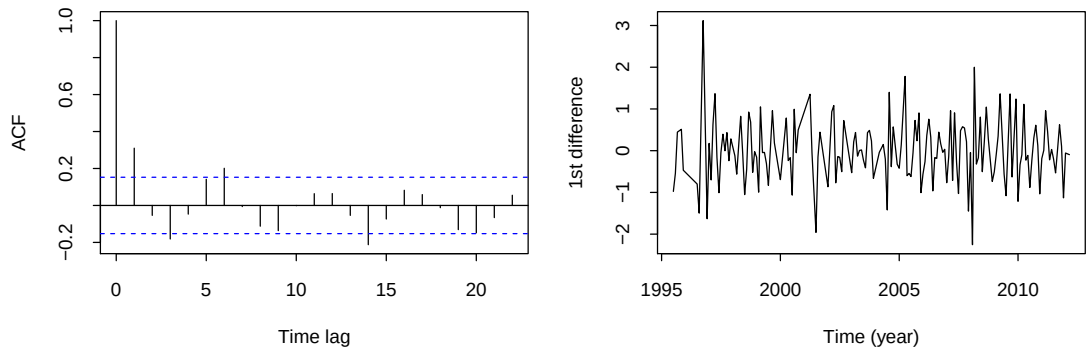

FIGURE 2.2: (Left) The ACF plot of the residuals from model (2.5) fitted to the LSWT time series of the pixels located at (33.275E°, −2.375N°). (Right) The plot of the first-order difference between the adjacent residuals.

In this analysis, the empirical ACF as defined in equation (2.6) was used to investigate the autocorrelations of time series from pixels with more than 75% data observed. The empirical version of the temporal variogram (2.7) was applied to the time series with less data available. Unfortunately, the sample variograms did not provide much useful information of correlation structures of the LSWT time series. Most of the plots were too wiggly to draw any conclusions from. This was the result of the relatively limited and unbalanced number of observations within each interval L_m, $m = 1, \cdots, M$. Therefore, only the sample ACFs computed directly using the `R` function `acf` from pixel locations with $\geq 75\%$ observations were investigated. Figure 2.2 presents the ACF plot of the residuals from fitting model (2.5) to the LSWT time series of a pixel located at (33.275E°, −2.375N°) in the left panel and the corresponding plot of the first-order difference, $\epsilon_t - \epsilon_{t-1}$, in the right panel. The ACF plot shows that, the autocorrelations of the majority of the time lags τ fall within the 95% confidence interval (indicated by the two dashed lines), apart from the lag-1 autocorrelation. This suggests that the autocorrelations for all $\tau > 1$ can be regarded as statistically not significant, which also eliminates the necessity of accounting for the subtle periodic features in the ACF plot. In this case, an AR(1) structure appears to be appropriate for the majority of the residual time series under study.

Based on the above information, the harmonic regression model was refitted with an additional AR(1) residual correlation structure

$$Z_t = A_0 + A_1 \cos\left(\frac{2\pi t}{12}\right) + A_2 \sin\left(\frac{2\pi t}{12}\right) + \epsilon_t \tag{2.8}$$
$$\epsilon_t = \psi \epsilon_{t-1} + v_t \ ,$$

where v_t are i.i.d $\mathcal{N}(0, \sigma_v^2)$ distributed. It can be easily shown that model (2.8) has a structured residual covariance matrix

$$\mathbf{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \psi & \cdots & \psi^{T-1} \\ \psi & 1 & \cdots & \psi^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \psi^{T-1} & \psi^{T-2} & \cdots & 1 \end{pmatrix}.$$

This means that A_0, A_1, A_2 can be estimated using generalized least squares described in section 1.3.2. The MLEs of ψ and σ^2 can be obtained accordingly as parameters of a mixed effect model. In this analysis, the `lme` function from `R` package `nlme` (Pinheiro *et al.*, 2016) was used to fit model (2.8). Continuing the example of the time series in the pixel located at $(33.275\text{E}^\circ, -2.375\text{N}^\circ)$, the estimated AR(1) coefficient of the residual model is $\hat{\psi} = 0.3368$. As shown in Table 2.2, the changes in the model coefficients are subtle, but estimates from model (2.8) have larger standard deviations. These changes in standard deviations are the result of accounting for the residual autocorrelation structure. It provides more reliable confidence intervals for statistical inference. The same analysis was carried out on pixels with $\geq 75\%$ data available.

TABLE 2.2: A comparison of the temporal regression model (2.5) and model (2.8) fitted to the LSWT time series from a pixel located at $(33.275\text{E}^\circ, -2.375\text{N}^\circ)$

	Intercept/$\hat{A}_0$	std $\hat{A}_0$	$\hat{A}_1$	std $\hat{A}_1$	$\hat{A}_2$	std $\hat{A}_2$
model (2.5)	24.81	0.0513	0.29	0.0731	1.01	0.0715
model (2.8)	24.81	0.1551	0.31	0.0904	1.01	0.0882

2.1.3 Temporal analysis summary

In general, model (2.8) is capable of capturing the main seasonal patterns in the LSWT data in individual pixels. An AR(1) structure is suitable for most of the residual series. However, such an analysis does not provide much information of the spatio-temporal patterns of the data. It is possible to compare the model results to examine how things change spatially. For example, the estimated coefficients, A_0, A_1 and A_2 from models fitted to different pixels can be compared, as well as the estimated AR(1) coefficient ψ and the residual variance σ^2. However, little information about spatial patterns can be concluded from this type of comparison. To some extent, the time series model (2.8) is only a basic interpretation of

the data. More advanced modelling techniques are required to uncover the spatio-temporal patterns in the LSWT data.

2.2 Investigating spatial patterns

Exploratory analysis of the remote-sensing LSWT data can also be carried out from the second perspective of modelling the spatial images. In this section, spatial patterns for individual LSWT images were investigated. Since most of the LSWT images are spatially smooth, spline regression models were considered. Analogous to section 2.1, the spatial autocorrelations were investigated using empirical variograms.

2.2.1 Bivariate spline regression

The modelling of the general spatial trend of the LSWT images was carried out using the spline regression technique. The model can be written as

$$Z_{(x,y)} = f(x,y) + \epsilon_{(x,y)} = \Phi(x,y)\boldsymbol{\beta} + \epsilon_{(x,y)}\,, \tag{2.9}$$

where $Z_{(x,y)}$ represents the LSWT in the pixel indexed with geographical coordinate (x,y)[ii], $f(x,y)$ is a smooth function, $\Phi(x,y) = (\phi_1(x,y), \ldots, \phi_K(x,y))$ is a vector of bivariate basis functions and $\boldsymbol{\beta}$ is the vector of basis coefficients β_k, $k = 1, \cdots, K$. Again the intercept of the model is included in the basis system, i.e. $\phi_1(x,y) = 1$. Various options are available for $\Phi(x,y)$, such as tensor spline basis and 2-dimensional Fourier basis. In this exploratory analysis, bivariate (or 2-dimensional) thin-plate regression splines were applied. Thin-plate spline is known for its adaptability to different dimensions. In terms of a 2-dimensional space, it has been shown to be equivalent to a kriging estimate of a spatial process with a special covariance function (Nychka, 2000). For this reason, it is an appropriate choice for modelling the spatial patterns.

The construction of the thin-plate regression splines begins with the basis and penalty of the full rank thin-plate splines. The full basis is then truncated to obtain a low rank smoother that 'optimally' approximates the full basis solution (Wood, 2003). Thin-plate spline smoothing finds the estimation of the smooth function $f(\boldsymbol{x})$, for $\boldsymbol{x} = (x_1, \cdots, x_q)$, by minimizing

[ii] It is appropriate to use the geographical coordinate directly here because Lake Victoria sits on the equator. Transformation would be needed for lakes in higher latitude, such as using the spherical trigonometry.

$$\| \boldsymbol{Z} - \boldsymbol{f} \|^2 + \omega \mathcal{P}_{mq}(f)\,, \tag{2.10}$$

where $\boldsymbol{Z}$ is the vector of observations, $\boldsymbol{f}$ is the vector of evaluated smooth functions, ω is the smoothing parameter and

$$\mathcal{P}_{mq}(f) = \int \cdots \int_{\mathcal{R}^q} \sum_{v_1+\cdots+v_q} \frac{m!}{v_1! \cdots v_q!} \left(\frac{\partial^m f}{\partial x_1^{v_1} \cdots \partial x_q^{v_q}} \right)^2 dx_1 \cdots dx_q \tag{2.11}$$

is the thin-plate penalty. Here $\mathcal{R}^q$ is the q-dimensional range space for $\boldsymbol{x}$ and m is chosen to satisfy $2m > q$. The minimizer of (2.10) is a function of the form

$$\sum_{i=1}^{N} b_i \varphi_{mq} \left(\| \boldsymbol{x} - \boldsymbol{x}_i \| \right) + \sum_{j=1}^{J} a_j \phi_j(\boldsymbol{x})$$

with basis $\varphi_{mq}(\centerdot)$ defined as in Wood (2003) and basis $\phi_j(\centerdot)$ being orthogonal to coefficient vector $\boldsymbol{b} = (b_1, \cdots, b_N)^\top$. By further introducing matrix notations, $\boldsymbol{\Phi} = (\phi_1, \cdots, \phi_J)^\top$, $\boldsymbol{a} = (a_1. \cdots, a_J)^\top$ and $\boldsymbol{E}$ with the (i,j)-th element $E_{ij} = \varphi_{mq} \left(\| \boldsymbol{x}_i - \boldsymbol{x}_j \| \right)$, the minimization criterion (2.10) becomes a constrained problem

$$\min_{b,a} \| \boldsymbol{Z} - \boldsymbol{E}\boldsymbol{b} - \boldsymbol{\Phi}\boldsymbol{a} \|^2 + \omega \boldsymbol{b}^\top \boldsymbol{E} \boldsymbol{b} \quad s.t. \quad \boldsymbol{\Phi}^\top \boldsymbol{b} = 0 \tag{2.12}$$

To achieve an 'optimal' low rank representation in the sense that minimal change is induced in the shape of the smooth function as determined by criterion (2.12), Wood (2003) showed that for a specific rank k, the appropriate solution is to set $\boldsymbol{b} = \boldsymbol{U}^{(k)} \boldsymbol{b}^{(k)}$, with $\boldsymbol{U}^{(k)}$ being the first k columns of the eigenvector matrix $\boldsymbol{U}$ from the eigen-decomposition $\boldsymbol{E} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$. Further constrain the vector $\boldsymbol{b}^{(k)}$ to the space satisfying $\boldsymbol{\Phi}^\top \boldsymbol{b} = 0$ by setting $\boldsymbol{b}^{(k)} = \Delta^{(k)} \tilde{\boldsymbol{b}}$, where $\boldsymbol{\Phi}^\top \boldsymbol{U}^{(k)} \Delta^{(k)} = 0$ for a certain column orthogonal matrix $\Delta^{(k)}$. Problem (2.12) can then be transformed into an unconstrained problem, with $\boldsymbol{D}^{(k)}$ denoting the top left $k \times k$ sub matrix of $\boldsymbol{D}$, which is

$$\min_{\tilde{b},a} \| \boldsymbol{Z} - \boldsymbol{U}^{(k)} \boldsymbol{D}^{(k)} \Delta^{(k)} \tilde{\boldsymbol{b}} - \boldsymbol{\Phi}\boldsymbol{a} \|^2 + \omega \tilde{\boldsymbol{b}}^\top \Delta^{(k)\top} \boldsymbol{D}^{(k)} \Delta^{(k)} \tilde{\boldsymbol{b}}\,. \tag{2.13}$$

Solving this optimization problem would give the thin-plate regression spline with degrees of freedom k. The selection of k is sometimes not very critical due to the presence of the smoothing parameter ω (Wood, 2003), as the actual EDF would be controlled by ω, which can be selected using methods such as (G)CV, REML and information criteria. Thin-plate regression splines can be implemented using the function `gam` in `R` package `mgcv`.

The investigation of LSWT images of Lake Victoria was carried out using model (2.9). The default setting in `gam` gives $q = 2$, $m = 2$ and the thin-plate penalty

$$\mathcal{P}_{22}(f) = \int\int_{\mathcal{R}^2} \left(\frac{\partial^2 f}{\partial x^2}\right)^2 + \left(\frac{\partial^2 f}{\partial x \partial y}\right)^2 + \left(\frac{\partial^2 f}{\partial y^2}\right)^2 dxdy.$$

A maximum basis degrees of freedom of $k = 20$ was used as input. The model estimation was again carried out using the REML method, where the smooth parameter was estimated as part of the random effect component. According to Wood (2011), this method is '*less prone to local minima than the other criteria, and may therefore be preferable*'. More details on the REML estimation of smooth component is given in section 2.2.3. The resulting EDF of the models range from 15 to 19, which produces a reasonable level of smoothness for the majority of the fitted images. The variance explained by the spline regressors ranges from 50% to 80%. Figure 2.3 shows a plot of the LSWT data in June 1997 in the left panel and the fitted LSWT from the thin-plate spline regression model in the right panel. The two plots were created using the same colour scheme, so that comparison of the spatial patterns can be made easier through colours. The estimated model has EDF = 18.48 and 75.1% of the variance is captured by the spline regressors. It can be said that the thin-plate regression spline generated a smooth image which captured the main patterns in the data.

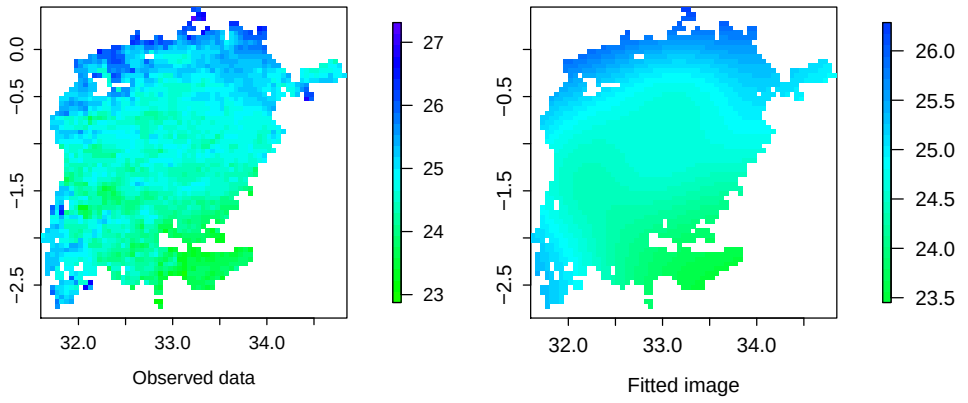

FIGURE 2.3: (Left) Image of the LSWT data in June 1997. (Right) The fitted image from the thin-plate spline regression model (2.9). The horizontal and vertical axes represent longitude and latitude respectively; the unit of the legend is C°.

2.2.2 Spatial correlation

While the bivariate spline regression model accounts for a substantial part of the spatial structure in the data, the residuals may still display a certain level of spatial correlation.

Ignoring this may lead to an under or overestimation of the standard errors and would have certain impact on statistical inference. Therefore, the model residuals were investigated for spatial correlation structures. This was achieved by examining the empirical variograms, under the second-order stationary assumption. Recall the definition of a (semi-)variogram as equation (1.8) in section 1.3.3. Its empirical version can be produced in the same way as that of the 'temporal variogram' $\hat{\gamma}(\tau)$ in section 2.1.1, with the distance in time τ replaced by the distance in space $\boldsymbol{h}$. However, unlike τ, which can only point to one direction, $\boldsymbol{h}$ can point to any direction in the spatial domain. That is, being the same distance apart from the north and from the east could be different. In consequence, the spatial correlations in different directions may have different natures. This property of a spatial process is referred to as anisotropic. Whereas an isotropic spatial process would display the same correlation structure for all directions.

Initially, directional variograms were produced to examine the isotropic property. Four directions, 0, $\pi/4$, $\pi/2$ and $3\pi/4$, were considered and empirical variograms were computed for each of these directions. The left panels of Figures 2.4 and 2.5 show two examples of the directional variograms computed using the residuals from the thin-plate spline regression model (2.9) fitted to the LSWT data in June 1997 and September 2006. The distances are measured using degrees in geographical coordinate. The variograms of four directions also appear to diverge at distances $\geq 1.5^\circ$ (i.e. 150km in the equatorial region). However, considering the distance the divergence begins and the unbalanced number of observations in different directions at very large distances, it is perhaps appropriate to truncate the empirical variogram and treat the spatial process as isotropic. To some extent, it makes little sense to expect strong spatial correlation of LSWT in two pixels more than 100 km apart, especially in a huge lake with a long retention time like Lake Victoria (Kayombo & Jorgensen, 2006). Since most of the directional variograms of the LSWT images show similar features as in Figures 2.4 and 2.5, the directional variograms are replaced by an isotropic variogram $\gamma(\|\boldsymbol{h}\|)$ in the remainder of the analysis, where $\|\,\cdot\,\|$ represents the Euclidean norm. For simplicity, denote $h = \|\boldsymbol{h}\|$ and $\gamma(h) = \gamma(\|\boldsymbol{h}\|)$ for all that follows.

The right panels of Figure 2.4 and 2.5 present the empirical variograms computed without distinguishing directions. The black curves are the variograms and the red curves are the Monte Carlo envelopes computed based on 100 permutations. The envelopes set the limits of the behavior of a random spatial process without significant spatial correlation. As the two variograms from the June 1997 and September 2006 models exceed the envelopes, it is sufficient to say that there is evidence of spatial correlation in both residual processes. Again,

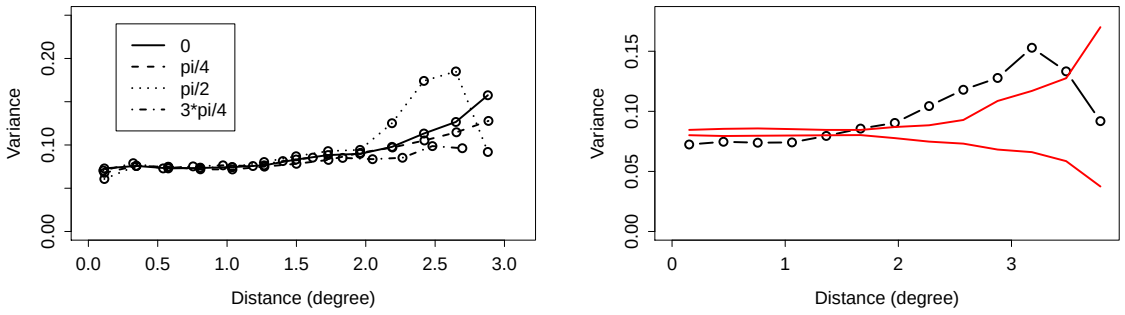

FIGURE 2.4: (Left) The directional variograms of residuals from model (2.9) fitted to the LSWT data in June 1997. The distances are measured in degrees. The four black curves show the directional variograms for 0, $\pi/4$, $\pi/2$ and $3\pi/4$. (Right) The omnidirectional variogram of residuals from the June 1997 model. The red curves represent the Monte Carlo envelop computed based on permutations.

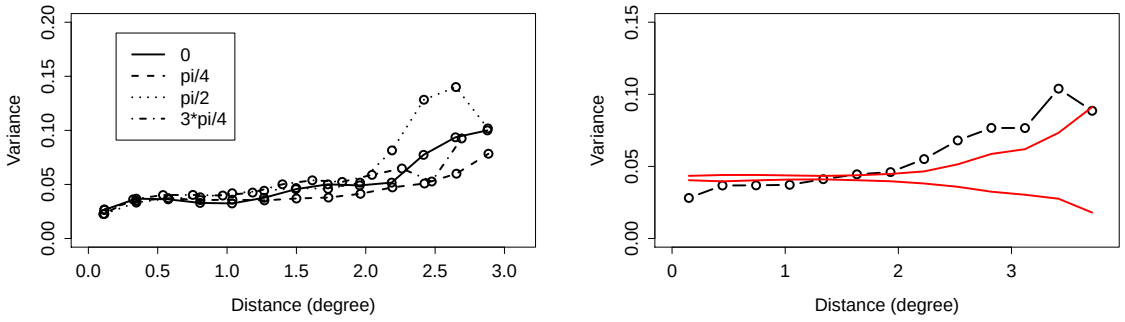

FIGURE 2.5: (Left) The directional variograms of residuals from the September 2006 model. The four black curves show the directional variograms for 0, $\pi/4$, $\pi/2$ and $3\pi/4$. (Right) The omnidirectional variogram of residuals from the September 2006 model. The red curves represent the Monte Carlo envelop computed based on permutations.

it is found that the majority of the variograms of the Lake Victoria LSWT images display similar patterns. Hence, the residual spatial correlations is to be incorporated to the spatial regression model (2.9).

There are a wide range of variogram models available for a stationary, isotropic spatial process. The general form of a variogram model can be written as

$$\gamma(h) = \sigma_{ng}^2 + \sigma_{ps}^2 \rho\left(\frac{h}{d}\right) . \tag{2.14}$$

In (2.14), σ_{ng} is the nugget effect, representing the variability at distances smaller than the sample spacing (including measurement errors). The parameter σ_{ps}^2 is sometimes referred to as the partial sill, which is the vertical distance between the the nugget and the value of $\gamma(h)$ as the distance $h \to \infty$. Function $\rho\left(\frac{h}{d}\right)$ describes the type of spatial correlation structure, with the range parameter d reflecting the distance from which the spatial correlation becomes zero (Cressie, 1993, Pinheiro & Bates, 2000). The most frequently used correlation functions

include the exponential, Gaussian, spherical and rational quadratic. A class of variogram models that offer great flexibility in modelling is the Matérn family,

$$\gamma(h) = \sigma^2_{ng} + \sigma^2_{ps}\left\{1 - \frac{2}{\Gamma(\nu)}\left(\frac{h}{2d}\right)^{\nu} K_\nu\left(\frac{h}{d}\right)\right\}. \tag{2.15}$$

The model is indexed by parameter ν, which governs the shape of the curve through the gamma function $\Gamma(\nu)$ and the modified Bessel function of order ν, $K_\nu(\cdot)$. Two special cases of the Matérn model are exponential ($\nu = 0.5$) and Gaussian ($\nu \to \infty$) (Cressie, 1993).

Model (2.15) was fitted to the empirical variograms of the residuals from the spline regression model (2.9) using function `variog` and `variofit` in the `geoR` package (Ribeiro Jr & Diggle, 2016). The results tend to be very sensitive to the initial inputs of ν, d and σ^2_{ng}. However, a relatively stable estimate can be achieved using a grid search of the optimal values of one or two parameters, so function `variofit` does not need to perform the optimization of all three parameters simultaneously. For example, to examine the residuals from the thin-plate regression model (2.9) of June 1997, a grid search of the optimal values ν and σ^2_{ng} was carried out and function `variofit` only estimated the range parameter d.

The investigation of the LSWT images with $\geq 77.8\%$ observations (1800 out of 2313 pixels) available showed that the majority of the models have an estimated index of the Matérn model at around 0.5, suggesting that an exponential correlation structure is appropriate for most of the LSWT images. Therefore, the exponential model was taken as representative of the residual spatial correlation structures of the Lake Victoria LSWT data. The following equations give the correlation and covariance functions of an exponential model,

$$\begin{aligned} \rho(h) &= \exp\left(-\frac{h}{d}\right), \\ \gamma(h) &= \sigma^2_{ng} + \sigma^2_{ps}\left\{1 - \exp\left(-\frac{h}{d}\right)\right\}. \end{aligned} \tag{2.16}$$

Model (2.16) is used in the modelling of the LSWT images with added spatial covariance structure in the next stage.

2.2.3 Bivariate spline regression with spatial covariance

In the previous model (2.9), the residuals were assumed to be i.i.d. $\mathcal{N}(0, \sigma^2)$ distributed. In this section, a spatial correlation structure was imposed on the residuals based on the

evidence obtained from the variogram modelling, giving the new model

$$Z_{(x,y)} = \Phi(x,y)\boldsymbol{\beta} + \epsilon_{(x,y)} = \Phi(x,y)\boldsymbol{\beta} + S(x,y) + \nu_{(x,y)} \,. \tag{2.17}$$

Here $S(x,y)$ is a zero mean stationary spatial process and the covariance matrix of the model is $\mathbf{Cov}[\boldsymbol{S}+\boldsymbol{\nu}] = \boldsymbol{\Sigma}_\gamma$, where $\boldsymbol{S}$ and $\boldsymbol{\nu}$ are the vectors of $S(x,y)$ and $\nu_{(x,y)}$ respectively. The (i,j)-th element of $\boldsymbol{\Sigma}_\gamma$ is determined by $\gamma(h_{ij})$, where $h_{ij} = \parallel (x_i,y_i) - (x_j,y_j) \parallel$. Specifically, the covariance matrix of the nugget effect component $\boldsymbol{\nu}$ is $\sigma^2_{ng}\boldsymbol{I}$ and the covariance matrix of $\boldsymbol{S}$ has its (i,j)-th element evaluated using function $\sigma^2_{ps}\rho(h_{ij})$. Model (2.17) can be regarded as a generalized additive mixed model (GAMM), which can be estimated using function `gamm` in the `R` package `mgcv`. The function is capable of modelling various spatial correlation structures, such as exponential, Gaussian, and spherical (Wood, 2011).

Some computational details are presented here. The estimation of the GAMM is carried out under the mixed model framework (Wood, 2006). The spline regressors $\Phi(x,y)\boldsymbol{\beta}$ are first divided into two parts, the fixed effect component $\Phi_f(x,y)\boldsymbol{\beta}_f$, which describe the linear spatial patten, and the random effect component $\Phi_r(x,y)\boldsymbol{\beta}_r$, which accounts for the spatial variation as higher order polynomials. This gives a model of the form

$$Z_{(x,y)} = \Phi_f(x,y)\boldsymbol{\beta}_f + \Phi_r(x,y)\boldsymbol{\beta}_r + \epsilon_{(x,y)} \,. \tag{2.18}$$

The random effect coefficient is assumed to follow the distribution $\boldsymbol{\beta}_r \sim \mathcal{N}(0, \frac{1}{\omega}\mathcal{S}_+)$, where ω is the smoothing parameter and $\mathcal{S}_+$ is a matrix associated with the eigen decomposition of the penalty matrix $\mathcal{S}$. The residuals are assumed to be normally distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$. Under this setting, the log-likelihood of the model can be written as

$$\mathcal{L}\left(\boldsymbol{Z}; \cdots\right) = -\frac{1}{2}\left\{\log(|\boldsymbol{\Sigma}|) + (\boldsymbol{Z} - \boldsymbol{\Phi}_f\boldsymbol{\beta}_f)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - \boldsymbol{\Phi}_f\boldsymbol{\beta}_f)\right\} + \text{constant} \tag{2.19}$$

where $\boldsymbol{\Sigma} = \frac{1}{\omega}\boldsymbol{\Phi}\mathcal{S}_+\boldsymbol{\Phi}^\top + \boldsymbol{\Sigma}_\epsilon$. The smoothing parameter can be selected using cross validation or the likelihood based methods. Imposing a spatial correlation structure on the residual process is equivalent to further decomposing $\epsilon_{(x,y)}$ into a spatial random component and a noise component as $S_{(x,y)} + \nu_{(x,y)}$ and parameterizing the covariance matrix $\boldsymbol{\Sigma}_\epsilon$ as $\boldsymbol{\Sigma}_\gamma$. For the purpose of model fitting, $\boldsymbol{S}$ is written as a product of a random effect design matrix $\boldsymbol{\Gamma}$ and coefficient vector $\boldsymbol{\eta}$, where the elements of $\boldsymbol{\Gamma}$ are determined through a spatial correlation function $\rho(h)$ associated with $\gamma(h)$ (Kammann & Wand, 2003). As a result, in practice, model (2.17) becomes

$$
\begin{aligned}
\boldsymbol{Z} &= \boldsymbol{\Phi}_f\boldsymbol{\beta}_f + \boldsymbol{\Phi}_r\boldsymbol{\beta}_r + \boldsymbol{\Gamma}\boldsymbol{\eta} + \boldsymbol{\nu} \qquad (2.20)\\
\boldsymbol{\beta}_r &\sim \mathcal{N}(\mathbf{0}, \frac{1}{\omega}\mathcal{S}_+)\\
\boldsymbol{\eta} &\sim \mathcal{N}(\mathbf{0}, \sigma^2_{ps}\boldsymbol{I})\\
\boldsymbol{\nu} &\sim \mathcal{N}(\mathbf{0}, \sigma^2_{ng}\boldsymbol{I})\,.
\end{aligned}
$$

It can be seen that the only thing that needs to be changed in the log-likelihood function (2.19) is the covariance matrix, which is now $\boldsymbol{\Sigma} = \frac{1}{\omega}\Phi\mathcal{S}_+\Phi^\top + \boldsymbol{\Sigma}_\gamma$. In other words, the additive model with spatial covariance structure as in model (2.20), or equivalently model (2.17), can be estimated using the same approach as that of the more general additive model (2.18). This is an elegant method, but its application to the Lake Victoria LSWT data did not necessarily result in a conclusive model. Below are two examples from applying model (2.17) to the LSWT images.

In the first example, model (2.17) with an exponential covariance structure was fitted to the LSWT data in June 1997. The initial values of σ^2_{ng} and d were gauged from the empirical variogram. The computation time of the model was about 1 hour 15 minutes. The estimated parameters are $\hat{d} = 0.1194$ and $\hat{\sigma}^2_{ng} = 6.5 \times 10^{-9}$. The EDF of the additive model is 13.3, which is smaller than that of the spline regression model (2.9). This is expected as part of the spatial variation has now been accounted for by the spatial covariance model. Table 2.3 provides a detailed comparison of some statistics from model (2.9) and (2.17). The fitted LSWT images are not presented because the difference between the images is small.

TABLE 2.3: A comparison of the spline regression model (2.9) and the spline regression model with spatial covariance structure (2.17), fitted to the LSWT data of June 1997.

	EDF	variance of ϵ	adjusted R^2	range d	nugget σ^2_{ng}
simple model (2.9)	18.48	0.0832	0.749	×	×
spatial model (2.17)	13.3	0.1155	0.739	0.1194	6.5e-9

To further assess the fit of the spatial covariance model, the normalized residuals were examined. Normalized residuals are model residuals after taking into account the covariance structure. In terms of model (2.17), the model residuals are $\hat{\boldsymbol{r}} = \boldsymbol{Z} - \boldsymbol{\Phi}_f\hat{\boldsymbol{\beta}}_f - \boldsymbol{\Phi}_r\hat{\boldsymbol{\beta}}_r$ and the corresponding normalized residuals are $\boldsymbol{r}^* = \widehat{\boldsymbol{\Sigma}}_\gamma^{-1/2}\hat{\boldsymbol{r}}$, where $\hat{\boldsymbol{r}}$ are assumed to follow the distribution $\mathcal{N}(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}_\gamma)$. If the spatial covariance structure in $\widehat{\boldsymbol{\Sigma}}_\gamma$ truly reflects the spatial structure of $\hat{\boldsymbol{r}}$, then the normalized residuals should follow the standard normal distribution,

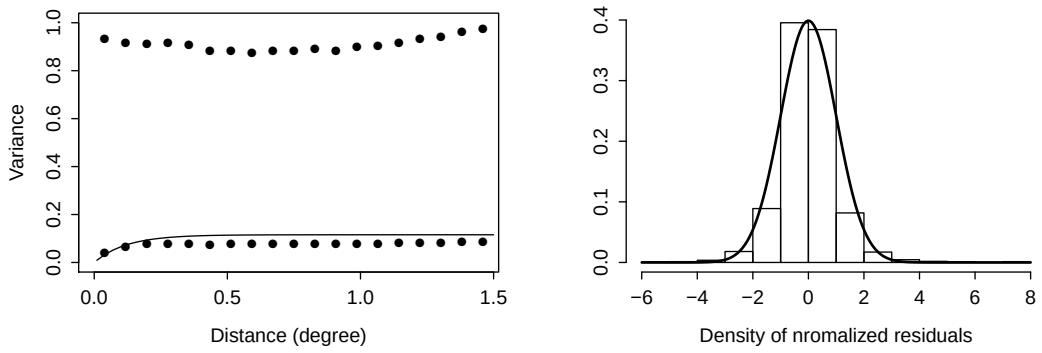

FIGURE 2.6: (Left) The empirical variogram of the model residuals (black dots at the bottom) and the normalized residuals (black dots at the top). The black curve represents the fitted variogram. (Right) The histogram of the normalized residuals. The black curve represents the $\mathcal{N}(0,1)$ density.

i.e. $\boldsymbol{r}^*$ should be a vector of white noises. Figure 2.6 provides a comparison of the spatial correlation structures of the model residuals and the normalized residuals for the model fitted to the June 1997 data. The left panel shows the empirical variograms of the model residuals (dots at the bottom) and the normalized residuals (dots at the top), with the fitted variogram plotted as a black curve. The right panel shows the histogram of the normalized residuals, with the imposed $\mathcal{N}(0,1)$ density as the black curve. Based on the information from the plots, it is sufficient to say that the normalized residuals are free from spatial structure and are approximately $\mathcal{N}(0,1)$ distributed.

The second example is an unsatisfactory fit as a result of the trade-off between various model components. It is taken from applying model (2.17) to the LSWT data in May 2007. Table 2.4 shows the comparison of the some statistics from model (2.17) and (2.9). In the model with spatial covariance structure, the spline regressors has EDF = 2 and residual variance 1.7043. Whereas the model without covariance structure has EDF = 18.27 and a much smaller residual variance of 0.1068. The estimated range parameter is 1.97 ($\approx$ 200km), which seems rather impractical for the Lake Victoria LSWT data.

TABLE 2.4: A comparison of the spline regression model (2.9) and the spline regression model with spatial covariance structure (2.17), fitted to the LSWT data of May 2007.

	EDF	variance of ϵ	adjusted R^2	range d	nugget σ^2_{ng}
simple model (2.9)	18.27	0.1068	0.67	×	×
spatial model (2.17)	2	1.7043	0.248	1.9727	8.9e-11

This problem is associated with the identifiability of $\boldsymbol{\Phi}_f\boldsymbol{\beta}_f$, $\boldsymbol{\Phi}_r\boldsymbol{\beta}_r$ and $\boldsymbol{\Gamma}\boldsymbol{\eta}+\boldsymbol{\nu}$, a phenomenon

linked to the spatial confounding of the covariates and random effects. In some situations, when spatially correlated errors (i.e. spatial random effect) are used to account for the spatial structure not explained by the model covariates (i.e. fixed effect), the parameter estimates would change substantially. This happens to many spatial regression models, including those estimated using a Bayesian approach. Discussion with respect to this issue can be found in Paciorek (2010), Reich & Hodges (2008), Wakefield (2007), etc. Hodges & Reich (2010) summarised several different interpretations of this phenomenon, including situations where the spatial random effects introduce or remove bias in the fixed effect coefficient, where there exists collinearity between the design matrices of the spatial random component and the fixed effect component, where the errors are correlated with the fixed effect, etc. It is not easy to attribute the identifiability problem in this exploratory analysis to one of these interpretations, especially when it applies to some, but not all LSWT images.

Strategies have been proposed over the years to deal with this issue, such as restricting the spatial random effect in a space orthogonal to the covariate space (Hodges & Reich, 2010, Hughes & Haran, 2013), investigating the scales of spatial variations of covariates and random effect to avoid spatial confounding (Paciorek, 2010) and the global/local smoothness of the spatial component (Lee *et al.*, 2014). However, it takes a lot of computational effort to implement these methods, which might not be practical when it comes to hundreds of high-resolution remote-sensing images. In addition, the results may not always improve the fit of the model (Pannullo *et al.*, 2016).

2.2.4 Spatial analysis summary

Due to the long computation time, 10 images with $\geq 65\%$ of data observed were analysed. Applying model (2.17) produced sensible results for most of the LSWT images investigated. Model computation time ranges from 40 minutes to 1 hour 20 minutes. EDF of the spline regressors ranges from 12 to 18 and is generally smaller than that of the model without covariance structure. An exponential variogram model is appropriate for the majority of the models. Model initialization appears to have a big influence on the final results, but robust estimates can be reached after trial and error. However, problems as illustrated in Table 2.4 could occur. In this investigation, 3 out of 10 images appeared to have this problem, where the fitted smooth function only contains the linear terms. The drawbacks with respect to the application of model (2.17) to the remote-sensing data, such as the Lake Victoria LSWT data, are summarised as follows.

First of all, the application of this model on sparse image data can be a problem. If there is a large area in the image without observations, then the estimation of the basis coefficients can be difficult due to the lack of information, which could affect the entire model fitting process. That is why the exploratory analysis was only conducted on images with relatively low percentages of missing observations.

Secondly, the identifiability problem presents another disadvantage. To some extent, it is hard to distinguish between the fixed and random effects due to spatial confounding and the complexity of the algorithm, even with distinct assumptions on each model component. Since the aim of the analysis is to understand the spatial and temporal patterns, a conclusive result would be far more appealing than a result which only provides a good fit to the data.

Finally, with the images explained by different models as a result of varying degrees of smoothness and covariance structures, the investigation of spatial patterns and their evolution is difficult. Just as the problems with the harmonic regression models in section 2.1.1, it makes little sense to compare the coefficients and the residual covariance structure from models fitted separately with no universal assumptions. Meanwhile, the computation time for these models is relatively long. As the size of remote-sensing data scales up quickly in both space and time, this method could eventually become computationally infeasible.

Therefore, more efficient methods are required to model the remote-sensing image time series. In particular, two aspects needs to be considered in terms of the alternative modelling strategy. (a) It helps to seek a more flexible and computationally efficient method to describe the covariance structure of the data. (b) It is better that the entire remote-sensing image time series can be handled simultaneously, i.e. to build a spatio-temporal model. One could follow the route of modelling the spatial or spatio-temporal covariance functions further. Lindgren & Rue (2015) described a flexible and efficient method based on the connection of the stochastic PDEs and the Gaussian fields. However, despite its fast computation using integrated nested Laplace approximation (INLA), the interpretation of the fitted model is not straightforward, which could be a problem of this analysis. In view of this, a different approach to investigating the spatial/temporal variation is proposed, the functional principal component analysis.

2.3 Functional principal component analysis (FPCA)

As introduced in section 1.3.1, the 'observation' in functional data analysis (FDA) is a smooth function representing the observations of an individual object. Statistical analysis is carried out at the function level. In terms of the remote-sensing data, this means the time series or images are first transformed into a collection of smooth univariate or bivariate functions. FDA techniques are then applied to these functions. With this approach, all the time series or images can be studied simultaneously, instead of 'one at a time'. Examples of the application of FDA to spatio-temporal environmental data include functional principal component analysis (FPCA) in Di Salvo *et al.* (2015), functional regression in Giraldo *et al.* (2009) and functional clustering in Haggarty *et al.* (2015). Among the statistical methods in the FDA family, FPCA is taken as the main approach to the investigation of the remote-sensing lake data.

FPCA is designed to provide an '*indication of the complexity of the data*' in the sense of the characteristics of functions (Ramsay & Silverman, 1997). This is a model-free approach for investigating the patterns of variations in the data and is often accompanied by a lower dimensional representation using the leading principal components (PCs). Although the interpretation may not be straightforward, the method is helpful in identifying the sources of variations in the data. The estimated results can be regarded as a non-parametric representation of the covariance structure of the data and may be used in further analysis.

2.3.1 The FPCA approach

Without loss of generality, consider data represented using univariate functions, $Z_i(t)$, $i = 1, \cdots, n$. According to Ramsay & Silverman (1997), the analysis begins with finding such a representation. A frequently used approach is to express the unknown function as a linear combination of a set of known basis functions $\phi_k(t)$,

$$Z_i(t) = \sum_{k=1}^{K} \beta_{ik}\phi_k(t) \ , \qquad i = 1, \ldots, n \ .$$

This representation can be written using matrix notation as $\boldsymbol{Z}(t) = \boldsymbol{B}\,\Phi(t)$, where $\boldsymbol{Z}(t)$ is a vector of data functions $Z_i(t)$, $i = 1, \ldots, n$, $\Phi(t)$ is a vector of basis functions $\phi_k(t)$, $k = 1, \ldots, K$ and $\boldsymbol{B}$ is a $n \times K$ coefficient matrix with its i-th row being $\boldsymbol{\beta}_i = (\beta_{i1}, \ldots, \beta_{iK})$. Different continuous and periodic constraints can be added to the basis representation. The

coefficients $\boldsymbol{\beta}_i$ are usually estimated by minimizing the (weighted) least squares criterion. Some frequently used bases include the Fourier basis, spline basis, polynomial basis, wavelet basis, etc. Attention is paid to how many features are to be retained from the data. This is often determined by the resolution and the curvature of the data and the type of question addressed with respect to the modelling. Model selection criteria may apply and some trade-offs might be required as well.

Analogous to a conventional PCA, the 'variables' in the FPCA are $Z(t)$ evaluated at all possible values of t. In theory, this means the number of 'variables' is infinite for continuous t. While in practice, there are usually a finite number of observations available at $t_1, \cdots, t_T$, so the 'variables' in the FPCA are $Z(t_1), \ldots, Z(t_T)$. Assuming zero mean for $Z(t_1), \cdots, Z(t_T)$, the covariance function for each pair of 'variables' $Z(t_j)$ and $Z(t_m)$ can be written as

$$\begin{aligned} V(t_j, t_m) &= \frac{1}{n} \sum_{i=1}^{n} \left[Z_i(t_j) - 0 \right] \left[Z_i(t_m) - 0 \right] \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[\sum_{k=1}^{K} \beta_{ik} \phi_k(t_j) \times \sum_{k=1}^{K} \beta_{ik} \phi_k(t_m) \right] , \end{aligned}$$

or in matrix notation,

$$V(t_j, t_m) = \frac{1}{n} \Phi(t_j)^\top \boldsymbol{B}^\top \boldsymbol{B} \Phi(t_m) \, . \tag{2.21}$$

The covariance matrix is then a square matrix with its elements being $V(t_j, t_m)$, for $j, m = 1, \ldots T$. The main idea of FPCA is to solve the eigenproblem

$$\int V(t_j, t) \xi(t) dt = \lambda \, \xi(t_j) \, , \tag{2.22}$$

subject to the orthonormal conditions, $\int \xi(t)^2 dt = 1$ and $\int \xi_p(t) \xi_q(t) dt = 0$ for all $p \neq q$. Solving equation (2.22) requires another finite approximation of the eigenfunction $\xi(t)$. This is done through another basis expansion using often the same basis as the one for constructing the functional data, i.e. $\xi(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \Phi(t)^\top \boldsymbol{c}$. Define the $K \times K$ matrix $\boldsymbol{W} = \int \Phi(t) \Phi(t)^\top dt$. The left hand side of equation (2.22) can be written as

$$\frac{1}{n} \int \Phi(t_j)^\top \boldsymbol{B}^\top \boldsymbol{B} \Phi(t) \Phi(t)^\top \boldsymbol{c} \, dt = \frac{1}{n} \Phi(t_j)^\top \boldsymbol{B}^\top \boldsymbol{B} \boldsymbol{W} \boldsymbol{c} \, .$$

Hence the approximated eigenproblem becomes

$$\frac{1}{n} \Phi(t_j)^\top \boldsymbol{B}^\top \boldsymbol{B} \boldsymbol{W} \boldsymbol{c} = \lambda \Phi(t_j)^\top \boldsymbol{c} \, . \tag{2.23}$$

Using the fact that the above equation holds for all values of t_j, along with the substitution $\boldsymbol{u} = \boldsymbol{W}^{1/2}\boldsymbol{c}$, the equivalent symmetric eigenproblem to equation (2.23) would be

$$\frac{1}{n}\boldsymbol{W}^{1/2}\boldsymbol{B}^\top\boldsymbol{B}\boldsymbol{W}^{1/2}\boldsymbol{u} = \lambda\boldsymbol{u}\,. \tag{2.24}$$

The maximum number of eigenvalues λ and eigenfunctions $\xi(t)$ that can be extracted from equation (2.24) is K (the degrees of freedom of the basis). The coefficient vector $\boldsymbol{c}$ in equation (2.23) can be obtained using $\boldsymbol{c} = \boldsymbol{W}^{-1/2}\boldsymbol{u}$ and the eigenfunction using $\xi(t) = \Phi(t)^\top\boldsymbol{c}$. The principal component score associated with the i-th function can be computed as

$$\alpha_i = \int \xi(t)Z_i(t)dt\,, \quad i = 1, \ldots, n, \tag{2.25}$$

In general, eigenfunctions $\xi(t)$ would carry information about the sources of variation in the data. Eigenvalues λ would indicate the proportion of variation explained by each principal component. Principal component scores α_i, which are mathematical realizations of the variation pattern, reflect the strength of the pattern in the i-th functional object $Z_i(t)$.

2.3.2 Extension to 2-dimensional data

The FPCA described above can be applied to 2-dimensional functional data through a straightforward generalization. Replace the univariate basis with a bivariate basis as

$$\boldsymbol{Z}(x, y) = \boldsymbol{B}\,\Phi(x, y)$$

and update the variance functions accordingly as

$$V(x_j, y_j, x_m, y_m) = \frac{1}{n}\Phi(x_j, y_j)^\top\boldsymbol{B}^\top\boldsymbol{B}\Phi(x_m, y_m)\,.$$

The eigenproblem with respect to bivariate functions becomes

$$\int V(x_j, y_j, x, y)\xi(x, y)dxdy = \lambda\,\xi(x_j, y_j)\,, \tag{2.26}$$

which can be solved using exactly the same approach as that used in solving eigenproblem (2.22). The code for computing the 2-dimensional FPCA was developed based on function `pca.fd` in the `R` package `fda` (Ramsay *et al.*, 2013). The trapezoidal rule is adopted here to approximate the double integrals essential to solving eigenproblem (2.26). Denote

$\Phi(x,y)\Phi(x,y)^\top = W(x,y)$, the integral can be approximated as

$$\begin{aligned}\boldsymbol{W} &= \int_{x_1}^{x_a}\int_{y_1}^{y_b} W(x,y)dxdy \qquad (2.27)\\ &\approx \triangle_x\triangle_y \left\{ \frac{1}{4}[W(x_1,y_1)+W(x_1,y_b)+W(x_a,y_1)+W(x_a,y_b)] + \sum_{i=2}^{a-1}\sum_{j=2}^{b-1} W(x_i,y_j) \right.\\ &\quad \left. + \frac{1}{2}\sum_{j=2}^{b-1}[W(x_1,y_j)+W(x_a,y_j)] + \frac{1}{2}\sum_{i=2}^{a-1}[W(x_i,y_1)+W(x_i,y_b)] \right\}\end{aligned}$$

where $(x_1,y_1),(x_2,y_1),\ldots,(x_1,y_2),(x_2,y_2),\ldots,(x_a,y_b)$ are quadrature points and $\triangle_x$ and $\triangle_y$ are the lengths of the intervals. While this is an approximation, a sensitivity analysis on univariate functions shows that there is no significant difference between the results using the trapezoidal rule and those using functions in package `fda`.

2.3.3 2-dimensional FPCA for reconstructed LSWT data

There are two ways of conducting FPCA on remote-sensing image time series data such as the LSWT, (a) transforming the time series data in each pixel into univariate functions and performing an analysis on the temporal curves, i.e. 1-dimensional FPCA, (b) constructing a collection of bivariate functions for the image at each time point and conducting an analysis on spatial images, i.e. 2-dimensional FPCA. The preference in this thesis is the second approach, as the 2-dimensional analysis has advantages over the 1-dimensional analysis in terms of the questions the thesis is trying to answer.

First of all, the majority of the remote-sensing images studied in this thesis are smooth by nature, so it is feasible to find a bivariate functional representation for the data. For example, it has been illustrated in section 2.2 that the patterns in the LSWT images can be captured using bivariate thin-plate regression splines. The situation might be slightly different for the Chlorophyll images with an algal bloom, but a smooth representation can be constructed for the majority of the images.

Secondly, as the remote-sensing data analysed in this thesis are more densely recorded in space than in time, bivariate functions are favoured in terms of dimension reduction. For example, for the Lake Victoria LSWT data, the representation of images using bivariate functions would result in 203 functional observations; whereas the representation of time series using univariate functions would give 2313 functional observations.

In addition, the influence of the basis needs to be considered. Since many environmental data appear to have a periodic pattern, the most straightforward choice of basis for the FPCA on temporal curves would be a Fourier basis. However, this would result in cyclical eigenfunctions that are only capable of identifying periodic patterns. Other interesting patterns, such as the long-term trend, would be beyond the capacity of the Fourier basis. It is possible to use other univariate bases, such as a spline basis, but this would require much higher degrees of freedom, which can be problematic for time series that covers a long period but is infrequently observed. This problem can be overcome by using a bivariate basis. As long as the images are relatively smooth, the degrees of freedom of the bivariate basis can be kept at a value much smaller than the number of observations per image. At the same time, the resulting eigenfunctions would have the flexibility to describe various types of spatial pattern in the data; the PC scores may also carry some information about the temporal patterns.

For an illustration, the 2-dimensional FPCA was applied to the 'Re LSWT' data set introduced at the end of Chapter 1. It is extracted from the ARC-Lake reconstructed LSWT data of Lake Victoria and is of dimension $26 \times 27 \times 203$. It was used here to avoid the computational problems brought to the FPCA by the high percentages of missing observations. Additional information on this issue is given in section 2.3.4.

The 'Re LSWT' data set was first centered by removing a monthly mean. Bivariate functional data were constructed as $Z_t(x, y) = \sum_{k=1}^{K} \beta_{tk}\phi_k(x, y)$, where $Z_t(x, y)$ is the reconstructed LSWT in the pixel indexed by longitude x, latitude y, at time point t. The bivariate basis used in this example is the tensor spline basis $\Phi(x, y)$, produced by meshing two univariate B-spline bases $\Phi_x(x)$ and $\Phi_y(y)$ through the Kronecker product. That is, $\mathbf{\Phi} = \mathbf{\Phi}_x \otimes \mathbf{\Phi}_y$, where $\mathbf{\Phi}_x$ and $\mathbf{\Phi}_y$ are the matrices of the univariate bases $\Phi_x(x)$ and $\Phi_y(y)$ respectively. For demonstration purpose, one knot each was placed in the median of the two coordinates x and y. This gives degrees of freedom of 5 to both $\Phi_x(x)$ and $\Phi_y(y)$ and degrees of freedom $K = 25$ to $\Phi(x, y)$. A formal way of selecting the basis dimension would involve methods such as cross-validation, information criteria, penalized regression, etc. This topic is discussed in detail in Chapter 3. Also note that the tensor spline basis is only one type of bases available for the 2-dimensional FPCA; other basis systems may be used for different data. In this example, 203 smooth bivariate functions were constructed using the tensor spline basis. Illustrations of constructing the functional representations $Z_t(x, y)$ from the LSWT data in June 1997 and September 2006 are shown in Figure 2.7.

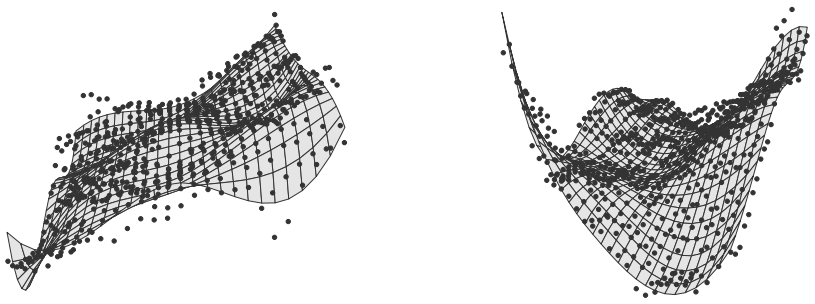

FIGURE 2.7: Illustrations of constructing functional representations using bivariate basis $\Phi(x, y)$ from June 1997 (left) and September 2006 (right). The dark grey dots represent the LSWT data and the light grey surfaces represent the functional observations.

The basis dimension $K = 25$ also suggests that the maximum number of PCs that can be extracted from the FPCA is 25. Applying the 2-dimensional FPCA algorithm gave the following results. Table 2.5 presents the eigenvalues of the first five functional PCs, along with their contributions towards the total variance explained. In this case, the first two PCs play a dominant role, accounting for 36.89% and 31.94% of the total variation respectively. The first five PCs together explain 92.80% of the total variation, which is sufficient to represent the entire data. Using the first five PCs to reconstruct the images results in a residual sum of squares (RSS) of 0.0078. Given that the variance in the centered data is 0.3514, this RSS value can be considered as relatively small. Note that a smaller RSS can be achieved by increasing the number of PCs used in the reconstruction.

TABLE 2.5: Eigenvalues of the first five functional principal components and their contribution towards the total variations evaluated in proportions.

	PC1	PC2	PC3	PC4	PC5
Eigenvalues	0.0223	0.0193	0.0077	0.0049	0.0020
Variance proportions	36.89%	31.94%	12.66%	8.08%	3.23%

The top two panels of Figure 2.8 present the eigenfunctions (or PC loadings) of the first two functional PCs, $\xi_1(x, y)$ and $\xi_2(x, y)$. In both plots, the blue end of the palette corresponds to positive loadings and the green end corresponds to negative loadings. A straightforward interpretation would be, the first eigenfunction displays a contrast between the north and south of the grid; the second eigenfunction shows a contrast between the east and west. In other words, PC1 and PC2 highlight the difference in the variation patterns between different parts of the lake area under study. The bottom two panels of Figure 2.8 display the scores of PC1 and PC2 obtained using the discretized version of equation (2.25). The scores can

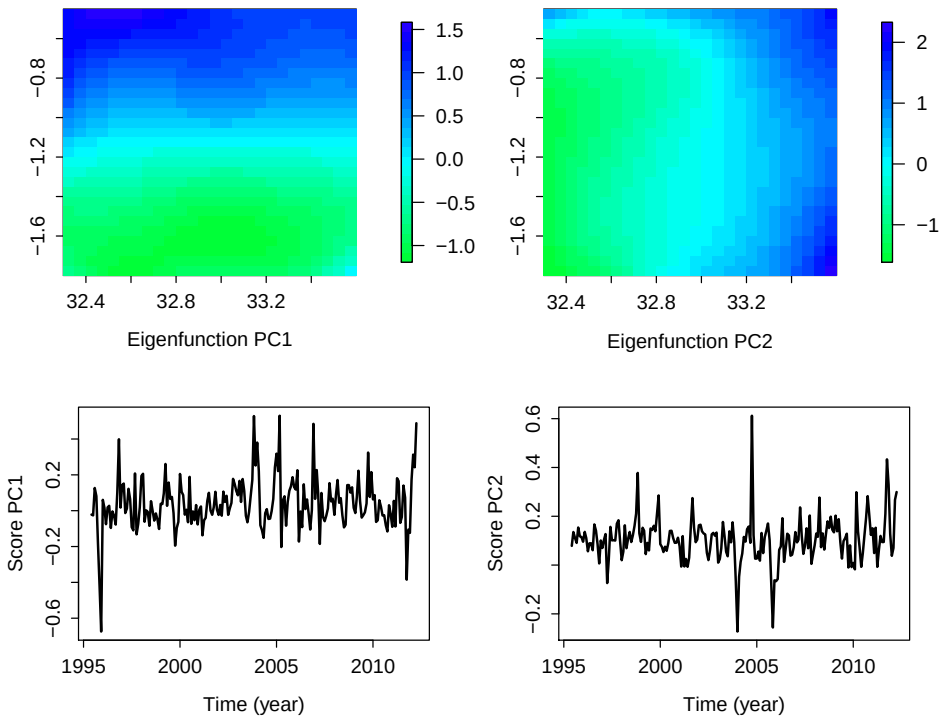

FIGURE 2.8: (Top) Illustrations of the eigenfunctions of PC1 and PC2. The horizontal and vertical axes are longitude and latitude respectively. (Bottom) Illustrations of the scores of PC1 and PC2 over time.

be regarded as indications of the temporal variations of the patterns shown in $\xi_1(x, y)$ and $\xi_2(x, y)$. Time series models may be applied to the scores to detect the existence of long-term trend, change points, etc. In this example, the score time series do not appear to show distinctive trend or change point.

As shown above, the 2-dimensional FPCA provides an efficient way to analyse the remote-sensing image time series. The analysis can be applied to the entire data set, not just a single time series or image and the computation of the example above took only 1 second. Even if the selection of the degrees of freedom of the basis is considered, the computational cost would still be much lower than that of the spatial regression model in section 2.2. Through functional data representation and keeping only the leading functional PCs, a dimension reduction can be achieved. For the above analysis, if the first five PCs are retained, then all the information required to reconstruct the original data are simply 25×5 basis coefficients and 5×203 PC scores. This is a significant reduction compared to the original data, which is of dimension $26 \times 27 \times 203$. In addition, the 2-dimensional FPCA can help to identify the common spatial patterns in the image time series using the extracted PCs. The temporal evolution of these spatial patterns may also be investigated through the PC scores.

2.3.4 Problems with respect to sparse data

Despite its efficiency, the FPCA described in section 2.3.1 and 2.3.2 may not be able to handle data with a high percentage of missing observations. The problem lies in the estimation of the coefficient matrix $\boldsymbol{B}$. As each column of $\boldsymbol{B}$ represents the coefficient vector of one functional data object, it could be impossible to estimate the coefficients if the observations corresponding to certain objects are too sparse. Unfortunately, this is an inevitable problem in the remote-sensing data, which is why the illustration in section 2.3.3 was presented using the ARC-Lake reconstructed LSWT data, not the real measurements, because the algorithm for FPCA simply cannot be implemented. Modifications are required to accommodate the missing observations.

James (2011) provided a summary of how to deal with missing observations in functional data analysis, where the approaches to FPCA were discussed in detail. Two methods widely discussed in research papers are the 'mixed effect model' and the 'local smoothing'. Both methods are designed to extract information from the entire data set when modelling the individual functions.

The mixed effect model approach was first proposed by James *et al.* (2000), where observations from individual objects are modelled as functions with random coefficients,

$$\begin{aligned} Z_i(t) &= \Phi(t)\boldsymbol{\beta} + \Phi(t)\boldsymbol{\eta}_i + \epsilon_i(t) \\ &= \Phi(t)\boldsymbol{\beta} + \sum_{p=1}^{P} \Phi(t)\boldsymbol{\theta}_p\alpha_{pi} + \epsilon_i(t)\ . \end{aligned} \tag{2.28}$$

The first part of (2.28) is a fixed mean function for the population; the second part of (2.28) is a random effect component, which models the variation unique up to the i-th object. The construction of the random effect $\sum_{p=1}^{P} \Phi(t)\boldsymbol{\theta}_p\alpha_{pi}$ is based on a Karhunen-Loève (K-L) expansion of a random process using a sequence of orthogonal functions. The authors named it the 'reduced rank principal component model', as only the leading P terms in the K-L expansion are used to approximate the process. The covariance structure can be modelled through $\boldsymbol{\eta}_i$, or equivalently $\boldsymbol{\theta}_p\alpha_{pi}$. It can be shown that $\Phi(t)\boldsymbol{\theta}_p$ is the equivalence of the eigenfunction $\xi_p(t)$ and α_{pi} is essentially the PC score (James *et al.*, 2000). Estimation of model (2.28) employs the EM algorithm, where α_{pi} is treated as missing information. Further development of the model can be found in Rice & Wu (2001), which offered more discussion on this method, Peng & Paul (2009), which proposed a Newton-Raphson algorithm for

model estimation, Gervini (2009), which described the model in a more general t-distribution setting, and Zhou & Pan (2014), which extended the method to a 2-dimensional case.

The local smoothing approach was described in detail in Yao *et al.* (2005), where a sparse longitudinal data set was analysed. The same idea was discussed in Di *et al.* (2009). The key to this method is to model the sparse functional data as noisy sampled points from a collection of trajectories with mean function $\mathbf{E}[Z(t)] = \mu(t)$ and covariance function $\mathbf{Cov}[Z(t), Z(u)] = V(t,u)$. First, observations Z_{it}, $i = 1, \ldots, n$, $t = 1, \ldots, T_i$, are stacked into a column vector to produce a mean function $\hat{\mu}(t)$ using a local linear smoother (kernel). Next the element in the raw covariance matrix is computed as

$$\widehat{V}(t_i, u_i) = [Z_{it} - \hat{\mu}(t_i)][Z_{iu} - \hat{\mu}(u_i)]$$

A second kernel is then applied to $\widehat{V}(t_i, u_i)$ to produce the smoothed covariance function $\widetilde{V}(t,u)$. Finally, eigenvalues $\hat{\lambda}_p$ and eigenfunctions $\hat{\xi}_p(t)$ are extracted from $\widetilde{V}(t,u)$. The PC scores α_{pi} are computed using the principal analysis by condition estimation (PACE) (Yao *et al.*, 2005) as

$$\hat{\alpha}_{pi} = \hat{\lambda}_p \hat{\boldsymbol{\xi}}_p^\top \widehat{\boldsymbol{\Sigma}}_i^{-1} \left(\boldsymbol{Z}_i - \hat{\boldsymbol{\mu}} \right) ,$$

where $\widehat{\boldsymbol{\Sigma}}_i = \widetilde{\boldsymbol{V}}_i + \hat{\sigma}^2 \boldsymbol{I}$ and $\hat{\sigma}^2$ is the estimated residual variance from the kernel smoothing. The subscript i indicates that $\widehat{\boldsymbol{\Sigma}}_i$, $\widetilde{\boldsymbol{V}}_i$ are different for each i due to missing observations. Further development of this method can be found in Di *et al.* (2014), Goldsmith *et al.* (2013), Zipunikov *et al.* (2011), where topics related to modelling high-dimensional multilevel data and constructing confidence bands for the estimated PCs were discussed.

In this thesis, the mixed model is favoured over the local smoothing approach due to the sparse features of the remote-sensing data. Recall the discussion in section 1.3.2 on different types of missing data mechanisms. It has been assumed that the type of missingness in the remote-sensing data in this thesis is missing at random (MAR). That is, the probability of the LSWT/Chl data being missing may depend on other observed variables, such as the longitude and latitude, but it is irrelevant to the values of the unobserved data. Further assuming that the parameters governing the missing data mechanism are distinct from the parameters in the model, the MAR condition means that the missing data mechanism can be ignored in the likelihood based inference process (Heitjan & Rubin, 1991, Lu & Copas, 2004). In view of this, the mixed model approach, implemented using the maximum likelihood method, is considered as an appropriate choice to analyse the remote-sensing data in this thesis.

According to Allison (2009), likelihood based inference minimizes the bias, maximizes the use of information in the data and provides asymptotic results for assessing the parameter estimates. It also has the advantage of automatically assigning the weights to each individual function to account for the impact of sparsity (James *et al.*, 2000).

The mixed model FPCA is explained in full detail in the next chapter, including its estimation method and a simulation study on the influence of sparsity on model fitting. The extension to accounting for the temporal correlations between remote-sensing images is investigated in Chapters 4 and 5.

Chapter 3

The mixed model FPCA for sparse image series

At the end of Chapter 2, two different approaches for performing FPCA on sparse data were introduced and the mixed model approach was favoured in terms of the analysis in this thesis. In this chapter, the mixed model FPCA method and its estimation procedure are described in detail. A comparison of the FPCA computed using direct matrix decomposition and the mixed model framework is carried out, which is followed by a simulation study assessing the method's performance with respect to sparse images. Applications of the method on the Lake Victoria LSWT and Chl data are presented at the end of the chapter.

3.1 The mixed model FPCA (MM-FPCA)

3.1.1 Model specification

Without loss of generality, consider a mixed model of n univariate random functionss $Z_i(t)$, $i = 1, \ldots, n$ and $t \in \mathcal{T}$,

$$Z_i(t) = \Phi(t)\boldsymbol{\beta} + \Phi(t)\boldsymbol{\eta}_i + \epsilon_i(t), \tag{3.1}$$

In this model, function $Z_i(t)$ is modelled through a collection of basis functions $\Phi(t)$, a fixed effect coefficient vector $\boldsymbol{\beta}$ and a random effect coefficient vector $\boldsymbol{\eta}_i$. The fixed effect $\Phi(t)\boldsymbol{\beta}$ is usually a mean function; whereas the random effect $\Phi(t)\boldsymbol{\eta}_i$ describes the unique effect of the i-th function. The covariance structure of the functions can be modelled through imposing constraints on $\boldsymbol{\eta}_i$. Based on this framework, James *et al.* (2000) proposed a reduced rank

functional principal component model. The idea is to represent the random effect using a truncated Karhunen-Loéve expansion (K-L expansion)

$$\begin{aligned} Z_i(t) &= \Phi(t)\boldsymbol{\beta} + \sum_{p=1}^{\infty} \xi_p(t)\alpha_{pi} \\ &\approx \Phi(t)\boldsymbol{\beta} + \sum_{p=1}^{P} \xi_p(t)\alpha_{pi} + \epsilon_i(t) \\ &= \Phi(t)\boldsymbol{\beta} + \Phi(t)\boldsymbol{\Theta}\boldsymbol{\alpha}_i + \epsilon_i(t). \end{aligned} \tag{3.2}$$

The first line of (3.2) is essentially a mean function $\Phi(t)\boldsymbol{\beta}$ plus a infinite order K-L expansion of a random process with zero mean and finite variance. Functions $\xi_p(t)$, $p = 1, \ldots, \infty$, are orthonormal functions, which form the basis of the K-L expansion. The component α_{pi}, $i = 1, \ldots, n$, are defined as $\int Z_i(t)\xi_p(t)dt$ following the properties of the expansion. The inclusion of the K-L expansion suggests that the representation using $\xi_p(t)$ and α_{pi} converges in mean square to the original random process as the expansion order goes to infinity (Alexanderian, 2013). A truncation is then applied so that only a finite number P of functions $\xi_p(t)$ are retained. The last line of (3.2) is simply to decompose $\xi_p(t)$ into a basis $\Phi(t)$ and the corresponding coefficient vector $\boldsymbol{\theta}_p$, so that $\sum_{p=1}^{P} \xi_p(t)\alpha_{pi}$ becomes $\sum_{p=1}^{P} \Phi(t)\boldsymbol{\theta}_p\alpha_{pi} = \Phi(t)\boldsymbol{\Theta}\boldsymbol{\alpha}_i$, where $\boldsymbol{\Theta}$ is a basis matrix with column vectors $\boldsymbol{\theta}_p$, $p = 1, \cdots, P$, and $\boldsymbol{\alpha}_i$ is a vector consisting of α_{pi}, $p = 1, \cdots, P$. As a result, the problem becomes a mixed model with random coefficient $\boldsymbol{\eta}_i = \boldsymbol{\Theta}\boldsymbol{\alpha}_i$. To ensure that the random effect is equivalent to a K-L expansion, the following model assumptions are required.

(a) The parameter matrix $\boldsymbol{\Theta}$ and the basis matrix $\boldsymbol{\Phi}$, which consists of $\Phi(t)$ evaluated at different values of t, are both column orthonormal, i.e. $\boldsymbol{\Theta}^\top\boldsymbol{\Theta} = \boldsymbol{I}$, $\boldsymbol{\Phi}^\top\boldsymbol{\Phi} = \boldsymbol{I}$. This is to make sure that the orthonormal constraints on functions $\xi_p(t)$ in eigenproblem (2.22) are satisfied.

(b) The random coefficient $\boldsymbol{\alpha}_i$ has distribution $\boldsymbol{\alpha}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda})$, where $\boldsymbol{\Lambda}$ is a diagonal matrix with diagonal elements λ_p, $p = 1, \cdots, P$

(c) The model residuals are i.i.d normal, i.e. $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})$.

(d) There is also a hidden assumption that the n functions $Z_i(t)$, $i = 1, \cdots, n$, are supposed to be independent.

This model is referred to as the MM-FPCA in all the content that follows, in order to distinguish from the FPCA described in section 2.3.1 of Chapter 2.

The connection between the MM-FPCA (3.2) and the FPCA in section 2.3.1, though not straightforward, can be explained by the properties of the K-L expansion, reproducing kernel Hilbert space and Mercer's representation theorem. Additional details are given in Appendix A.1. In general, the orthonormal function $\xi_p(t) = \Phi(t)\boldsymbol{\theta}_p$ is equivalent to the p-th eigenfunction; the random coefficient $\alpha_{pi} = \int Z_i(t)\xi_p(t)dt$ is equivalent to the score of the p-th principal component. As $\mathbf{Cov}[\boldsymbol{\alpha}_i] = \boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \cdots, \lambda_P\}$ suggests that $\mathbf{Var}[\alpha_{pi}] = \lambda_p$, it can be deduced that λ_p, $p = 1, \ldots, P$, are equivalent to the eigenvalues of the FPCA.

James *et al.* (2000) described the advantages of the method as being able to estimate the individual functions using all the observed data rather than just those from one individual object (e.g. time series and image). At the same time, it automatically adjusts the influence of the missing percentages for each individual object. Potential drawbacks of this method are the large number of parameters to be estimated and the occasional failure of convergence of the EM to a global maximum. These can sometimes be avoided by careful choice of initial values, which is discussed later in this chapter.

As the MM-FPCA (3.2) was inspired by sparse longitudinal data sets, most of the pioneering studies were carried out on univariate functional data, i.e. curves. However, apart from the potential computational cost, there is no restriction on the dimension of the functions in theory. In some situations, using multivariate functions may even be advantageous, such as modelling of a sequence of smooth images. In this thesis, a MM-FPCA using bivariate functions was proposed to model the sparse remote-sensing image time series. The model generalizes equation (3.2) to

$$Z_t(x, y) = \Phi(x, y)\boldsymbol{\beta} + \Phi(x, y)\boldsymbol{\Theta}\boldsymbol{\alpha}_t + \epsilon_t(x, y), \tag{3.3}$$

for $t = 1, \cdots, T$ and $(x, y) \in \mathcal{D}$. The change of the individual function index from i in model (3.2) to t in model (3.3) is to emphasize that the model is going to be applied to a time series of images. The basis vector $\Phi(x, y)$ is now a collection of bivariate functions defined on a 2-dimensional domain $\mathcal{D}$. The same assumptions as in model (3.2) apply, which are

$$\begin{aligned} &\boldsymbol{\Phi}^\top\boldsymbol{\Phi} = \boldsymbol{I}\ , \quad \boldsymbol{\Theta}^\top\boldsymbol{\Theta} = \boldsymbol{I}\ , \\ &\boldsymbol{\alpha}_t \sim\ \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda})\ , \quad \boldsymbol{\epsilon}_t \sim\ \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I})\ , \end{aligned}$$

where $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \cdots, \lambda_P\}$ is a diagonal covariance matrix. .

The design of the bivariate basis $\Phi(x, y)$ is usually motivated by the problem under study. For a regular shaped field, such as a rectangular grid, the basis $\Phi(x, y)$ can be constructed by taking the tensor product of two univariate bases and then applying an orthonormalization process so that $\boldsymbol{\Phi}^\top \boldsymbol{\Phi} = \boldsymbol{I}$ is satisfied. For an irregular field, triangulation is often applied and the bivariate basis is defined on each triangle. This technique has been presented in Ettinger *et al.* (2012), Guillas & Lai (2010) for modelling the ozone concentration using functional regression and Zhou & Pan (2014) for a 2-dimensional FPCA on Texas temperature data. Both applications take into account the effect of irregular boundaries. Penalty matrices and smoothing parameters may be used to control the smoothness of the functions. In Zhou & Pan (2014), a thin-plate penalty $\mathcal{P}$ is used to control the smoothness of both the mean function and the functional PCs. This gives $\omega_1 \boldsymbol{\beta}^\top \mathcal{P} \boldsymbol{\beta} + \omega_2 \sum_{p=1}^{P} \boldsymbol{\theta}_p^\top \mathcal{P} \boldsymbol{\theta}_p$ as an addition to the usual estimation criterion of the model. In circumstances where selecting smoothing parameters ω_1 and ω_2 is computationally intensive, alternative methods for establishing an appropriate degrees of freedom for the basis may be required. One approach could be directly specifying the degrees of freedom of the basis based on scientific/application background of the problem under study.

The estimated eigenfunctions $\xi_p(x, y)$, $p = 1, \ldots, P$, are the counterparts of the PC loadings in a PCA. In particular, the bivariate eigenfunctions assign weights to each point (x, y) in the range of support $\mathcal{D}$. It measures how much 'load' each point has on the p-th principal component. Under the scenario that $\mathcal{D}$ is a spatial field, the eigenfunctions can be regarded as the spatial patterns common to all functional objects. By default, $\xi_p(x, y)$, $p = 1, \ldots, P$, are ordered by the magnitude of the eigenvalues λ_p, showing their contributions to the total variation in decreasing order. The leading eigenfunctions usually display the most distinctive spatial variations in the data. In the MM-FPCA (3.3), the PC scores are estimated as the random components α_{pt}. They reflect how strong the pattern shown by $\xi_p(x, y)$ is in terms of the t-th functional objects. However, the scores need to be interpreted carefully as some distinctive values might be induced by the high proportions of missing observations.

3.1.2 Estimation of MM-FPCA

The main approach used here to estimate the MM-FPCA is maximum likelihood. Specifically, the EM algorithm is applied, with the coefficient of the random effect component estimated as the missing information (Rice & Wu, 2001). The log-likelihood functions and their expectations for the E-step and M-step iteration have been derived in (James *et al.*,

2000). Although the authors presented their results as 1-dimensional functional data, the extension to 2-dimensional functional data involves only a small change of the estimating equations. However, some modifications of the computational details are required. In the following paragraphs, the complete EM algorithm for the 2-dimensional MM-FPCA (3.3) is presented.

The EM algorithm is a general method for obtaining MLEs in incomplete data problems (Little & Rubin, 2002). As described in section 1.3.2 in Chapter 1, the algorithm consists of two steps, an E-step for the conditional expectation of the complete data log-likelihood and a M-step where the MLEs are produced by maximizing the E-step expectation with respect to the model parameters.

For the MM-FPCA, the complete data of the problem are $\boldsymbol{Z}_{com} = \{\boldsymbol{Z}_{1:T}, \boldsymbol{\alpha}_{1:T}\}$, where the subscript $1:T$ stands for the collection of data from time point 1 to T. The parameter set is denoted as $\Psi = \{\boldsymbol{\beta}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}, \sigma^2\}$. The complete data distribution of the model is then

$$f(\boldsymbol{Z}_{1:T}, \boldsymbol{\alpha}_{1:T}; \Psi) = \prod_{t=1}^{T} f(\boldsymbol{Z}_t, \boldsymbol{\alpha}_t; \Psi) = \prod_{t=1}^{T} f(\boldsymbol{Z}_t | \boldsymbol{\alpha}_t; \Psi) f(\boldsymbol{\alpha}_t; \Psi) \,, \tag{3.4}$$

where the product comes from the assumption that the functional data at different time points are independent. The conditional distribution $f(\boldsymbol{\alpha}_t | \boldsymbol{Z}_t; \Psi)$ can be derived from the joint distribution of data at time t. In the it-th iteration, the E-step calculates the expectation of the complete data log-likelihood given the observed data $\boldsymbol{Z}_t$ and the current parameter estimate $\Psi^{(it-1)}$

$$\mathcal{Q}\left(\Psi; \Psi^{(it-1)}\right) = \mathbf{E}\left[\mathcal{L}(\Psi; \boldsymbol{Z}_{1:T}, \boldsymbol{\alpha}_{1:T}) \,\middle|\, \boldsymbol{Z}_{1:T}, \Psi^{(it-1)}\right] \,, \tag{3.5}$$

where the conditional expectation is taken with respect to the missing information $\boldsymbol{\alpha}_t$, as $\mathbf{E}[\boldsymbol{\alpha}_t | \boldsymbol{Z}_t, \Psi^{(it-1)}]$. The M-step then updates the parameter set to $\Psi^{(it)}$, so that the condition

$$\mathcal{Q}\left(\Psi^{(it)}; \Psi^{(it-1)}\right) \geq \mathcal{Q}\left(\Psi; \Psi^{(it-1)}\right), \quad \forall\, \Psi \in \mathcal{W} \,, \tag{3.6}$$

is satisfied. The iterations of E-step and M-step terminate when the difference between certain measure of the fit of the model is smaller than a pre-determined threshold. Given this outline, a detailed algorithm can be established as follows.

Step 1: model distributions According to the model assumptions above, the distributions of $\boldsymbol{Z}_t$ and $\boldsymbol{Z}_t|\boldsymbol{\alpha}_t$ are

$$\begin{aligned}\boldsymbol{Z}_t &\sim \mathcal{N}(\boldsymbol{\Phi}_t\boldsymbol{\beta},\ \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top + \sigma^2\boldsymbol{I})\ ,\\ \boldsymbol{Z}_t|\boldsymbol{\alpha}_t &\sim \mathcal{N}(\boldsymbol{\Phi}_t\boldsymbol{\beta} + \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t,\ \sigma^2\boldsymbol{I}).\end{aligned}$$

Here $\boldsymbol{\Phi}_t$ is the basis matrix for the t-th function, where the subscript t is used to reflect the impact of missing data on the model. As the observed pixels vary with the images, the evaluated basis matrix $\boldsymbol{\Phi}_t$ would change accordingly. The joint density function $f(\boldsymbol{Z}_t, \boldsymbol{\alpha}_t)$ can be obtained using $f(\boldsymbol{Z}_t|\boldsymbol{\alpha}_t)f(\boldsymbol{\alpha}_t)$ as

$$\begin{aligned}f(\boldsymbol{Z}_t, \boldsymbol{\alpha}_t) = &\frac{1}{(2\pi)^{(n_t+K)/2}\sigma^{n_t}|\boldsymbol{\Lambda}|^{1/2}} \\ &\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta} - \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t)^\top(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta} - \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t) - \frac{1}{2}\boldsymbol{\alpha}_t^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\alpha}_t\right\},\end{aligned} \tag{3.7}$$

where n_t is the number of observations at time t. The conditional density function $f(\boldsymbol{\alpha}_t|\boldsymbol{Z}_t)$ can be derived using $f(\boldsymbol{Z}_t, \boldsymbol{\alpha}_t)/f(\boldsymbol{Z}_t)$ as

$$\begin{aligned}f(\boldsymbol{\alpha}_t|\boldsymbol{Z}_t) \propto \exp\Bigg\{&-\frac{1}{2\sigma^2}(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta} - \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t)^\top(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta} - \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t) - \frac{1}{2}\boldsymbol{\alpha}_t^\top\boldsymbol{\Lambda}^{-1}\boldsymbol{\alpha}_t \\ &+\frac{1}{2}(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta})^\top\left(\boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top\right)^{-1}(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta})\Bigg\}.\end{aligned}$$

Rearranging this with regard to random vector $\boldsymbol{\alpha}_t$ and recognizing the fact that it follows a normal distribution, gives the conditional distribution of $\boldsymbol{\alpha}_t|\boldsymbol{Z}_t$ as in the supplemental document of James *et al.* (2000)

$$\mathcal{N}\left(\left(\sigma^2\boldsymbol{\Lambda}^{-1} + \boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t\boldsymbol{\Theta}\right)^{-1}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}),\ \left(\boldsymbol{\Lambda}^{-1} + \frac{1}{\sigma^2}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t\boldsymbol{\Theta}\right)^{-1}\right). \tag{3.8}$$

Alternatively, conditional distribution (3.8) can be derived using the property of the following multivariate normal distribution (Zhou & Pan, 2014)

$$\begin{pmatrix}\boldsymbol{\alpha}_t \\ \boldsymbol{Z}_t\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{0} \\ \boldsymbol{\Phi}_t\boldsymbol{\beta}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Lambda} & \boldsymbol{\Lambda}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top \\ \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\Lambda} & \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top + \sigma^2\boldsymbol{I}\end{pmatrix}\right)$$

and then applying the Woodbury identity to the conditional expectation and variance,

$$\mathbf{E}[\boldsymbol{\alpha}_t|\boldsymbol{Z}_t] = \boldsymbol{\Lambda}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top\left(\boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top + \sigma^2\boldsymbol{I}\right)^{-1}(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta})$$

$$\mathbf{Cov}[\boldsymbol{\alpha}_t|\boldsymbol{Z}_t] = \boldsymbol{\Lambda} - \boldsymbol{\Lambda}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top\left(\boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\Lambda}\,.$$

Both the distributions in equation (3.7) and (3.8) are essential to the computation of the expectation in the E-step, which is then passed onto the M-step to get the MLEs.

Step 2: E-step equations Based on the conditional distribution (3.8), in the it-th iteration, the conditional expectations of the missing data $\boldsymbol{\alpha}_t$ and its quadratic $\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^\top$ evaluated at the current estimates of the parameter set, $\Psi^{(it-1)} = \{\boldsymbol{\beta}^{(it-1)}, \boldsymbol{\Theta}^{(it-1)}, \boldsymbol{\Lambda}^{(it-1)}, \sigma^{2(it-1)}\}$, can be computed as

$$\begin{aligned}\hat{\boldsymbol{\alpha}}_t &= \mathbf{E}\left[\boldsymbol{\alpha}_t \,\middle|\, \boldsymbol{Z}_t,\, \Psi^{(it-1)}\right] \\ &= \left[\sigma^{2(it-1)}\left(\boldsymbol{\Lambda}^{(it-1)}\right)^{-1} + \left(\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\right)^\top\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\right]^{-1}\left(\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\right)^\top\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}^{(it-1)}\right)\end{aligned} \tag{3.9}$$

$$\begin{aligned}\widehat{\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^\top} &= \mathbf{E}\left[\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^\top \,\middle|\, \boldsymbol{Z}_t, \Psi^{(it-1)}\right] \\ &= \mathbf{E}\left[\boldsymbol{\alpha}_t \,\middle|\, \boldsymbol{Z}_t, \Psi^{(it-1)}\right]\mathbf{E}\left[\boldsymbol{\alpha}_t \,\middle|\, \boldsymbol{Z}_t, \Psi^{(it-1)}\right]^\top + \mathbf{Cov}\left[\boldsymbol{\alpha}_t \,\middle|\, \boldsymbol{Z}_t, \Psi^{(it-1)}\right] \\ &= \hat{\boldsymbol{\alpha}}_t\hat{\boldsymbol{\alpha}}_t^\top + \left[\left(\boldsymbol{\Lambda}^{(it-1)}\right)^{-1} + \frac{1}{\sigma^{2(it-1)}}\left(\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\right)^\top\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\right]^{-1}.\end{aligned} \tag{3.10}$$

Plugging in (3.9) and (3.10) to the conditional expectation of the complete data log-likelihood at the current iteration $\mathcal{L}(\Psi^{(it-1)}; \boldsymbol{Z}_{1:T}, \boldsymbol{\alpha}_{1:T})$

$$\begin{aligned}&-\frac{1}{2}\sum_{t=1}^{T}\Bigg\{n_t\log\left(\sigma^{2(it-1)}\right) + \log\left(\left|\boldsymbol{\Lambda}^{(it-1)}\right|\right) + \boldsymbol{\alpha}_t^\top\left(\boldsymbol{\Lambda}^{(it-1)}\right)^{-1}\boldsymbol{\alpha}_t \\ &+\frac{1}{\sigma^2}(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}^{(it-1)} - \boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\alpha}_t)^\top(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}^{(it-1)} - \boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\alpha}_t\Bigg\} + \text{constant}\end{aligned} \tag{3.11}$$

gives the target function $\mathcal{Q}(\Psi; \Psi^{(it-1)})$ as defined in equation (3.5).

Step 3: M-step equations The $\mathcal{Q}(\Psi; \Psi^{(it-1)})$ function obtained above is then maximized with respect to each parameter component to obtain their MLEs. The estimating equations can be derived by solving the equations of the partial derivatives with respect to each parameter being zero. Particularly, the partial derivative with respect to $\boldsymbol{\Theta}$ is computed for each column $\boldsymbol{\theta}_p$ of $\boldsymbol{\Theta}$, because $\boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t$ is essentially $\sum_{p=1}^{P}\boldsymbol{\Phi}_t\boldsymbol{\theta}_p\alpha_{pt}$. The partial derivative of $\boldsymbol{\Lambda}$ is also derived for each diagonal element λ_p, $p = 1, \ldots, P$, based on the fact that

$$\log\left(|\boldsymbol{\Lambda}|\right) + \boldsymbol{\alpha}_t^\top \boldsymbol{\Lambda}^{-1} \boldsymbol{\alpha}_t = \log\left(\prod_{p=1}^{P} \lambda_p\right) + \sum_{p=1}^{P} \alpha_{pt}^2 \lambda_p \, .$$

As a result, the M-step equations based on the E-step predictions $\hat{\boldsymbol{\alpha}}_t$ and $\widehat{\boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^\top}$ are

$$\begin{aligned} \sigma^{2(it)} = \frac{1}{\sum n_t} \sum_{t=1}^{T} \Bigg[& \left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t \boldsymbol{\beta}^{(it-1)}\right)^\top \left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t \boldsymbol{\beta}^{(it-1)}\right) \\ & - 2\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t \boldsymbol{\beta}^{(it-1)}\right)^\top \boldsymbol{\Phi} \boldsymbol{\Theta}^{(it-1)} \hat{\boldsymbol{\alpha}}_t + \operatorname{tr}\left\{\boldsymbol{\Phi} \boldsymbol{\Theta}^{(it-1)} \widehat{\boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^\top} \boldsymbol{\Theta}^{(it-1)\top} \boldsymbol{\Phi}^\top\right\} \Bigg] \, , \end{aligned} \tag{3.12}$$

$$\lambda_p^{(it)} = \frac{1}{T} \sum_{t=1}^{T} \widehat{\boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^\top}_{(p,p)} \, , \tag{3.13}$$

for $p = 1, \ldots, P$, with $\widehat{\boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^\top}_{(p,p)}$ indicates the p-th diagonal element of $\widehat{\boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^\top}$, and

$$\boldsymbol{\beta}^{(it)} = \left(\sum_{t=1}^{T} \boldsymbol{\Phi}_t^\top \boldsymbol{\Phi}_t\right)^{-1} \sum_{t=1}^{T} \boldsymbol{\Phi}_t^\top \left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t \boldsymbol{\Theta}^{(it-1)} \hat{\boldsymbol{\alpha}}_t\right) \, , \tag{3.14}$$

$$\boldsymbol{\theta}_p^{(it)} = \left[\sum_{t=1}^{T} \widehat{\boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^\top}_{(p,p)} \boldsymbol{\Phi}_t^\top \boldsymbol{\Phi}_t\right]^{-1} \sum_{t=1}^{T} \boldsymbol{\Phi}_t^\top \left[\hat{\boldsymbol{\alpha}}_{t(p)} \left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t \boldsymbol{\beta}^{(it)}\right) - \sum_{j \neq p} \widehat{\boldsymbol{\alpha}_t \boldsymbol{\alpha}_t}_{(p,j)} \boldsymbol{\Phi}_t \hat{\boldsymbol{\theta}}_j\right] \, , \tag{3.15}$$

with $\hat{\boldsymbol{\alpha}}_{t(p)}$ represents the p-th element in vector $\hat{\boldsymbol{\alpha}}_t$, $\hat{\boldsymbol{\theta}}_j = \boldsymbol{\theta}_j^{(it)}$ for $j < p$ and $\hat{\boldsymbol{\theta}}_j = \boldsymbol{\theta}_j^{(it-1)}$ for $j > p$, for the basis coefficients. Note that the estimation of $\boldsymbol{\theta}_p$ needs to be done iteratively for all $p = 1, \ldots, P$. Strictly speaking, the M-Step in this algorithm is essentially the conditional maximization (CM) steps (Meng & Rubin, 1993) as the estimation of parameter $\boldsymbol{\theta}_p$ is conditioned on the estimations of $\boldsymbol{\theta}_j$, $j \neq p$. However, James *et al.* (2000) and some other authors still referred to it as the M-step. This thesis chooses to follow this tradition; whereas an explanation of the CM-steps is given in Chapter 4. After running through above equations, the current parameter set is updated to $\Psi^{(it)} = \{\boldsymbol{\beta}^{(it)}, \boldsymbol{\Theta}^{(it)}, \boldsymbol{\Lambda}^{(it)}, \sigma^{2(it)}\}$.

Step 4: evaluate convergence Choices of convergence criteria for the EM iterations include relative changes of the expected complete data log-likelihood, RSS, specific parameters, etc. For example, the criterion using relative change in the RSS values from two consecutive iterations is

$$\frac{\mathrm{RSS}^{(it)} - \mathrm{RSS}^{(it-1)}}{\mathrm{RSS}^{(it-1)}} \leq \varepsilon,$$

where ε is a pre-determined small value, such as 0.005, 0.0001.

Step 5: orthonormalize the results According to the assumptions of the MM-FPCA (3.3), the coefficient matrix $\boldsymbol{\Theta}$ is required to be orthonormal. However, the resulting MLE $\boldsymbol{\Theta}^*$ from the EM iterations is not guaranteed to have this property. Therefore, a final step of orthonormalizing $\boldsymbol{\Theta}^*$ is carried out. This is done through computing the covariance matrix of the estimated random effect component and then applying an eigen-decomposition to the covariance matrix,

$$\mathbf{Cov}[\boldsymbol{\Phi}\boldsymbol{\Theta}^*\boldsymbol{\alpha}^*] = \boldsymbol{\Phi}\boldsymbol{\Theta}^*\boldsymbol{\Lambda}^*\boldsymbol{\Theta}^{*\top}\boldsymbol{\Phi}^\top = \boldsymbol{\Xi}\boldsymbol{\Lambda}^{(new)}\boldsymbol{\Xi}^\top.$$

The columns of matrix $\boldsymbol{\Xi}$ give the final estimation of the orthonormal eigenfunctions $\xi_p(x, y)$ and $\boldsymbol{\Lambda}^{(new)}$ is the final approximation to the covariance matrix of the PCs. In practice, the eigen-decomposition $\boldsymbol{\Theta}^*\boldsymbol{\Lambda}^*\boldsymbol{\Theta}^{*\top} = \boldsymbol{\Theta}^{(new)}\boldsymbol{\Lambda}^{(new)}\boldsymbol{\Theta}^{(new)\top}$ is computed to avoid the manipulating of a very high dimensional matrix, as a result of the dimensionality of basis matrix $\boldsymbol{\Phi}$. The results are exactly the same since $\boldsymbol{\Phi}$ is orthonormal. The final version of the eigenfunctions is then computed as $\boldsymbol{\Xi} = \boldsymbol{\Phi}\boldsymbol{\Theta}^{(new)}$. In the end, the PC scores α_{pt} are re-estimated with the orthonomalized matrix $\boldsymbol{\Theta}^{(new)}$.

3.1.3 MM-FPCA initialization

Due to the complexity of the complete data likelihood in equation (3.11), the choice of starting values for the EM iterations is important. A sensible initialization method is essential to the convergence of the algorithm. Laird *et al.* (1987) suggested that '*Criteria for good starting values are (a) initial estimates can be obtained under all configurations of data and models, (b) if the closed form expressions of* $\hat{\sigma}$ *and* $\widehat{D}$ *exist, the method of obtaining starting values should find them*', where σ^2 represents the residual variance and D represents the random effect covariance matrix of the repeated measures model studied in the paper.

For the MM-FPCA (3.3), an initialization method based on the `R` package `fpca` is adopted with small modifications to $\sigma^{2(0)}$ and $\boldsymbol{\Lambda}^{(0)}$. The package is developed by Peng & Paul (2009) to implement the method described in their paper. It handles only univariate functional data, but the idea can be generalized to bivariate functional data in the following ways.

(a) The initial value of $\boldsymbol{\beta}^{(0)}$ is computed through fitting the model $\boldsymbol{Z} = \boldsymbol{\Phi}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ using vectorized data $\boldsymbol{Z} = \mathbf{vec}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_T)$.

(b) The residuals plus random effects are then calculated by subtracting the mean function from the data as $\hat{\boldsymbol{r}}_t = \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t = \boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}^{(0)}$.

(c) Rewriting $\boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t$ as $\boldsymbol{\Phi}_t\boldsymbol{\eta}_t$ and fitting the linear model $\hat{\boldsymbol{r}}_t = \boldsymbol{\Phi}_t\boldsymbol{\eta}_t + \boldsymbol{\epsilon}_t$ gives the least square estimates $\hat{\boldsymbol{\eta}}_t = (\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t)^{-1}\boldsymbol{\Phi}_t^\top\hat{\boldsymbol{r}}_t$. The fact that $\mathbf{Cov}[\boldsymbol{\eta}_t] = \boldsymbol{\Theta}\mathbf{Cov}[\boldsymbol{\alpha}_t]\boldsymbol{\Theta}^\top$ suggests that an eigenvalue decomposition of $\mathbf{Cov}[\hat{\boldsymbol{\eta}}_t] = \boldsymbol{U}\boldsymbol{\Sigma}_\alpha\boldsymbol{U}^\top$ can be used to initialize $\boldsymbol{\Theta}$ as $\boldsymbol{\Theta}^{(0)} = \boldsymbol{U}$. Note that a perturbation is sometimes added to the least square estimator of $\boldsymbol{\eta}_t$ to prevent $\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t$ from being singular, i.e. $\hat{\boldsymbol{\eta}}_t = (\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t + \kappa\boldsymbol{I})^{-1}\boldsymbol{\Phi}_t^\top\hat{\boldsymbol{r}}_t$, where κ is a small positive real number.

(d) The initial values of $\boldsymbol{\Lambda}$ and σ^2 are obtained as $\boldsymbol{\Lambda}^{(0)} = \boldsymbol{\Sigma}_\alpha$ and $\sigma^{2(0)} = \frac{1}{\sum n_t}\sum_{t=1}^{T}\hat{\boldsymbol{r}}_t^\top\hat{\boldsymbol{r}}_t$. The main idea is to avoid setting $\sigma^{2(0)}\boldsymbol{\Lambda}^{(0)-1}$ overwhelmingly larger than the product $\boldsymbol{\Theta}^{(0)\top}\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(0)}$. Otherwise, the conditional mean of $\boldsymbol{\alpha}_t|\boldsymbol{Z}_t$ in equation (3.9) would be driven towards zero by the factor $\sigma^{2(0)}\boldsymbol{\Lambda}^{(0)-1}$ and the algorithm might shortly converge to a biased solution.

3.1.4 MM-FPCA implementation

One of the major assumptions of the MM-FPCA is that the basis functions are orthogonal. There are several bivariate bases which are orthogonal by design and are capable of incorporating the shapes of the images, such as the bivariate B-spline, simplex splines, etc. However, building such bases usually involves complicated geometric partition of the spatial domain, e.g. the triangulation, and the quality of result often depends on the specific geometric design. As far as the problems in this thesis are concerned, the gains from using the advanced basis systems may not compensate the costs in implementing such bases. There are two main reasons. (a) The remote-sensing images are recorded in regularly spaced pixels, so the basis can be evaluated without additional geometric partition of the domain. (b) Due to the higher uncertainties in pixels towards the boundaries of the images[i], the modelling of the shapes of the images is not considered as a priority. Instead, the grid is trimmed to remove pixels that are irrelevant to the lake so the influence of the shape on the model can be minimized. A relatively simple method is then applied, which takes the tensor product of two univariate B-spline bases to construct a bivariate basis on a rectangular grid.

Since the bivariate bases created using the above method are usually not orthogonal, a transformation is also applied. Two bivariate functions being orthogonal refers to

$$\int_{\mathcal{D}} \phi_k(x,y)\phi_l(x,y)dxdy = 0\ , \quad \text{for } k \neq l\ .$$

[i]The retrieved remote-sensing data in the boundary pixels are often considered as not very reliable, because of the uncertainty in identifying whether a pixel is for land or water.

This integral can be approximated as

$$\sum_{i=1}^{n} \phi_k(x_i, y_i)\phi_l(x_i, y_i) \triangle_x \triangle_y, \qquad n \to \infty. \tag{3.16}$$

The orthonormalization of basis functions $\phi_1(x, y), \ldots, \phi_K(x, y)$ is carried out in discrete forms using approximation (3.16). The process involves evaluating the basis functions on a fine grid to obtain a basis matrix and then applying the transformation using the Cholesky decomposition. Some details of this process are presented below.

- The transformation using the Cholesky decomposition follows two steps. First decompose the product of the basis matrix as $\boldsymbol{\Phi}^\top \boldsymbol{\Phi} = \boldsymbol{L}\boldsymbol{L}^\top$, where $\boldsymbol{L}$ is an orthogonal lower triangular matrix. Then construct the orthogonal basis matrix as $\boldsymbol{\Phi}(\boldsymbol{L}^\top)^{-1}$. This is the method proposed in package `fpca`, which is essentially a linear transformation defined by Cholesky decomposition.

- Unlike univariate basis functions, there are two ways of constructing the basis matrix of bivariate basis functions $\phi_k(x, y)$. For basis functions defined on a 2-dimensional range space $\mathcal{D}$, the k-th column of the basis matrix $\boldsymbol{\Phi}$ can be created by concatenating the evaluations of $\phi_k(x, y)$ by either x or y and the results would be different. That is, a bivariate basis matrix constructed using $\boldsymbol{\Phi}_x \otimes \boldsymbol{\Phi}_y$ is different from that constructed using $\boldsymbol{\Phi}_y \otimes \boldsymbol{\Phi}_x$. However, it can be shown that the results after orthonormalization are essentially the same, subject only to a permutation of rows and columns. In other words, it will not affect the model fitting.

In order to implement the model, the values of two additional parameters need to be specified before starting the EM iterations described above. They are the degrees of freedom (or dimension) of the basis K and the order of the K-L expansion P. These two parameters control the smoothness of the functions and can be regarded as the 'smoothing' parameters of the MM-FPCA (3.3), although they function in a slightly different way as the smoothing parameter ω introduced in section 1.3.1. The basis dimension K can be chosen using model selection criteria, such as AIC/BIC, cross-validation or alternatively using a penalized approach (Zhou & Pan, 2014). The expansion order P can be selected using similar methods. James *et al.* (2000) also proposed the use of a plot of the expected log-likelihood evaluated at the MLEs against the expansion order, i.e. $\mathcal{L}(\Psi^*; \boldsymbol{Z}_{1:T}, \boldsymbol{\alpha}_{1:T}) \sim P$. The optimal choice is the value of P at which the curve becomes flat. For an approach that follows the tradition

of PCA, P can be selected by inspecting the magnitude of the variance of PCs relative to the total variance (Rice & Wu, 2001), i.e. the variance proportion criterion.

Considering the dimension of the remote-sensing data in general, both the cross-validation and penalized approach would be computationally expensive. Therefore, the selection of K and P based on information criteria is preferred. The variance proportion criterion is also considered for choosing expansion order P. Ideally, the two parameters should be selected simultaneously through a grid search. However, this again would be computationally intensive if higher basis dimension is required for the problem, as it could result in numerous combinations of P and K to search through. To overcome this problem, a simplified 2-stage approach is proposed. This approach handles the choice of K and P as two successive problems.

(a) The basis dimension K is selected first using the AIC/BIC. In order to select K initially, a sufficiently large P is used and is fixed throughout this stage. In practice, the sufficiently large P can be chosen by fitting a MM-FPCA with an arbitrary basis and inspecting the variance explained by various numbers of PCs. This idea of selecting the basis dimension K regardless of the expansion order P is similar to the method used in the FPCA described in section 2.3, where the basis dimension only depends on functional data representation and is not affected by the PCA that follows.

(b) Next the expansion order P is selected using the optimal basis decided in stage (a). The selection using AIC/BIC is relatively straightforward. Although, in some situations, a more practical approach may be used where a truncated expansion which provides a high enough approximation power is used instead of the one selected by the information criteria. The selection using variance proportion criterion is even easier to implement. First fit a full rank (or high rank) model and then select P so that at least $\delta\%$ of the total variation is explained, i.e.

$$\frac{\sum_{p=1}^{P} \lambda_p}{\sum_{p=1}^{K} \lambda_p} \geq \delta\% \ , \quad \text{for } P \leq K \ . \tag{3.17}$$

Another approach to the variance proportion criterion is illustrated in Zhou & Pan (2014). The authors fitted a series of models with increasing expansion order P until a PC with variance significantly smaller than other leading PCs appeared, then they set the expansion order as the current P.

In general, the selection of K and P should not be treated too rigidly. It is better to adapt the selection criteria to the purpose of statistical analysis. For example, the relative changes

of RSS and mean integrated squared error (MISE) from the fitted model may be used as they can be helpful in assessing whether it is signal or noise the model is trying to capture. Scientific knowledge associated with the application background may also play a part in the selection of the 'optimal' combination of K and P.

Code for implementing the MM-FPCA has been developed based on the `R` package `fpca` (Peng & Paul, 2013). An extension from univariate functions to bivariate functions has been made, which involves modifications of the basis matrix and its orthonomalization.

3.2 MM-FPCA investigation using image series

Several investigations on the MM-FPCA (3.3) were carried out to examine its performance on sparse remote-sensing image series. The first two studies were based on the 'Re LSWT' data set introduced at the end of Chapter 1, including a comparison between the MM-FPCA and the FPCA by eigenvalue decomposition (referred to as 'direct FPCA') and an investigation with respect to the basis dimension and expansion order. A simulation study was then carried out to assess the performance of the model under different levels of missing percentages and spatial missing patterns (i.e. missingness appearing as spatial regions).

3.2.1 MM-FPCA and direct FPCA

For a comparison between the MM-FPCA and the direct FPCA, model (3.3) was fitted to the 'Re LSWT' data set as used in section 2.3.3. The orthonormal basis was constructed by first creating the tensor spline basis $\mathbf{\Phi} = \mathbf{\Phi}_x \otimes \mathbf{\Phi}_y$, then applying the orthonormalization process described in section 3.1.4. The same degrees of freedom $K = 25$ as in section 2.3.3 was used. An example with 6 out of 25 resulting orthonormal bivariate basis functions is given in Figure 3.1. For comparison purpose, the full rank model with $P = 25$ was fitted. This gives a random effect component describing a space spanned by 25 PCs. It was also assumed that the mean function $\Phi(x, y)\boldsymbol{\beta} = 0$, which, after centering the data by removing the monthly means, can be regarded as appropriate.

The computation of the MM-FPCA took 136.8 seconds. The EM algorithm converged after 7 iterations. The estimated residual variance is $\hat{\sigma}^2 = 0.0049$. The covariance matrix of the estimated random effect was computed using the results from the EM iterations and the final eigen-decomposition $\mathbf{\Theta}^* \mathbf{\Lambda}^* \mathbf{\Theta}^{*\top} = \mathbf{\Theta}^{(new)} \mathbf{\Lambda}^{(new)} \mathbf{\Theta}^{(new)\top}$ was then applied to give

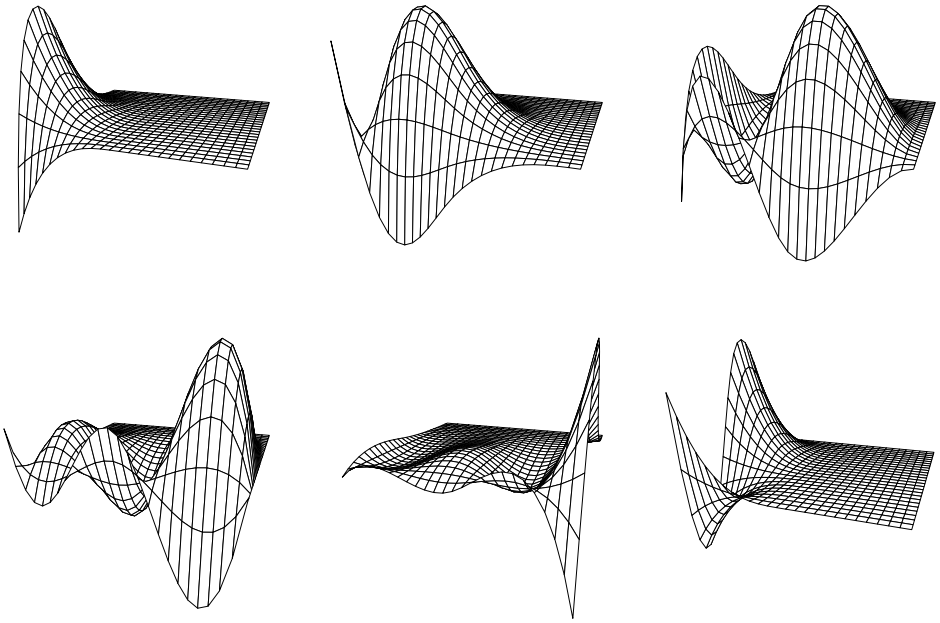

FIGURE 3.1: An example of six orthogonal bivariate basis functions from the basis (degrees of freedom = 25) defined on the rectangular grid of the 'Re LSWT' data set.

the eigenvalues $\hat{\lambda}_p$ as the diagonal elements of $\mathbf{\Lambda}^{(new)}$ and the eigenfunctions $\hat{\xi}_p(x, y)$ as the column vectors of $\mathbf{\Phi\Theta}^{(new)}$. Table 3.1 summarises the eigenvalues of the first five PCs and their contribution to the variance in proportions. In this case, the first PC is the most influential one, accounting for 35.54% of the total variation. The second and the third PC appear to be equally important, accounting for 21.06% and 19.48% of the total variation respectively. The leading five PCs in total explains 91.38% of the variation. The same measure from the direct FPCA is 92.80%. These results from the MM-FPCA are different from those extracted from the direct FPCA. This is understandable because the MM-FPCA contains a residual component $\boldsymbol{\epsilon}_t$, which does not exist in the direct FPCA. This component is certain to have some effect on the estimated eigenvalues and eigenfunctions. In this case, the estimated residual variance is $\hat{\sigma}^2 = 0.0049$. The RSS from reconstructing the original image using the first five PCs is 0.0081, which is almost the same as the RSS from the direct FPCA (0.0078) in section 2.3.3.

TABLE 3.1: The first five eigenvalues and their variance proportions from the MM-FPCA

	PC1	PC2	PC3	PC4	PC5
Eigenvalues	9.52	5.64	5.22	2.99	1.10
Variance proportions	35.54%	21.06%	19.48%	11.17%	4.11%

For a complete comparison, the eigenfunctions and scores of PC1 and PC2 extracted from

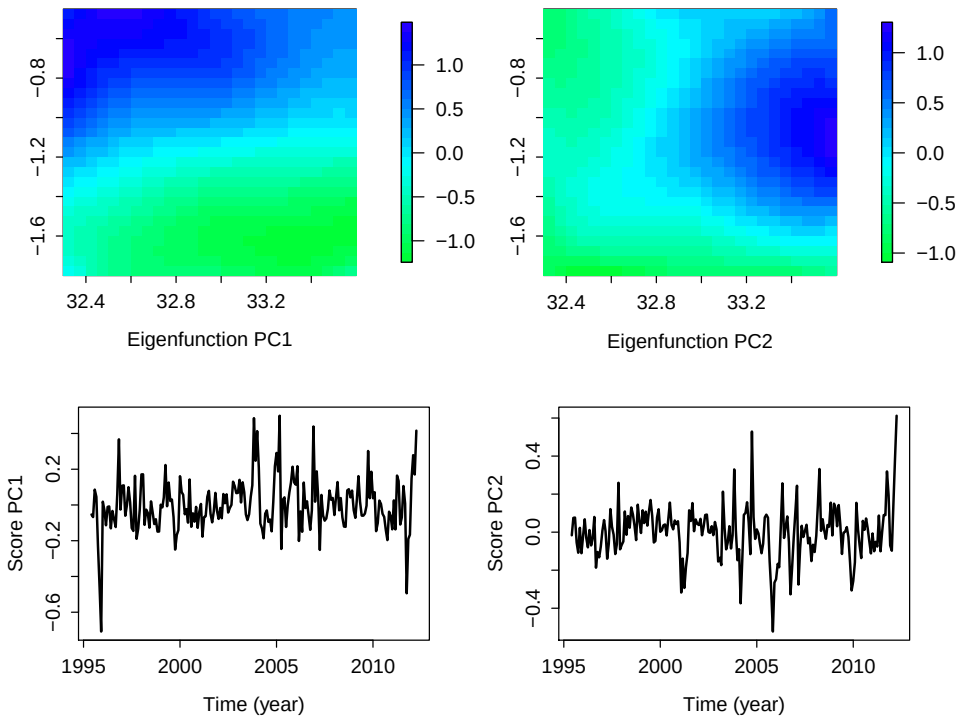

FIGURE 3.2: (Top) Illustrations of the eigenfunctions of PC1 and PC2. The horizontal and vertical axes represent longitude and latitude respectively. (Bottom) Illustrations of the scores of PC1 and PC2 over time.

the MM-FPCA were plotted in Figure 3.2. In order to make the comparison easy, these eigenfunctions were rescaled to match the eigenfunctions extracted from the direct FPCA in section 2.3.3. The rescaling was carried out using the following equation,

$$\tilde{\xi}_p(x,y) = \sqrt{\frac{\hat{\lambda}_p^{(m)}}{\hat{\lambda}_p^{(d)}}}\,\hat{\xi}_p(x,y)\ ,$$

where $\hat{\lambda}_p^{(m)}$ is the eigenvalues from the MM-FPCA and $\hat{\lambda}_p^{(d)}$ is the eigenvalues from the direct FPCA. The same was done to the scores, by changing the rescaling factor to $\sqrt{\hat{\lambda}_p^{(d)}/\hat{\lambda}_p^{(m)}}$. The resulting eigenfunctions and scores are different from their counterparts in section 2.3.3. However, the rescaled eigenfunction $\tilde{\xi}_1(x,y)$ describes a similar contrast between the north and south as that in the left panel of Figure 2.8. In terms of $\hat{\xi}_2(x,y)$, although the pattern is not exactly reflecting the contrast between East and West as the right panel in Figure 2.8, it conveys a similar idea.

It could be difficult to examine the similarity between the results from the two methods applied to sparse data, because the direct FPCA cannot even be implemented if the missingness is substantial. However, based on the performance of the two methods on complete data and

their theoretical connections, it is appropriate to used the MM-FPCA as an alternative to the direct FPCA. In fact, the MM-FPCA may even be a superior method, because it makes optimal use of the available information (James *et al.*, 2000). In practice, the assumption $\Phi(x, y)\boldsymbol{\beta} = 0$ is not required and $\boldsymbol{\beta}$ is estimated within the EM iterations (section 3.1.2).

3.2.2 Basis dimension and expansion order

The basis dimension K and the K-L expansion order P need to be selected before launching the EM algorithm. The following paragraphs continue the investigation using the 'Re LSWT' data set, but with emphasis put on the selection of K and P using the 2-stage approach described in section 3.1.4 and the influence of the choices on the model.

The first step in the 2-stage approach is to choose basis dimension K. The selection procedure starts with the 5×5 basis, which is the smallest possible basis with only one interior knot along each coordinate. Then one knot is added to one of the two coordinates each time to increase the basis dimension. That is, testing a sequence of basis of dimension 5×5, 6×5, 5×6, etc, until the maximum basis dimension considered is reached. For each basis tested, a MM-FPCA is fitted with a sufficiently large initial expansion order P_{ini}. The log-likelihood, AIC, BIC and RSS values are recorded. In this investigation, the maximum dimension was taken to be 7×7 and an initial $P_{ini} = 20$ was used[ii]. Table 3.2 presents some detail from this selection. In this case, the AIC and BIC failed to give an explicit answer as both criteria gave decreasing values as the basis dimension increases. However, there is a big drop in the AIC and BIC values after the basis dimension reaches 6×5, corresponding to a large increase in the log-likelihood. It also appears that the rapid decrease of AIC and BIC values slows down after the 7×6 basis. Therefore, a 7×6 basis is selected.

TABLE 3.2: The log-likelihood, AIC, BIC and RSS from the MM-FPCA fitted with increasing degrees of freedom.

Basis	5×5	6×5	5×6	6×6	7×6	6×7	7×7
likelihood	179470	181256	192678	197226	202064	202602	208166
AIC	-357848	-361211	-384055	-392899	-402323	-403398	-414232
BIC	-352461	-354788	-377631	-385232	-393413	-394488	-403872
RSS	0.0048	0.0046	0.0038	0.0035	0.0033	0.0033	0.0030

After determining the basis, the magnitude of the expansion order was investigated. For the likelihood based approach, both the method using the information criteria and the method

[ii]The maximum degrees of freedom of the basis can be chosen based on initial analysis, e.g. fitting spline regression models to individual images and examine the smoothness. It can be increased during the selection process, if the initial choice appears to be too low.

using the log-likelihood against expansion order plot were considered. For the 7×6 basis, models with expansion order ranging from 2 to 20 were tested. The log-likelihood, AIC, BIC and RSS were recorded and reported in Table 3.3. Again, the AIC and BIC did not give an explicit answer as the values keep on decreasing. However, from the plot of log-likelihood against expansion order P in Figure 3.3, it is possible to identify a point after which the log-likelihood curve becomes almost horizontal. Specifically, the dashed vertical line, corresponding to $P = 15$, indicates the point after which the increase of log-likelihood becomes smaller than 0.5%, i.e. $(\mathcal{L}_{P+1} - \mathcal{L}_P)/\mathcal{L}_P \leq 0.5\%$, where $\mathcal{L}_P$ is the log-likelihood of the model with expansion order P. It turns out that $P = 15$ is also the point where the AIC and BIC values reach an asymptotic. As a result, the expansion order $P = 15$ can be regarded as an appropriate choice.

TABLE 3.3: The log-likelihood, AIC, BIC and RSS from the MM-FPCA fitted with increasing expansion orders, when the basis dimension is fixed as 7×6.

P	2	...	13	14	15	...	19	20
log-like	91845	...	196609	198863	199907	...	202107	202194
AIC	-183432	...	-391928	-396351	-398353	...	-402494	-402583
BIC	-182160	...	-385564	-389563	-391140	...	-394009	-394736
RSS	0.0159	...	0.0035	0.0034	0.0033	...	0.0033	0.0033

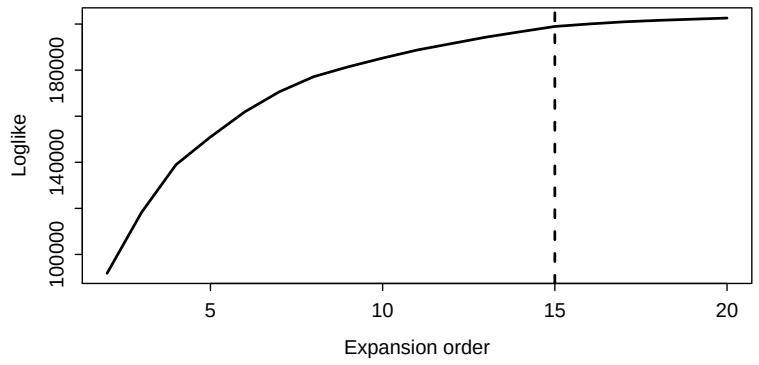

FIGURE 3.3: Illustrations of the selection of expansion order P using the log-likelihood. The black solid curve represents the log-likelihood; the dashed line indicates the point where the increase of log-likelihood becomes smaller than 0.5%. The horizontal axis represents P.

An alternative way to select the expansion order P is based on considering the variance proportion criterion, which is a method widely used in the PCA for multivariate analysis. To implement this criterion, a high rank model with $P = 30$ was fitted. According to this model, a 90% threshold for the variance explained gives expansion order $P = 6$; a 95% threshold gives $P = 8$; a 99% threshold would require $P = 15$. The variance proportion criterion can be attractive due to its computational efficiency, as it requires fitting a high rank model

only once. In this example, if the modelling purpose is to reduce the data complexity by retaining a small number of PCs, or to identify the PCs explaining the main patterns, then the variance proportion criterion would be helpful. If data imputation or reconstruction is of main interest, then the likelihood based approach might be preferred for a measure of the fit of the data.

To examine the impacts of the selected degrees of freedom, a model with $K = 7 \times 6$, $P = 15$ (denoted as the P15 model) and another with $K = 7 \times 6$, $P = 6$ (denoted as the P6 model) were fitted and the imputations were computed. The computation of the P15 model took 85.6s; the timer for the P6 model showed 20.6s. The estimated σ^2 for the P15 model is 0.0035 and that of the P6 model is 0.0058. The first two eigenfunctions $\xi_1(x, y)$ and $\xi_2(x, y)$ estimated from the P15 and P6 models are presented in Figure 3.4; examples of reconstructions from the two models are shown in Figure 3.5. The plots were produced using the same colour scheme for the purpose of comparison. The eigenfunctions from the two models are very similar to each other, so are the reconstructed images, indicating there is no substantial difference between the P15 and the P6 model.

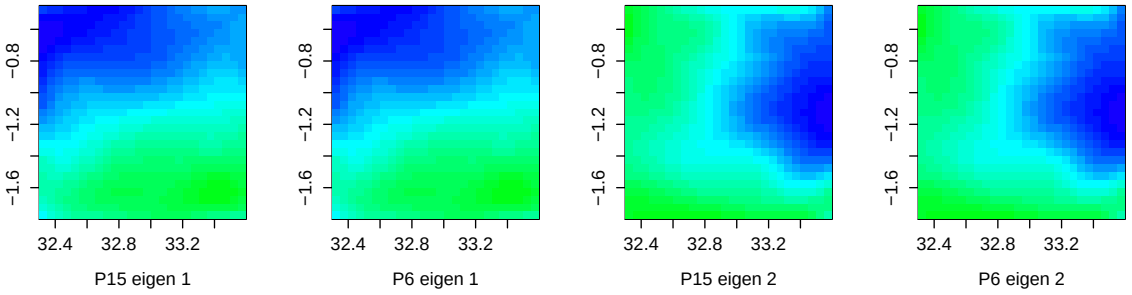

FIGURE 3.4: Examples of eigenfunctions from the MM-FPCA. The two left panels show the first eigenfunction from the P15 and P6 models. The two right panels show the second eigenfunction from the P15 and P6 models. TThe horizontal and vertical axes represent longitude and latitude respectively.

The above investigation suggests that if the main interest is in dimension reduction or to identify the dominant spatial patterns in the data, then a small model such as the P6 model would be sufficient. If the detail of the spatial reconstruction is of interest, then a larger model such as the P15 model may be a better choice. In addition, while the likelihood based approach and the variance proportion approach provide some information on the selection of the expansion order, the optimal choice in a real application may also be guided by the scientific background of the problem. It may be essential to consider the trade-off between the fit of the data and the identification of the actual signal, so that the dimension of the model is not increased merely for explaining the noise. Computation time and the number of

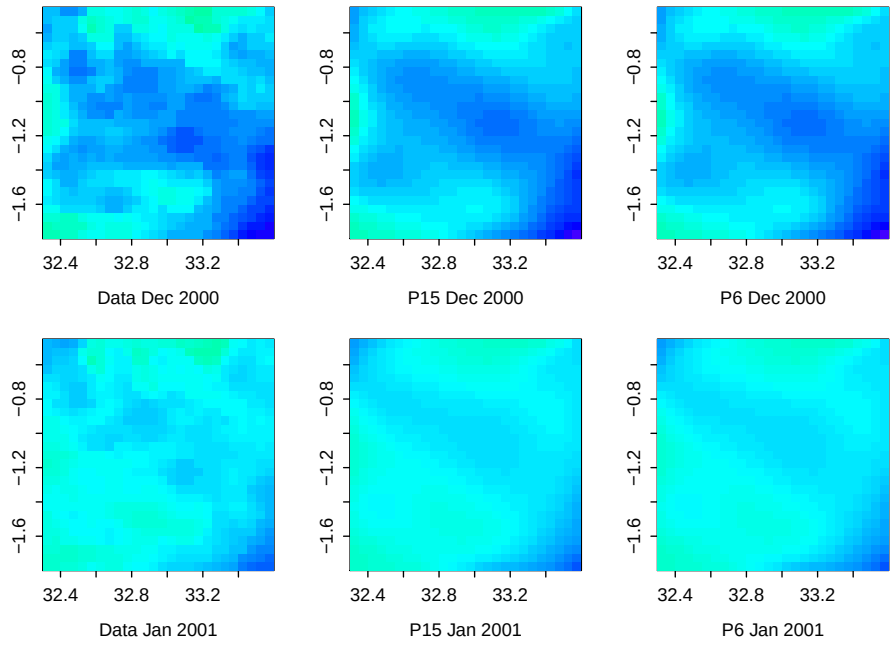

FIGURE 3.5: Examples of reconstructions from the MM-FPCA with $P = 15$ and $P = 6$ from December 2000 (top) and January 2001 (bottom). In each row, the left panel represents the data, the middle panel represents the P15 model and right panel represents the P6 model. The horizontal and vertical axes are longitude and latitude respectively.

observations available are among the other factors to be considered when choosing the most appropriate model.

3.2.3 Simulation study on missing conditions

Through its specification, the MM-FPCA handles the problem of missing observations automatically. However, the proportions of missing observations and the patterns of the missing data are presumed to have some impact on the model (Allison, 2009). The missing observations in remote sensing data are often the result of meteorological conditions and the satellite orbit. The percentage of missing observations in satellite images can be relatively high. In the Lake Victoria LSWT data, more than 1/5 of the images have less than 30% of the data observed and the total percentage of missing data reaches almost 50%. Another problem about the remote-sensing data is that missing observations often appear as spatial regions (recall Figure 1.2). This feature is referred to as spatial missing patterns in this thesis. It does not affect the assumption of missing at random (MAR) made earlier in the thesis as the probability the data are missing does not depend on the unobserved values, yet these missing regions would make recognizing the spatial patterns difficult.

To the author's knowledge, these issues have not received much investigation so far. Therefore, a simulation study on the impact of various conditions of sparsity on the model was carried out. The aim of this study is to investigate the applicability of the MM-FPCA to sparse data such as the remote-sensing LSWT. It also attempts to find the potential threshold for percentage of missing where the application of the MM-FPCA is not appropriate.

Part 1: simulation design Using the properties of Lake Victoria LSWT data as a guide, a 30×40 rectangular grid is defined and 120 images are simulated on this grid. Three levels, 0%, 30% and 50%, are assessed for the percentage of missing data. The last one is slightly higher than the missing proportion of the Lake Victoria LSWT data. Two types of sparsity are considered, one with spatial missing patterns and one without. The missing scenarios are then paired with four levels of spatial variation, giving 20 different scenarios in total (see Figure 3.6).

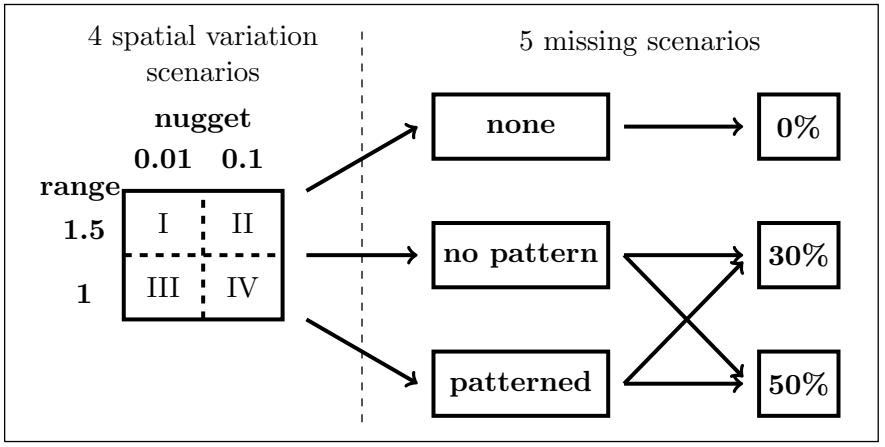

FIGURE 3.6: A diagram showing the settings of 20 simulation scenarios

Data are simulated pixel by pixel using function (3.18)

$$Z_t(x, y) = A(x, y) \cos\left[2\pi\nu(t - \varphi)\right] + S_t(x, y) + \epsilon_t(x, y) \ , \tag{3.18}$$

where $t = 1, 2, \ldots, 120$ and $(x, y) \in \mathcal{D}$, which covers a grid of size 30×40. The $A(x, y)$ component is the main spatial pattern designed for the simulated data, which corresponds to the leading eigenfunction. The sinusoid is used to mimic the seasonal fluctuation of the data with cycle length $\frac{1}{\nu} = 12$. An isotropic Gaussian random field (GRF) $S_t(x, y)$ adds noise in the form of spatial variations to the data. The component $\epsilon_t(x, y)$ is the i.i.d. random noise, which can be merged into the GRF as a nugget effect. The GRF is generated using the

covariance function (2.14), $\gamma(h) = \sigma^2_{ng} + \sigma^2_{ps}\rho(\frac{h}{d})$, where σ^2_{ng} is the nugget effect and $\rho(\frac{h}{d})$ is the exponential correlation function. The levels of spatial variation in the simulation study are created by varying the range and nugget parameters, d and σ^2_{ng}.

Simulation studies in the literature have focused mainly on the MM-FPCA's ability to identify the covariance structure and the corresponding orthogonal patterns. In Rice & Wu (2001), orthogonal designs were introduced to the covariance matrix in the simulation study and it has been shown that the leading eigenfunctions from their model (which is essentially a MM-FPCA) were able to capture these orthogonal functions. Similar results can be found in the simulation study in Peng & Paul (2009). The simulation study in this thesis does not intend to assess the model's ability to capture the patterns in the data covariance structure again. On the contrary, the emphasis is put on the fit of the model under different conditions for missing data. Therefore, only one spatial pattern $A(x, y)$ is included in the data generating function (3.18). No other spatial pattern orthogonal to $A(x, y)$ is involved for simplicity. This can be regarded as a problem with only one eigenfunction, with the GRF being responsible for the additional spatial variation[iii]. Efforts are made to design the appropriate missing data scenarios in this thesis.

To mimic the missing patterns in the remote-sensing LSWT data, the spatial missing clusters are introduced using the following method.

- First assign the missing percentages to each individual image. Assume that the cold season ($\frac{1}{4}$ of all time points) contributes 40% to the total missing percentage, the transitional period ($\frac{1}{2}$ of all time points) contributes 50% and the warm season ($\frac{1}{4}$ of all time points) contributes 10% to the total missing percentage. Under the 30% missing in total scenario, this is equivalent to assigning the missing percentage $(30\% \times 40\%) \div (\frac{1}{4}) = 48\%$ to the images corresponding to the cold season. The same applies to the rest of the seasons and the resulting missing percentages per image are shown in Table 3.4.

- Based on the above structure, the spatial missing clusters are created using probability maps. The maps are created by first generating several Gaussian random fields and then assigning different missing probabilities to different regions in the fields. For example, to create spatial clusters with three levels of missingness and a total of 48% missing observations, divide the GRF into regions using the 40-th and 65-th percentiles.

[iii] It is possible to include more than one eigenfunctions. There are several ways of generating orthogonal bivariate functions as described in literatures. Alternatively, eigenfunctions from a real problem may be used. Data can then be generated using the corresponding covariance matrix. This method is not used here due to different priority. A version of simulation using three bivariate eigenfunctions is available upon request.

Then assign probability 0.9 to regions with values below the 40-th percentile, 0.48 to regions with values between the 40-th and 65-th percentile and 0 to regions with values greater than the 65-th percentile. Table 3.4 summarises the critical percentiles and probabilities used to create the missing probability maps.

TABLE 3.4: Missing percentages per image, critical percentiles and probabilities used to create the missing probability maps.

Total %		**30%**			**50%**		
Cold	per image		48%			80%	
	percentile	40-th	65-th	100-th	20-th	80-th	100-th
	probability	0.9	0.48	0	1	0.9	0.3
Transitional	per image		30%			50%	
	percentile	25-th	50-th	100-th	35-th	68-th	100-th
	probability	0.9	0.3	0	0.9	0.6	0
Warm	per image		12%			20%	
	percentile	10-th	30-th	100-th	15-th	40-th	100-th
	probability	0.9	0.15	0	0.9	0.26	0

- Five missing probability maps are created for the missing 30% and 50% with spatial missing patterns scenarios respectively. Examples are shown in Figure 3.7, where the darker areas have higher missing probabilities and lighter areas have better data availability. The cold and transitional seasons are both assigned with two missing probability maps; whereas the warm season receives one map.

Using this design, the sparse images can be simulated. Figure 3.8 shows some examples of the generated images under the spatial variation scenario $d = 1$, $\sigma^2_{ng} = 0.01$ and the total missing percentage of 30%.

For each of the 20 scenarios, 200 replicates were simulated by varying $S_t(x, y)+\epsilon_t(x, y)$, where $S_t(x, y)$ were generated using the `R` package `RandomFields` (Schlather *et al.*, 2016). The MM-FPCA was fitted to each replicate. The number of replicates is chosen for computational efficiency. Evidence showing that 200 replicates are enough to produce robust estimation is given in Appendix A.2. The selection of basis and expansion order is not addressed in this study. Instead, a basis with degrees of freedom $K = 5 \times 5$ and an expansion order $P = 20$ are used throughout the replicates. Initial analysis show that $P = 20$ is sufficiently large for the simulated data under various scenarios. The number of PCs used in the data reconstruction is chosen using the variance proportion criteria, which in this case is $\delta\% = 90\%$. The following quantities are recorded for comparing the simulation scenarios,

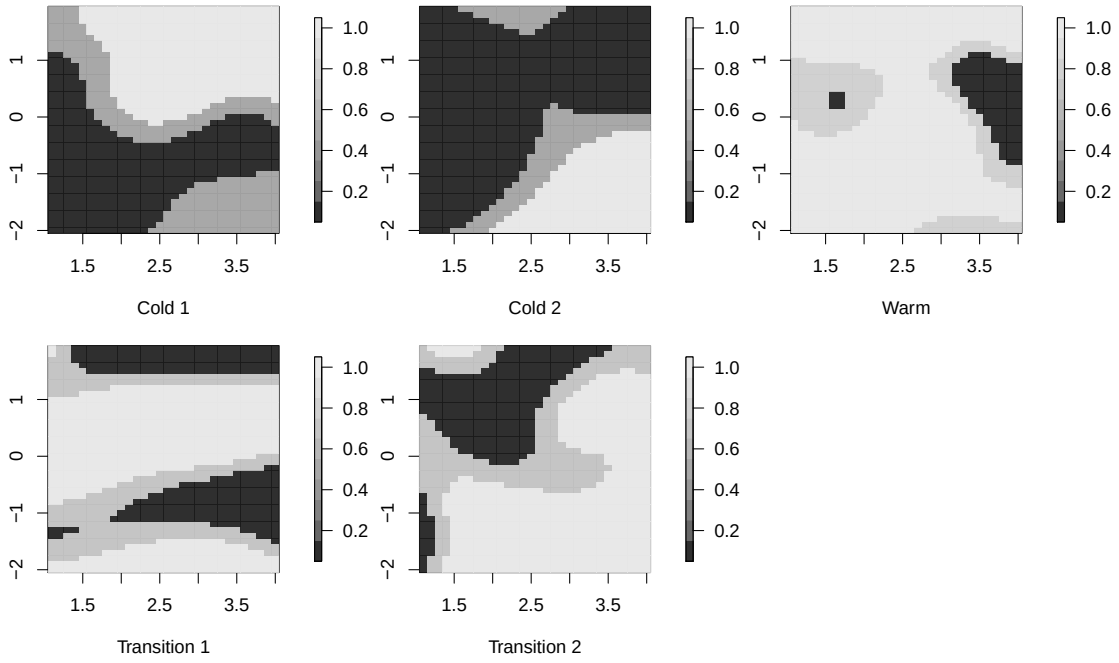

FIGURE 3.7: Missing probability maps for the 30% missing with spatial pattern scenarios. The legends show the probability that an observation is available. This means, the darker areas have higher missing probabilities and lighter areas have better data availability. The horizontal and vertical axes are longitude and latitude respectively.

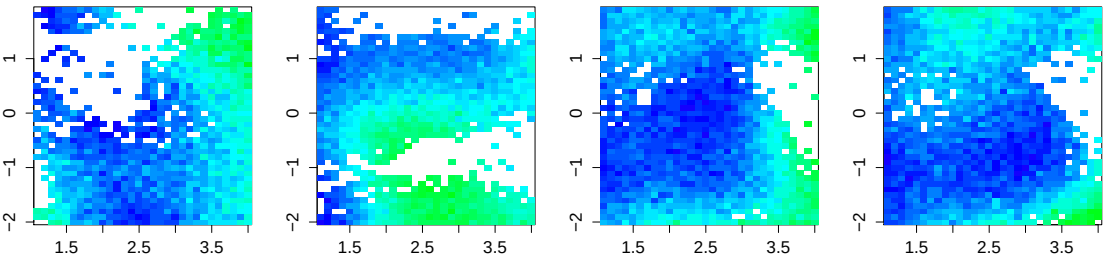

FIGURE 3.8: Examples of four simulated images from spatial scenario I, $d = 1$, $\sigma_{ng} = 0.01$. The missing condition is 30% missing with spatial pattern. The horizontal and vertical axes are longitude and latitude respectively.

(a) the estimated residual variance $\hat{\sigma}^2$ and the coefficient of the first eigenfunction $\hat{\boldsymbol{\theta}}_1$;

(b) the number of functional PCs kept for data reconstruction, P, which gives a minimum of 90% of the total variation;

(c) the mean integrated squared error (MISE) from reconstructing $\hat{Z}_t(x, y)$ using P functional PCs (Cardot, 2000, Ivanescu, 2013) for a global measure of the model performance. The detailed expression of the MISE can be written as

$$\mathbf{E}\left[\int_{\mathcal{D}}\left\{\hat{Z}(x,y)-Z(x,y)\right\}^2 dxdy\right] \approx \frac{1}{T}\sum_{t=1}^{T}\int_{\mathcal{D}}\left\{\widehat{Z}_t(x,y)-Z_t(x,y)\right\}^2 dxdy\,. \quad (3.19)$$

The means and standard deviations of these quantities based on the 200 replicates are then produced. In addition, the coefficient of the sinusoidal $A(x,y)$ at pixel location (x,y) is estimated from the reconstructions $\widehat{Z}_t(x,y)$ using a linear regression model

$$\widehat{Z}_{t,(x,y)} = A_{(x,y)}\cos\left[2\pi\nu(t-\varphi)\right] + \epsilon_{t,(x,y)}\,.$$

The resulting $\widehat{A}_{(x,y)}$ can be regarded as an approximation of function $A(x,y)$ evaluated at (x,y). A version of normed bias can be computed as

$$\widehat{\Delta}_{(x,y)} = \frac{1}{200}\sum_{m=1}^{200}\left|\widehat{A}_{m,(x,y)} - A(x,y)\right| \quad (3.20)$$

where m is the index of replicate. This measure is helpful in examining the impact of the spatial missing patterns.

Part 2: simulation results Table 3.5 summarises the means/medians and standard deviations of $\hat{\sigma}^2$, P, MISE for all 20 scenarios. The means of $\hat{\sigma}^2$ reflect the scale of the nugget effects. The small standard deviations of $\hat{\sigma}^2$ suggest that the estimates are robust in all 20 scenarios. The means and medians of P are consistent with the complexities in the simulation designs. The level of spatial variation appears to be the most influential on the expansion order. The larger the spatial variation (smaller d, larger σ^2_{ng}), the more PCs are required to reach a specified percentage of total variance. Meanwhile, the spatial missing patterns further weaken the signal, resulting in higher expansion orders being included in corresponding scenarios. The standard deviations of P turn out to vary greatly among 20 scenarios. This can be explained by the fact that P is chosen by a threshold. If the expansion order P happens to give a variance proportion at around the critical point, e.g. 89.9% and 90.1%, then a case with 89.9% would need to settle for $(P+1)$ PCs despite the small discrepancy in percentage values. The occurrences of the two cases out of 200 repetitions would affect the standard deviation of P.

After fitting the MM-FPCA, the data reconstructions are computed using P principal components. The MISE is then computed using these reconstructions. The results are also shown

TABLE 3.5: Means, medians (numbers in italics) and standard deviations (numbers in brackets) of $\hat{\sigma}^2$, expansion order P and MISE, rounded to three decimal places.

	spatial I $d = 1.5$, $\sigma^2_{ng} = 0.01$	**spatial II** $d = 1.5$, $\sigma^2_{ng} = 0.1$	**spatial III** $d = 1$, $\sigma^2_{ng} = 0.01$	**spatial IV** $d = 1$, $\sigma^2_{ng} = 0.1$
		None		
$\hat{\sigma}^2$	0.011	0.101	0.020	0.110
	(≤ 0.001)	(≤ 0.001)	(≤ 0.001)	(≤ 0.001)
P	6.010, *6*	6.005, *6*	9.445, *9*	9.540, *10*
	(0.099)	(0.122)	(0.498)	(0.499)
MISE	0.257	1.155	0.377	1.266
	(0.007)	(0.010)	(0.018)	(0.020)
		No pattern 30%		
$\hat{\sigma}^2$	0.011	0.101	0.020	0.110
	(≤ 0.001)	(≤ 0.001)	(≤ 0.001)	(≤ 0.001)
P	6.010, *6*	6.005, *6*	9.455, *9*	9.495, *9*
	(0.099)	(0.122)	(0.499)	(0.501)
MISE	0.258	1.153	0.378	1.273
	(0.007)	(0.010)	(0.018)	(0.020)
		No pattern 50%		
$\hat{\sigma}^2$	0.011	0.101	0.020	0.110
	(≤ 0.001)	(≤ 0.001)	(≤ 0.001)	(≤ 0.001)
P	6.000, *6*	6.000, *6*	9.380, *9*	9.355, *9*
	(0.000)	(0.141)	(0.486)	(0.473)
MISE	0.259	1.162	0.384	1.293
	(0.007)	(0.010)	(0.018)	(0.019)
		Pattern 30%		
$\hat{\sigma}^2$	0.011	0.101	0.019	0.109
	(≤ 0.001)	(≤ 0.001)	(≤ 0.001)	(≤ 0.001)
P	6.985, *7*	6.945, *7*	10.275, *10*	10.205, *10*
	(0.121)	(0.228)	(0.447)	(0.404)
MISE	0.260	1.162	0.406	1.310
	(0.009)	(0.013)	(0.018)	(0.016)
		Pattern 50%		
$\hat{\sigma}^2$	0.011	0.101	0.019	0.109
	(≤ 0.001)	(≤ 0.001)	(≤ 0.001)	(≤ 0.001)
P	7.000, *7*	6.985, *7*	10.670, *10*	10.360, *10*
	(0.100)	(0.157)	(0.471)	(0.481)
MISE	0.304	1.238	0.491	1.429
	(0.014)	(0.021)	(0.024)	(0.025)

in Table 3.5. The means of MISE do not appear to increase significantly in the missing without spatial pattern scenarios. However, the means of MISE do tend to inflate when the spatial missing patterns are introduced, especially in the missing 50% with spatial pattern scenario. This is highlighted in Table 3.6, where the percentage increments of the mean of MISE under different missing data scenarios from the complete data scenario are presented. The percentage increments under the missing 50% with pattern scenario are much higher than those from the rest of the scenarios. If the main purpose of the analysis is data imputation or prediction, then a 30% increase in the MISE may not be satisfactory. Although the MISE can be made smaller with a higher expansion order P, fitting a larger model using a

limited amount of data may itself be problematic.

TABLE 3.6: The percentage increments in the mean of MISE under different missing scenarios from the complete data scenario.

	spatial I $d = 1.5$, $\sigma^2_{ng} = 0.01$	**spatial II** $d = 1.5$, $\sigma^2_{ng} = 0.1$	**spatial III** $d = 1$, $\sigma^2_{ng} = 0.01$	**spatial IV** $d = 1$, $\sigma^2_{ng} = 0.1$
		No pattern		
30%	0.17%	0.26%	0.19%	0.57%
50%	0.79%	0.89%	1.84%	2.10%
		Pattern		
30%	1.07%	0.91%	7.58%	3.50%
50%	18.02%	7.52%	30.07%	12.87%

The simulated results of $\widehat{\Delta}_{(x,y)}$, which is the bias of estimated coefficient $\widehat{A}_{(x,y)}$ from 200 replicates, are presented as box plots in Figure 3.9. The two panels correspond to spatial variation levels I and IV respectively. The boxes show the distributions of the bias from all 1200 pixels and are ordered from left to right according to the complexity in the missing pattern design. The medians of bias from different scenarios are similar, but the chances of getting large values of $\widehat{\Delta}_{(x,y)}$ increase substantially in the missing 50% with spatial pattern scenario. The left panel of Figure 3.10 flags the the pixels with large bias from the spatial variation level IV plus missing 50% with spatial pattern scenario. The locations of these pixels match the darkest areas in the missing probability map of the season with the highest missing probability (i.e. the cold season), shown in the right panel of Figure 3.10. Since $\widehat{A}_{(x,y)}$ was estimated through a linear model with covariate, $\cos[2\pi\nu(t-\varphi)]$, the high missing probability in these areas during the cold season would make the estimation less accurate, introducing a large bias to $\widehat{A}_{(x,y)}$. Things can be worse for areas with few observations throughout the time. Inferences with respect to very sparse regions need to be carried out with extreme caution.

In addition to the above, the estimates of the first eigenfunction from all 20 scenarios are investigated via $\hat{\boldsymbol{\theta}}_1$ and the results tend to be robust throughout the simulation replicates. In particular, the first eigenfunction appear to show a similar pattern as the main spatial pattern $A(x, y)$ after rescaling using the standard deviation of the first PC. Some details can be found in Appendix A.2. In general, this study shows that the estimates from the MM-FPCA are robust throughout the simulation replicates. The analysis of MISE and the bias $\widehat{\Delta}_{(x,y)}$ suggests that the spatial missing patterns tend to be more influential than the percentages of missing. This is highlighted by the statistics from the missing 50% with

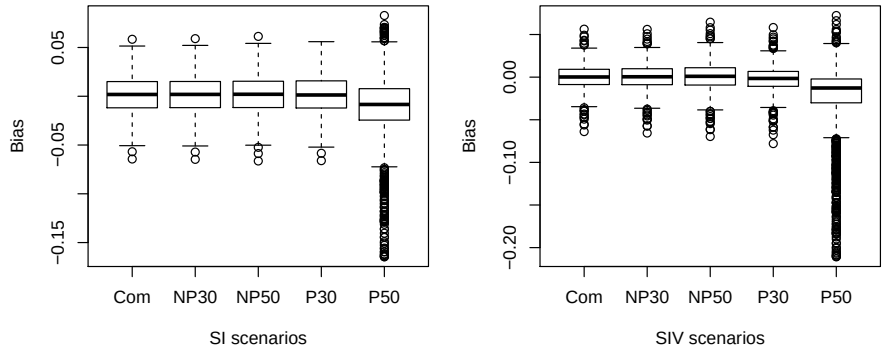

FIGURE 3.9: Box plots of the bias $\widehat{\Delta}_{(x,y)}$ from spatial scenarios I (left) and spatial scenarios IV (right). In each panel, from left to right are complete data, missing 30% without pattern, missing 50% without pattern, missing 30% with pattern and missing 50% with pattern.

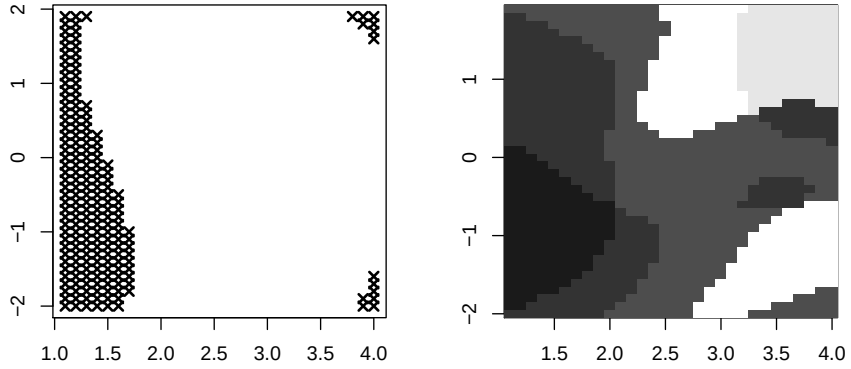

FIGURE 3.10: (Left) A map indicating pixels with large bias of refitted coefficient $\widehat{A}_{(x,y)}$ from the spatial variation VI and missing 50% with spatial pattern scenario. (Right) The missing probability map of the season with the highest missing percentage under this scenario. The horizontal and vertical axes are longitude and latitude respectively.

pattern scenario. The results also suggest that the MM-FPCA be applied with caution if the data set has $\geq 50\%$ observations missing with clear spatial pattern.

3.3 Application to the sparse Lake Victoria data

Application 1: LSWT The MM-FPCA was applied to the sparse Lake Victoria LSWT data. Before applying the model, the grid was trimmed to remove the redundant land pixels. The trimming involves removing 2 pixels to the left, 14 pixels to the right, 5 pixels to the top and 4 pixels to the bottom of the original 65×66 grid. The left panel of Figure 3.11 shows this trimming, where the target area is inside the four red lines. The last image of

April 2012 is not considered in this application as the observations are clearly outliers due to a satellite breakdown. The resulting data set is of dimension $49 \times 57 \times 202$. The missing percentage in the lake pixels is 46.8%, which is lower than the worst case scenario tested in the simulation study. The right panel of Figure 3.11 shows the proportion of data missing in each pixel. The darker areas have more data available and the pale areas display higher missing percentages. The pixels marked with red cross have more than 70% data missing. In general, the situation with respect to the main body of the lake is much less problematic than the boundary of the lake. Considering that the interest of this analysis lies in the lake body, it is appropriate to apply the MM-FPCA in spite of the poorly observed lake borders (refer to explanations in section 3.2.3).

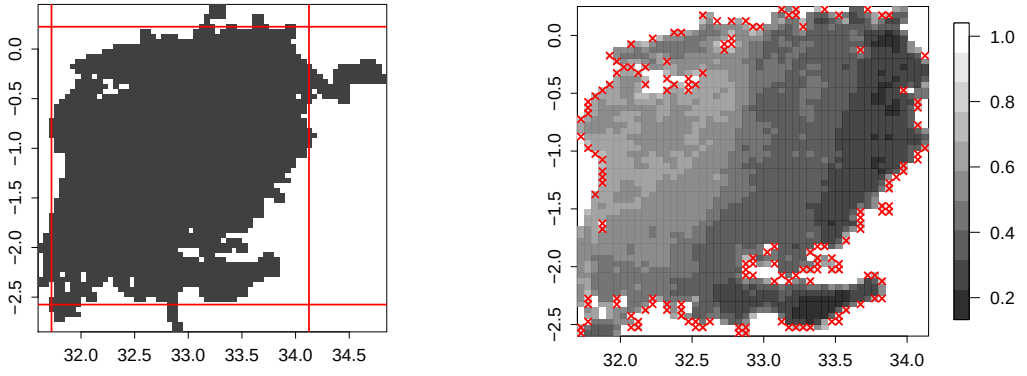

FIGURE 3.11: (Left) The trimming of the grid of the Lake Victoria LSWT data. (Right) The map of proportion of data available in each pixel in the trimmed Lake Victoria data set. The red corsses indicate the pixels with $\leq 30\%$ data available. The horizontal and vertical axes are longitude and latitude respectively.

The monthly mean was removed from the LSWT data before the analysis. The centered images are assumed to be independent realizations from a random spatial process. This is not the most appropriate assumption, but can be justified since the main temporal structure in the LSWT data, which is the seasonality, has been removed. The influence of the remaining temporal dependence is not supposed to be substantive. A tensor spline basis was used to construct the bivariate functions representing the images. A transformation was applied to the tensor spline basis matrix, giving the orthonormal $\Phi(x, y)$. The selection of K and P was processed using the two-stage approach described in section 3.1.3. The variance proportion criterion was used to choose the expansion order. In this application, the basis selected by the AIC and BIC criteria is of dimension 7×7, i.e. 3 knots each along the longitude and the latitude coordinates. Figure 3.12 shows some details of this selection, where the dip can

be found at index 7 on the vertical axis, corresponding to a 7×7 basis[iv]. The threshold of variance proportion 95% suggests that $P = 4$ is appropriate for this problem.

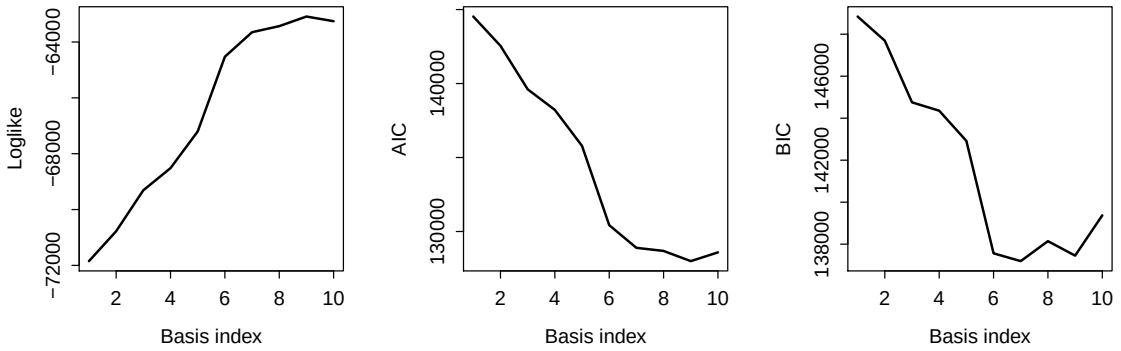

FIGURE 3.12: The selection of basis dimension for the MM-FPCA applied to the sparse Lake Victoria LSWT data. The three panels are the log-likelihood (left), AIC (middle) and BIC (right) against the index of basis of increasing degrees of freedom, from 5×5 to 8×8.

The estimated residual variance of the mixed model is $\hat{\sigma}^2 = 0.1165$. Two principal components with relatively large eigenvalues were identified. They contribute 66.7%, 29.2% each and a sum of 95.9% to the total variations. The plots of the leading two eigenfunctions, along with their scores are given in Figure 3.13. The first eigenfunction displays a contrast between the northeast and the southeast/northwest edge of the lake. The 66.7% variance contribution indicates that this is the dominant spatial pattern in the data over the monitoring period. The second eigenfunction shows a contract between the middle/south of the lake and the north half, plus the southeast corner of the lake. The PC scores are measures of the strength of the corresponding spatial patterns at each time point, which in this case can be interpreted as the evolution of the spatial patterns throughout time. In this example, there is no sign of clear temporal trend or change point in the PC scores.

The reconstructions of LSWT were then produced using the estimated PCs. An example of two imputed LSWT images along with the observed images is given in Figure 3.14. The images were plotted using the same colour scheme for ease of comparison. The reconstructions have captured the main spatial patterns in the data. The RSS from the MM-FPCA reconstructions is compared to those from the reconstructions provided by ARC-Lake, which are derived using EOF-based techniques MacCallum & Merchant (2013). The RSS from the MM-FPCA with four PCs is 0.1207, which is smaller than the RSS value 0.1571 from the

[iv]The indexes of the bases are, 1 for the 5×5 basis, 2 for the 6×5 basis, 3 for the 5×6 basis, 4 for the 6×6 basis, so on and so forth until 10 for the 8×8 basis.

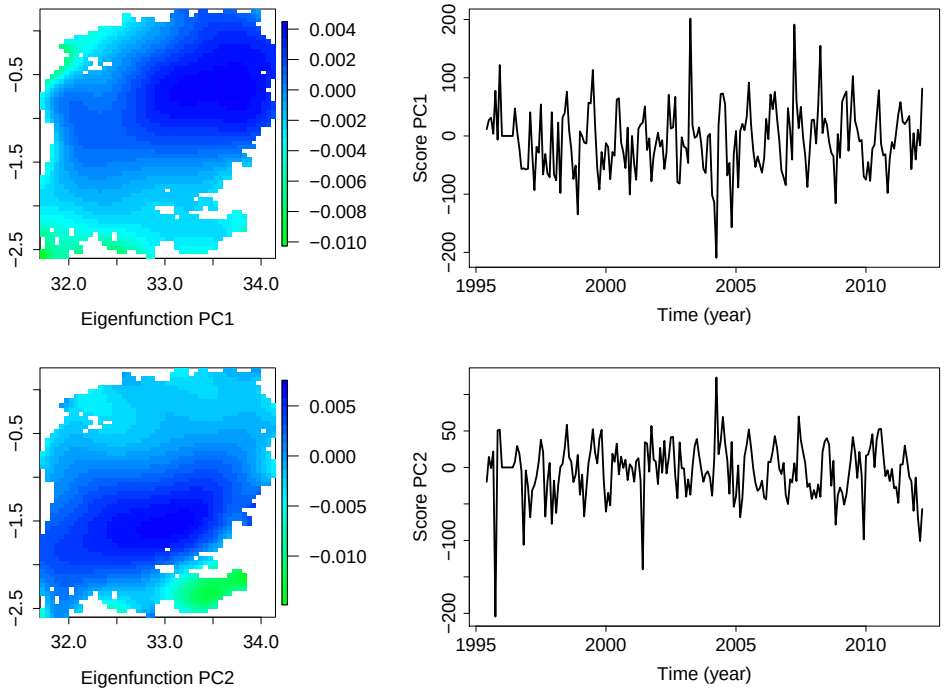

FIGURE 3.13: The plots of the eigenfunctions and the scores of the PC1 (top) and PC2 (bottom). The horizontal and vertical axes of the eigenimages represent longitude and latitude respectively.

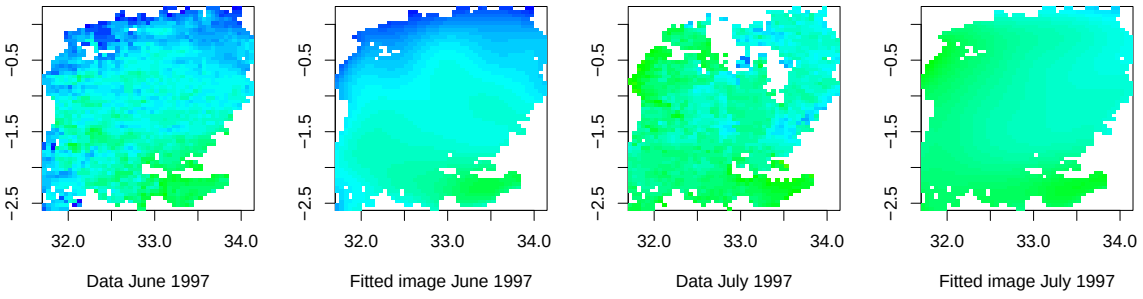

FIGURE 3.14: Examples of the MM-FPCA reconstructions of the Lake Victoria LSWT data. From left to right are the data and the imputation of June 1997 and July 1997. The horizontal and vertical axes are longitude and latitude respectively.

ARC-Lake reconstructions. However, since the EOF analysis is essentially a PCA, the residuals from the ARC-Lake reconstruction can also be made smaller by increasing the number of EOFs involved in the data imputation. A comparison purely based on RSS measures is not always convincing. Therefore, the regional fit of the MM-FPCA is also investigated. In particular, the RSS for individual pixels is computed and the results are shown in Figure 3.15. For the ease of comparison, the two RSS images are plotted using the same color scheme.

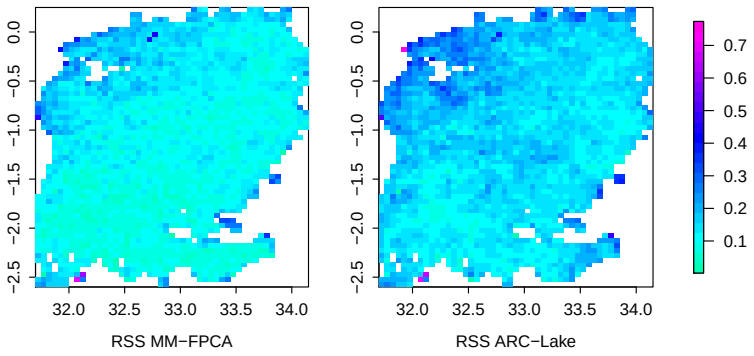

FIGURE 3.15: Images of the RSS of each pixel in the grid from the MM-FPCA (left) and the ARC-Lake reconstructions (right). The legend is for both images. The horizontal and vertical axes of the two images are longitude and latitude respectively.

Although the RSS in pixels with higher missing percentage are generally larger, it appears that the discrepancies between the RSS in these pixels and the better observed pixels are smaller in the MM-FPCA imputation (the left panel) than in the ARC-Lake reconstructions (the right panel). In fact, both the largest and the smallest RSS values come from the ARC-Lake reconstructions. This result shows the potential advantage of the MM-FPCA in terms of data imputation due to improved smoothness as compared to the EOF method. At the same time, information from neighbouring areas and time points can be 'borrowed' in the MM-FPCA to improve the estimation of the less observed area.

Application 2: Chlorophyll For additional information, an application of the MM-FPCA to the Lake Victoria Chlorophyll data is presented here. The data used in this analysis are the spatially aggregated Chl data, i.e. the average value of the Chl observations in a 3×3 grid is used as the observation of the lager pixel covering the 3×3 grid (see the explanation at the end of Chapter 1). Due to the massive size of the data set, only a subset defined on the grid from 32.8°E to 33.4°E, -1.6°N to -1°N is investigated. As the first 7 images were not observed due to a satellite problem, they were excluded from the analysis. This gives a data set of dimension $72 \times 72 \times 112$. The total missing percentage of this data set is 5.3%, which is substantially less than the LSWT data set.

The data was first transformed to the log scale and then centered by a monthly mean. The MM-FPCA using a 7×6 tensor basis and an expansion order of $P = 9$ was fitted. The computation time was 624.89s. The model identified two dominant PCs, explaining 34.62% and 33.42% of the total variation respectively, showing contrasts between the northeast

versus southwest and northwest versus the southeast corner. The rest of the PCs describe the variation patterns at relatively smaller spatial scales, yet are still common to the data. The resulting model has RSS of 0.0539. The reconstructed images captured the general patterns in the original data. However, the fitted model could not capture the occasional discontinuities in an image as a result of an algal bloom. This is highlighted by two panels on the left in Figure 3.16. To some extent, the discontinuities are beyond the capacity of the MM-FPCA as it is designed for modelling smooth data. It might not be a problem if the emphasis is on the general pattern, but would be problematic if the discontinuities are themselves of interest. Under such circumstances, the MM-FPCA might not be an appropriate choice for data that are not smooth by nature. However, if the discontinuity can be accounted for before applying the MM-FPCA, then the results from the MM-FPCA may still be able to provide some useful information.

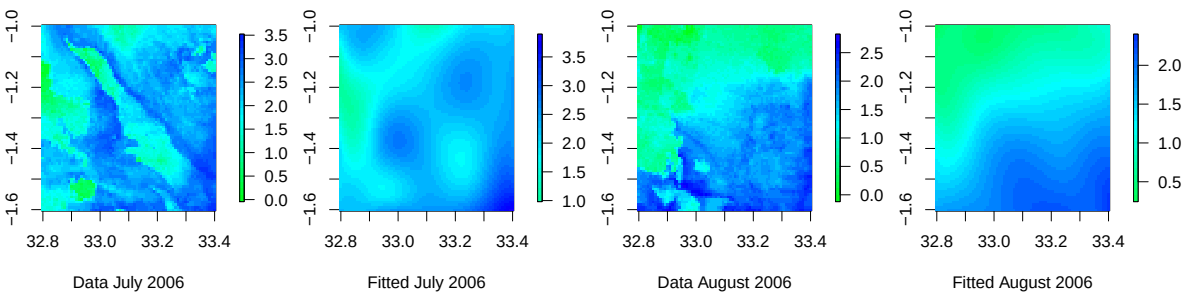

FIGURE 3.16: Examples of the MM-FPCA reconstructions of the Lake Victoria Chlorophyll data. From left to right are the data and the imputations of July 2006 and August 2006. The horizontal and vertical axes of are longitude and latitude respectively.

3.4 Remarks

This chapter presents the mixed model FPCA for the analysis of sparse remote-sensing image time series, when the direct FPCA cannot be implemented due to missing observations. The method treats the principal components as the random effect in a mixed effect model and estimates eigenfunctions and PC scores using maximum likelihood within an EM framework. The test using complete reconstructed LSWT data in section 3.2 shows that the results estimated by the MM-FPCA are comparable to the results extracted from the direct FPCA. The simulation study in 3.2.3 suggests that the MM-FPCA is capable of modelling the sparse data, provided there is no substantive missing and no region in the grid with extremely low data availability. The application of the method to the sparse LSWT data of Lake Victoria

shows encouraging results in terms of data imputation and the RSS values. The application to the Chl data suggests that the smoothness of the data is an important assumption for the MM-FPCA model.

The MM-FPCA is still a relatively new method, especially for image data. There are several interesting aspects for further investigation, e.g. the influence of the independent assumption of the individual functions, the irregular boundaries of the images and the level of smoothness of the original data. Among these, the most intriguing problem with respect to the modelling of time series of remote sensing images is the independence assumption. This is one of the fundamental assumptions of the FPCA, or in fact of any PCA. However, temporal correlations are almost unavoidable in time series data, be it point observations or images. Ignoring their influence would affect the statistical inference based on the model results (Zhou & Pan, 2014), such as the standard errors of the estimated model parameters. Investigating the effect of the temporal correlation is crucial to the generalization of the model for other applications. This problem is investigated in the next chapter.

Chapter 4

Towards a spatio-temporal framework

This chapter explores the statistical approaches that are essential to extending the MM-FPCA to temporally correlated data. These include the general spatio-temporal modelling framework, the temporal dynamic structures, the model distributional assumptions and the estimation methods. Following this investigation, a spatio-temporal model consisting of a state space component and a FPCA component is proposed based on the classic literature and the more recent development of spatio-temporal analysis.

4.1 The spatio-temporal modelling framework: DSTM

The dynamic spatio-temporal model (DSTM) framework is investigated initially. The DSTM is a family of widely used models for spatio-temporal data analysis. Cressie & Wikle (2011) describes the essence of this type of models as the '*hierarchical state space framework*'. It is assumed that the true process of interest cannot be observed perfectly, so the first level employs '*a mapping that relates a set of observations to the true process of interest*'. The second level would then specify '*a model for this true (hidden/latent/state) process*', which typically involves some forms of Markovian-dependency. There is usually a third level providing assumptions on the model parameters.

Following Cressie & Wikle (2011), the DSTM can be written schematically in terms of a data model, a process model and a parameter model. At the top level is the data model,

which associates the observed data $Z(\boldsymbol{x};r)$ in spatial domain $\mathcal{D}$ ($\boldsymbol{x}\in\mathcal{D}$) and time domain $\mathcal{T}$ ($r\in\mathcal{T}$) to a latent/hidden process $Y(\boldsymbol{s};t)$,

$$[\{Z(\boldsymbol{x};r):\ \boldsymbol{x}\in\mathcal{D}\,, r\in\mathcal{T}\}\,|\,\{Y(\boldsymbol{s};t):\ \boldsymbol{s}\in\mathcal{B}_s\,, t\in\mathcal{B}_t\}\,,\ \Psi_D]\ .$$

Here $\mathcal{B}_s$, $\mathcal{B}_t$ represents the neighbourhoods of $\boldsymbol{s}$ and t respectively; Ψ_D is the collection of parameters of this mapping. In the middle level is the process model,

$$\left[Y(\boldsymbol{s};t)\ \middle|\ \left\{Y(\boldsymbol{w};t-\tau_1):\ \boldsymbol{w}\in\mathcal{B}_s^{(1)}\right\},\ \cdots,\left\{Y(\boldsymbol{w};t-\tau_q):\ \boldsymbol{w}\in\mathcal{B}_s^{(q)}\right\},\ \Psi_P\right]\ ,$$

where $\tau_1,\cdots,\tau_q$ are the time lags, $\mathcal{B}_s^{(1)},\cdots,\mathcal{B}_s^{(q)}$ represent the neighbourhoods of $\boldsymbol{s}$ at different time lags and Ψ_P is the collection of process model parameters. The process model describes the spatio-temporal dynamic of the hidden process. Finally, the parameter model at the bottom level is,

$$[\Psi_D,\ \Psi_P|\Psi_H]\ ,$$

with Ψ_H representing the collection of 'hyperparameters'. Various types of models can be built based on this framework, through specifications of the data, process and parameter models and the associated hierarchy (Cressie & Wikle, 2011, Wikle & Hooten, 2010).

For the data model written specifically as

$$\boldsymbol{Z}(\boldsymbol{.};t)=\boldsymbol{A}_t\boldsymbol{Y}(\boldsymbol{.};t)+\boldsymbol{\epsilon}(\boldsymbol{.},t)\ ,$$

both linear and non-linear mapping between the observations $\boldsymbol{Z}(\boldsymbol{.};t)$ and the latent process $\boldsymbol{Y}(\boldsymbol{.};t)$ can be considered through the design of $\boldsymbol{A}_t$. It also provide the possibility of dimension reduction (Wikle & Cressie, 1999) through the basis representation of $\boldsymbol{Y}(\boldsymbol{.};t)$ as

$$\boldsymbol{Y}(\boldsymbol{.};t)=\boldsymbol{\Phi}(\boldsymbol{.})\boldsymbol{\beta}_t+\boldsymbol{\omega}(\boldsymbol{.};t)\,.$$

Typical choices of the basis functions $\boldsymbol{\Phi}(\boldsymbol{.})$ are Fourier, empirical orthogonal functions (EOF), wavelet, splines, bi-square, etc. A related approach using the idea of low rank representation can be found in Mardia *et al.* (1998), for the Kriged Kalman filter.

For the process model, a Markov-type dynamic is often used to describe the evolution of the latent process,

$$\boldsymbol{Y}(\boldsymbol{.};t)=\boldsymbol{M}\boldsymbol{Y}(\boldsymbol{.};t-1)+\boldsymbol{u}(\boldsymbol{.};t)\,,$$

where $\boldsymbol{M}$ is the propagator matrix. This has the advantage of avoiding the specification of a joint spatial-temporal covariance structure, which is usually impractical in real life (Cressie & Wikle, 2011). There are various designs of matrix $\boldsymbol{M}$, e.g. spatio-temporal random walk, 'lagged nearest-neighbour' models, vector auto-regressive (VAR) models, PDE/IDE based models and non-linear specifications (Wikle & Hooten, 2010). It is worthwhile pointing out that the estimation of $\boldsymbol{M}$ can be difficult for a high dimensional process, especially if the number of time points T is small. So parameterization is often considered to reduce the estimation complexity.

The specification of the parameter model is usually associated with the hierarchical designs of the DSTM. An example is parameterizing the error covariance matrix in the data or process model level (Xu & Wikle, 2007). Priors may be assigned to the parameters in a Bayesian setting. It is also possible to incorporate random parameters, since the deterministic model might not be able to describe a complex process. However, one needs to be aware of the interpretation and identifiability issues of such settings (Cressie & Wikle, 2011). In general, a sensible design of the parameter model can simplify the evaluation of the model distribution and its computation at the same time.

The estimation of the DSTM model falls into two general categories. Cressie & Wikle (2011) summarised them as empirical hierarchical modelling (EHM) and Bayesian hierarchical modelling (BHM). Both approaches estimate the models using sequential implementation in an iterative manner. In terms of the inference of the model parameters, EHM often adopts an EM-type algorithm; whereas BHM applies Gibbs samplers, MCMC or other sampling techniques to assist the inference. For the update of system states, EHM often uses Kalman filter/smoother in linear Gaussian models. BHM implementation usually involves sampling from the filtering and prediction distributions. A Kalman filter step can be added to the sampling procedure to update the system states and speed up convergence, provided the dependencies between the current states and previous states are relatively strong.

The DSTM has wide application in remote-sensing data, ranging from research on ocean water temperature by Berliner *et al.* (2000), Stroud *et al.* (2001), tropical ocean surface wind by Wikle *et al.* (2001), Wikle & Berliner. (2005), to global CO_2 by Katzfuss & Cressie (2011, 2012), Nguyen *et al.* (2014) and many others. The method shows distinctive advantages in terms of these applications, e.g. its power in dimensional reduction, flexibility in describing the system dynamics and ability to accommodate different spatial resolutions.

This thesis considers one particular type of DSTM, consisting of three levels.

(a) A data model exploits a 'dimension reduction' through basis representation, similar to the one proposed in Wikle & Cressie (1999),

$$Z(\boldsymbol{s};t) = Y(\boldsymbol{s};t) + \epsilon(\boldsymbol{s};t) \ , \tag{4.1}$$

with the latent process $Y(\boldsymbol{s};t)$ specified using basis function representation

$$Y(\boldsymbol{s};t) = \Phi_\beta(\boldsymbol{s})\boldsymbol{\beta}_t + \zeta(\boldsymbol{s};t) \ . \tag{4.2}$$

(b) A process model describing the dynamics through lagged temporal dependence, such as in a vector auto-regressive model

$$\boldsymbol{\beta}_t = \sum_q \boldsymbol{M}_q \boldsymbol{\beta}_{t-\tau_q} + \boldsymbol{u}_t \ . \tag{4.3}$$

(c) A parameter model putting constraints/priors on $\epsilon(\boldsymbol{s};t)$, $\zeta(\boldsymbol{s};t)$, $\boldsymbol{M}_q$ and $\boldsymbol{u}_t$, which completes the hierarchical design and makes the model identifiable.

Note that the component $\zeta(\boldsymbol{s};t)$ in equation (4.2) is introduced to the model to account for the remaining spatial or spatio-temporal variations which cannot be accommodated by the system dynamic component $\Phi_\beta(\boldsymbol{s})\boldsymbol{\beta}_t$. It is sometimes assumed that $\zeta(\boldsymbol{s};t)$ is a random component and only depends on the data at time t. This type of model is often referred to as a spatio-temporal random effect (STRE) model. The STRE model has received great interest in recent years, research on this model can be found in Cressie *et al.* (2010), Kang & Cressie (2010), Katzfuss & Cressie (2011), to name just a few.

It is possible to further decompose $\zeta(\boldsymbol{s};t)$ as

$$\zeta(\boldsymbol{s};t) = \Phi_\eta(\boldsymbol{s})\boldsymbol{\eta}_t + \omega(\boldsymbol{s};t) \ ,$$

where the basis representation $\Phi_\eta(\boldsymbol{s})\boldsymbol{\eta}_t$ is used to transform a high-dimensional process into a low-dimensional one. The choice of basis Φ_η can be different from Φ_β to reflect different spatial contents, such as the macro and micro spatial scales in Wikle *et al.* (2001). Alternatively, using the same basis yields the following,

$$\begin{aligned} Z(\boldsymbol{s};t) &= Y(\boldsymbol{s};t) + \epsilon(\boldsymbol{s};t) \\ &= \Phi(\boldsymbol{s})\boldsymbol{\beta}_t + \Phi(\boldsymbol{s})\boldsymbol{\eta}_t + \epsilon^*(\boldsymbol{s};t) \ , \end{aligned} \tag{4.4}$$

where $\epsilon^*(\boldsymbol{s};t) = \omega(\boldsymbol{s};t) + \epsilon(\boldsymbol{s};t)$.

There are two reasons why this approach is of interest. First of all, it is flexible in design. The model above allows dimension reduction through basis/spectral representation, as described in Wikle & Cressie (1999). Specifically, the state transition equation with respect to the high-dimensional vector, $\boldsymbol{Y}_t = (Y(s_1;t), \cdots, Y(s_n;t))^\top$, can be transformed into a low-dimensional transition equation of $\boldsymbol{\beta}_t$ without loss of information. This means, the functional representation used in Chapter 3 can be carried into this new setting. Meanwhile, the system dynamic $\Phi_\beta(\boldsymbol{s})\boldsymbol{\beta}_t$ can be efficiently estimated using the classical Kalman filter and smoother. All the parameters associated with the system dynamic can be estimated using an EM-type algorithm (Katzfuss & Cressie, 2011).

The DSTM described here is in its most general form. Various models can be built based on this framework through the specification of model components, which provides the possibility of describing many different spatio-temporal contents. Associated with these models are a variety of estimation methods. In the next two sections, several aspects of the DSTM framework are investigated, including its connection with the state space model, its estimation using the Kalman filter/smoother within the EM algorithm and the frequently used model specifications. These are crucial to the development of the spatio-temporal model for the remote-sensing image time series.

4.2 State space model and Kalman filter/smoother

4.2.1 The state space model and its estimation

As the DSTM is closely related to the state space model, the essentials of the state space modellig framework and the Kalman filter/smoother (KF/KS) are introduced first. Without loss of generality, consider a simple state space model,

$$\boldsymbol{Z}_t = \boldsymbol{\Phi}_t\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}) \tag{4.5}$$

$$\boldsymbol{\beta}_t = \boldsymbol{M}\boldsymbol{\beta}_{t-1} + \boldsymbol{u}_t, \qquad \boldsymbol{u}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{H}). \tag{4.6}$$

It consists of an observation equation (4.5) and a state transition equation (4.6), which are equivalent to the data model and the process model described in section 4.1. Note that if the design matrix $\boldsymbol{\Phi}_t$ is taken to be a basis matrix, then the model above becomes a dimension-reduced state space model (Wikle & Cressie, 1999), where the dimension of the

hidden process, $\boldsymbol{Y}_t = \boldsymbol{\Phi}_t\boldsymbol{\beta}_t$, is reduced to the dimension of the basis coefficient vector $\boldsymbol{\beta}_t$. The state space model can be estimated using the EM algorithm, where the E-step computes the expectation of $\boldsymbol{\beta}_t$ conditioned on all the data $\boldsymbol{Z}_{1:T} = \{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_T\}$ through the Kalman filter/smoother and the M-step produces the MLEs of $\boldsymbol{M}$, $\boldsymbol{G}$ and $\boldsymbol{H}$.

Following the derivation in Shumway & Stoffer (2006), denote $\boldsymbol{Z}_{1:t}$ as the collection of data $\{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_t\}$ and assume that $\boldsymbol{\beta}_t|\boldsymbol{Z}_{1:t-1}$ has distribution $\mathcal{N}(\boldsymbol{\beta}_{t|t-1}, \boldsymbol{B}_{t|t-1})$. The mean and variance of $\boldsymbol{Z}_t|\boldsymbol{Z}_{1:t-1}$ can be written as $\boldsymbol{\Phi}_t\boldsymbol{\beta}_{t|t-1}$ and $\boldsymbol{\Phi}_t\boldsymbol{B}_{t|t-1}\boldsymbol{\Phi}_t^\top + \boldsymbol{G}$ respectively. The joint conditional distribution of $\boldsymbol{\beta}_t, \boldsymbol{Z}_t$ given the information up to time point $t-1$ is thus

$$\begin{pmatrix} \boldsymbol{\beta}_t \\ \boldsymbol{Z}_t \end{pmatrix} \Bigg| \boldsymbol{Z}_{1:t-1} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\beta}_{t|t-1} \\ \boldsymbol{\Phi}_t\boldsymbol{\beta}_{t|t-1} \end{pmatrix}, \begin{pmatrix} \boldsymbol{B}_{t|t-1} & \boldsymbol{B}_{t|t-1}\boldsymbol{\Phi}_t^\top \\ \boldsymbol{\Phi}_t\boldsymbol{B}_{t|t-1} & \boldsymbol{\Phi}_t\boldsymbol{B}_{t|t-1}\boldsymbol{\Phi}_t^\top + \boldsymbol{G} \end{pmatrix} \right) . \tag{4.7}$$

From the above multivariate normal distribution, the filtering equations can be obtained by the well-known conditional distribution results as

$$\begin{aligned} \boldsymbol{\beta}_{t|t} &= \boldsymbol{\beta}_{t|t-1} + \boldsymbol{K}_t\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_{t|t-1}\right) \\ \boldsymbol{B}_{t|t} &= \left(\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{\Phi}_t\right)\boldsymbol{B}_{t|t-1} , \end{aligned} \tag{4.8}$$

where $\boldsymbol{K}_t = \boldsymbol{B}_{t|t-1}\boldsymbol{\Phi}_t^\top\left(\boldsymbol{\Phi}_t\boldsymbol{B}_{t|t-1}\boldsymbol{\Phi}_t^\top + \boldsymbol{G}\right)^{-1}$ is the Kalman gain. With the filtering results, the forecasting equations can be obtained as

$$\begin{aligned} \boldsymbol{\beta}_{t|t-1} &= \boldsymbol{M}\boldsymbol{\beta}_{t-1|t-1} \\ \boldsymbol{B}_{t|t-1} &= \boldsymbol{M}\boldsymbol{B}_{t-1|t-1}\boldsymbol{M}^\top + \boldsymbol{H} . \end{aligned} \tag{4.9}$$

The above equations defines the famous Kalman filter algorithm. The Kalman smoother is derived based on the distribution of $\boldsymbol{\beta}_t|\boldsymbol{Z}_{1:T}$ as

$$\begin{aligned} \boldsymbol{\beta}_{t-1|T} &= \boldsymbol{\beta}_{t-1|t-1} + \boldsymbol{J}_{t-1}\left(\boldsymbol{\beta}_{t|T} - \boldsymbol{\beta}_{t|t-1}\right) \\ \boldsymbol{B}_{t-1|T} &= \boldsymbol{B}_{t-1|t-1} + \boldsymbol{J}_{t-1}\left(\boldsymbol{B}_{t|T} - \boldsymbol{B}_{t|t-1}\right)\boldsymbol{J}_{t-1}^\top , \end{aligned} \tag{4.10}$$

where $\boldsymbol{J}_{t-1} = \boldsymbol{B}_{t-1|t-1}\boldsymbol{M}^\top\boldsymbol{B}_{t|t-1}^{-1}$. See Ansley & Kohn (1982) for more details. The filtering and forecasting algorithm is a forward process (from 1 to T); whereas the smoothing algorithm is a backward process (from T to 1). In Durbin & Koopman (2001), the smoothed versions of $\boldsymbol{u}_t$ and $\boldsymbol{\epsilon}_t$ are also derived to compute the expectation of the log-likelihood. This is not considered in this thesis as it requires a more complicated version of derivation; whereas the same expected log-likelihood can be computed using the smoothed $\boldsymbol{\beta}_t$.

The parameters of the state space model (4.5) and (4.6), $\boldsymbol{M}$, $\boldsymbol{G}$ and $\boldsymbol{H}$, are often estimated using the EM-algorithm. This is out of the concern for the difficulty of evaluating the observed log-likelihood $\mathcal{L}(\boldsymbol{Z}_{1:T}; \Psi)$. To be specific, the derivation of $\mathcal{L}(\boldsymbol{Z}_{1:T}; \Psi)$ requires the conditional distribution of $\boldsymbol{Z}_t|\boldsymbol{Z}_{1:(t-1)}$ because of the dependence $\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}$. The derivation itself might not be difficult, but evaluating the inverse and determinant of the conditional variance of $\boldsymbol{Z}_t$, which are crucial to $\mathcal{L}(\boldsymbol{Z}_{1:T}; \Psi)$, can be computationally infeasible for high-dimensional data. As a result, the complete data log-likelihood $\mathcal{L}(\boldsymbol{Z}_{1:T}, \boldsymbol{\beta}_{0:T}; \Psi)$ is used instead (Cressie & Wikle, 2011, Shumway & Stoffer, 2006). This function inherits the advantage of the hierarchical model, resulting in a much simpler expression.

Denote the distribution of $\boldsymbol{Z}_t$ given $\boldsymbol{\beta}_t$ as $f(\boldsymbol{Z}_t|\boldsymbol{\beta}_t)$, $\boldsymbol{\beta}_t$ given $\boldsymbol{\beta}_{t-1}$ as $f(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1})$ and that of the initial state $\boldsymbol{\beta}_0$ as $f(\boldsymbol{\beta}_0)$. The joint distribution of $\{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_T; \boldsymbol{\beta}_0, \cdots, \boldsymbol{\beta}_T\}$ can be written as

$$f(\boldsymbol{Z}_{1:T}, \boldsymbol{\beta}_{0:T}) = f(\boldsymbol{\beta}_0) \prod_{t=1}^{T} f(\boldsymbol{Z}_t|\boldsymbol{\beta}_t) f(\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}) \,. \tag{4.11}$$

This gives the expectation of the complete data log-likelihood in the E-step,

$$\begin{aligned}
&\mathbf{E}\left[-2\mathcal{L}\left(\Psi; \boldsymbol{Z}_{1:T}, \boldsymbol{\beta}_{0:T}\right) \middle| \boldsymbol{Z}_{1:T}, \Psi^{(it-1)}\right] \\
&= T\log(|\boldsymbol{G}|) + \sum_{t=1}^{T}\left\{\mathbf{E}\left[\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_t\right)^\top \boldsymbol{G}^{-1}\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_t\right) \middle| \boldsymbol{Z}_{1:T}, \Psi^{(it-1)}\right]\right\} \\
&\quad + T\log(|\boldsymbol{H}|) + \sum_{t=1}^{T}\left\{\mathbf{E}\left[\left(\boldsymbol{\beta}_t - \boldsymbol{M}\boldsymbol{\beta}_{t-1}\right)^\top \boldsymbol{H}^{-1}\left(\boldsymbol{\beta}_t - \boldsymbol{M}\boldsymbol{\beta}_{t-1}\right) \middle| \boldsymbol{Z}_{1:T}, \Psi^{(it-1)}\right]\right\} \\
&\quad + \log(|\boldsymbol{B}_0|) + \mathbf{E}\left[\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right)^\top \boldsymbol{B}_0^{-1}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right) \middle| \boldsymbol{Z}_{1:T}, \Psi^{(it-1)}\right] + \text{constant} \,,
\end{aligned} \tag{4.12}$$

where $\Psi^{(it-1)} = \{\boldsymbol{M}^{(it-1)}, \boldsymbol{H}^{(it-1)}, \boldsymbol{G}^{(it-1)}\}$ is the collection of parameter estimates at the current iteration. Note that $\boldsymbol{\beta}$ and $\boldsymbol{B}_0$, which characterize the distribution $\boldsymbol{\beta}_0 \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{B}_0)$, are not considered as model parameters. Replacing the expectations in (4.12) by the smoothed states $\{\boldsymbol{\beta}_{t|T}\}_{t=1}^T$ and the variances $\{\boldsymbol{B}_{t|T}\}_{t=1}^T$ based on $\Psi^{(it-1)}$, the computational form of the E-step equation is derived in Cressie & Wikle (2011) and Shumway & Stoffer (2006) as

$$\begin{aligned}
&\mathbf{E}\left[-2\mathcal{L}\left(\Psi; \boldsymbol{Z}_{1:T}, \boldsymbol{\beta}_{0:T}\right) \middle| \boldsymbol{Z}_{1:T}, \Psi^{(it-1)}\right] \\
&= T\log(|\boldsymbol{G}|) + \sum_{t=1}^{T} \operatorname{tr}\left\{\boldsymbol{G}^{-1}\left[\boldsymbol{\Phi}_t\boldsymbol{B}_{t|T}\boldsymbol{\Phi}_t^\top + \left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_{t|T}\right)\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_{t|T}\right)^\top\right]\right\} \\
&\quad + T\log(|\boldsymbol{H}|) + \operatorname{tr}\left\{\boldsymbol{H}^{-1}\left[\boldsymbol{V}_{11} - \boldsymbol{V}_{10}\boldsymbol{M}^\top - \boldsymbol{M}\boldsymbol{V}_{10}^\top + \boldsymbol{M}\boldsymbol{V}_{00}\boldsymbol{M}^\top\right]\right\} \\
&\quad + \log(|\boldsymbol{B}_0|) + \operatorname{tr}\left\{\boldsymbol{B}_0^{-1}\left[\boldsymbol{B}_{0|T} + \left(\boldsymbol{\beta}_{0|T} - \boldsymbol{\beta}\right)\left(\boldsymbol{\beta}_{0|T} - \boldsymbol{\beta}\right)^\top\right]\right\} + \text{constant} \,,
\end{aligned} \tag{4.13}$$

where

$$\begin{aligned}
\boldsymbol{V}_{11} &= \sum_{t=1}^{T} \left(\boldsymbol{B}_{t|T} + \boldsymbol{\beta}_{t|T} \boldsymbol{\beta}_{t|T}^{\top} \right) \\
\boldsymbol{V}_{00} &= \sum_{t=1}^{T} \left(\boldsymbol{B}_{t-1|T} + \boldsymbol{\beta}_{t-1|T} \boldsymbol{\beta}_{t-1|T}^{\top} \right) \\
\boldsymbol{V}_{10} &= \sum_{t=1}^{T} \left(\boldsymbol{B}_{t,t-1|T} + \boldsymbol{\beta}_{t|T} \boldsymbol{\beta}_{t-1|T}^{\top} \right)
\end{aligned}$$

A sequence of the lag-1 covariance smoother defined as

$$\boldsymbol{B}_{t,t-1|T} = \mathbf{E}\left[(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|T})(\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_{t-1|T}) \right] , \tag{4.14}$$

is required here. It is not part of the KF/KS routine, but can be computed using the output of KF/KS through a backward recursion starting with the last time point T (Shumway & Stoffer, 2006), which is

$$\begin{aligned}
\boldsymbol{B}_{T,T-1|T} &= (\boldsymbol{I} - \boldsymbol{K}_T \boldsymbol{\Phi}_T) \, \boldsymbol{M} \boldsymbol{B}_{T-1|T-1} \\
\boldsymbol{B}_{t-1,t-2|T} &= \boldsymbol{B}_{t-1|t-1} \boldsymbol{J}_{t-2}^{\top} + \boldsymbol{J}_{t-1} \left(\boldsymbol{B}_{t,t-1|T} - \boldsymbol{M} \boldsymbol{B}_{t-1|t-1} \right) \boldsymbol{J}_{t-2}^{\top} .
\end{aligned}$$

In standard cases where no special parameterization is involved, all three parameters have analytical solutions for their MLEs, giving the M-step equations as

$$\widehat{\boldsymbol{M}}^{(it)} = \boldsymbol{V}_{10} \boldsymbol{V}_{00}^{-1} \tag{4.15}$$

$$\widehat{\boldsymbol{H}}^{(it)} = \frac{1}{T} \left(\boldsymbol{V}_{11} - \boldsymbol{V}_{10} \boldsymbol{V}_{00}^{-1} \boldsymbol{V}_{10}^{\top} \right) \tag{4.16}$$

$$\widehat{\boldsymbol{G}}^{(it)} = \frac{1}{T} \sum_{t=1}^{T} \left[\boldsymbol{\Phi}_t \boldsymbol{B}_{t|T} \boldsymbol{\Phi}_t^{\top} + \left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t \boldsymbol{\beta}_{t|T} \right) \left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t \boldsymbol{\beta}_{t|T} \right)^{\top} \right] . \tag{4.17}$$

The algorithm then iterates until the log-likelihood and/or the parameter estimates converge.

4.2.2 Computational challenges of the Kalman filter

Despite being a classic approach that has been used for over 60 years since Kalman (1960), there are some computational challenges when the Kalman filter is applied to the sparse remote-sensing image time series. Two of these challenges with respect to the high dimensionality and the sparsity are discussed here.

(a) In terms of the filtering of high-dimensional data, although a dimension-reduced form can be used (Wikle & Cressie, 1999), the computation of some elements of the Kalman filter and the log-likelihood in the corresponding EM algorithm may still be difficult. To ease the computation burden, the Sherman-Morrison-Woodbury identity is used to simplify the matrix inversion as part of the Kalman gain,

$$\left(\boldsymbol{\Phi}_t \boldsymbol{B}_{t|t-1} \boldsymbol{\Phi}_t^\top + \boldsymbol{G}\right)^{-1} = \boldsymbol{G}^{-1} - \boldsymbol{G}^{-1}\boldsymbol{\Phi}_t \left(\boldsymbol{B}_{t|t-1}^{-1} + \boldsymbol{\Phi}_t^\top \boldsymbol{G}^{-1} \boldsymbol{\Phi}_t\right)^{-1} \boldsymbol{\Phi}_t^\top \boldsymbol{G}^{-1}. \tag{4.18}$$

The advantage of this identity lies in the dimension reduction of the matrix to be inverted. Provided that matrix $\boldsymbol{\Phi}_t \boldsymbol{B}_{t|t-1} \boldsymbol{\Phi}_t^\top + \boldsymbol{G}$ is of much higher dimension (i.e. the dimension of the data) than the matrix $\boldsymbol{B}_{t|t-1}^{-1} + \boldsymbol{\Phi}_t^\top \boldsymbol{G}^{-1} \boldsymbol{\Phi}_t$ (i.e. the dimension of the basis matrix), this could result in a substantial improvement in computational efficiency. In terms of the remote-sensing image data, this means a reduction from, for example, 2000×2000 to 25×25. The computation of the conditional expectation of the log-likelihood (4.12) can also be made easier through some matrix algebra results. For example, the Cholesky decomposition of matrix $\boldsymbol{G}$ and $\boldsymbol{H}$ may be used to improve the stability of the computation of the matrix inverse and the logarithm of determinant. Take matrix $\boldsymbol{G}$, this is

$$\begin{aligned} \boldsymbol{G}^{-1} &= \boldsymbol{G}_c^{-1} \left(\boldsymbol{G}_c^{-1}\right)^\top \\ \log\left(|\boldsymbol{G}|\right) &= \log\left(\prod_i g_i^2\right) = 2\sum_i \log\left(g_i\right) \ , \end{aligned}$$

where $\boldsymbol{G}_c$ is the Cholesky factor of matrix $\boldsymbol{G}$ and g_i is the i-th diagonal element of $\boldsymbol{G}_c$. This treatment would also reduce the risk of getting a singular matrix as a result of system rounding errors.

(b) When it comes to the filtering of sparse images, the missing data Kalman filter, which is described in Shumway & Stoffer (2006), is required. There are two different situations, each adopts different filtering equations. The first one applies to the situation where there is no data observed at time t. The solution to this problem is almost straightforward. According to Petris *et al.* (2009), simply replace the standard filter equations

$$\begin{aligned} \boldsymbol{\beta}_{t|t} &= \boldsymbol{\beta}_{t|t-1} + \boldsymbol{K}_t \left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t \boldsymbol{\beta}_{t|t-1}\right) \quad \text{with} \quad \boldsymbol{\beta}_{t|t} = \boldsymbol{\beta}_{t|t-1} \\ \boldsymbol{B}_{t|t} &= \left(\boldsymbol{I} - \boldsymbol{K}_t \boldsymbol{\Phi}_t\right) \boldsymbol{B}_{t|t-1} \quad \text{with} \quad \boldsymbol{B}_{t|t} = \boldsymbol{B}_{t|t-1} \ , \end{aligned} \tag{4.19}$$

and carry out the prediction step in the same way as

$$\begin{aligned} \boldsymbol{\beta}_{t+1|t} &= \boldsymbol{M}\boldsymbol{\beta}_{t|t} = \boldsymbol{M}\boldsymbol{\beta}_{t|t-1} \\ \boldsymbol{B}_{t+1|t} &= \boldsymbol{M}\boldsymbol{B}_{t|t}\boldsymbol{M}^\top = \boldsymbol{M}\boldsymbol{B}_{t|t-1}\boldsymbol{M}^\top . \end{aligned} \tag{4.20}$$

The second one applies to the situation where the data are partly observed at time t. According to Shumway & Stoffer (2006), the adjustment here is to filter with only the observed part of $\boldsymbol{Z}_t$ and change the likelihood accordingly with a prior distribution on the missing data. This means, the last line in equation (4.12) becomes

$$\sum_{t=1}^{T} \left\{ \mathbf{E}\left[\left(\boldsymbol{Z}_t^{(o)} - \boldsymbol{\Phi}_t^{(o)}\boldsymbol{\beta}_t \right)^\top \left(\boldsymbol{G}^{(o)} \right)^{-1} \left(\boldsymbol{Z}_t^{(o)} - \boldsymbol{\Phi}_t^{(o)}\boldsymbol{\beta}_t \right) \middle| \; \boldsymbol{Z}_{1:T}^{(o)}, \Psi^{(it-1)} \right] \right\}, \tag{4.21}$$

where $\boldsymbol{Z}_t^{(o)}$ and $\boldsymbol{\Phi}_t^{(o)}$ are the reordered $\boldsymbol{Z}_t$ and $\boldsymbol{\Phi}_t$ whose leading rows are observed and the rest are missing. The covariance matrix $\boldsymbol{G}^{(o)}$ is often assumed to be block diagonal, $\text{diag}\{\boldsymbol{G}_{obs}, \boldsymbol{G}_{mis}\}$, indicating that the missing elements and the non-missing elements are not correlated (which is consistent with the assumption of missing at random). This gives the following equation as the explicit form of (4.21),

$$\begin{aligned} \sum_{t=1}^{T} \text{tr} \Bigg\{ \left(\boldsymbol{G}^{(o)} \right)^{-1} \Bigg[& \left(\boldsymbol{Z}_t^{(o)} - \boldsymbol{\Phi}_t^{(o)}\boldsymbol{\beta}_{t|T} \right) \left(\boldsymbol{Z}_t^{(o)} - \boldsymbol{\Phi}_t^{(o)}\boldsymbol{\beta}_{t|T} \right)^\top \\ & + \boldsymbol{\Phi}_t^{(o)} \boldsymbol{B}_{t|T} \boldsymbol{\Phi}_t^{(o)\top} + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{G}_{mis} \end{pmatrix} \Bigg] \Bigg\} . \end{aligned} \tag{4.22}$$

Details of the derivation can be found in Shumway & Stoffer (2006). The log-likelihood of the model with partly unobserved $\boldsymbol{Z}_t$ requires a prior distribution on the missing data to initialize the filter. In practice, the mean vector of the unobserved data is often set to zero. As for covariance matrix $\boldsymbol{G}_{mis}$, a simple approach is provided in Shumway & Stoffer (1982), where $\boldsymbol{G}_{mis}$ in the it-th iteration consists of the corresponding rows and columns in the estimated $\boldsymbol{G}^{(it-1)}$ from the $(it-1)$-th iteration. This algorithm is of great importance to the problem in this thesis because the remote sensing images are usually only partly observed. `R` functions for the missing data filter have been developed based on the functions in package `dlm` (Petris, 2010).

The Kalman filter is a powerful tool to estimate the system states. However, there could be a problem of over-fitting when the filter is applied to extremely sparse images. The example highlighting this problem was carried out on a subset of the ARC-Lake reconstructed LSWT

data of Lake Victoria. It is defined on the same grid as the 'LSWT section' data set, which is introduced at the end of Chapter 1, and is of dimension $34 \times 24 \times 202$. To highlight the problem, about 25% of the images were replaced with their sparse counterparts in the 'LSWT section' data set.

The dimension-reduced state space model consisting of equation (4.5) and (4.6) was fitted to this data set. A comparison of the filtered states $\boldsymbol{\beta}_{t|t}$ and fitted values $\widehat{\boldsymbol{Z}}_t$ at $t-1, t, t+1$, where t is the time point corresponding to a sparse image, shows that the data at time t can play an important role in the filtered results due to the Kalman gain. Sometimes, this impact is much larger than the temporal dynamics specified in the state transition equation. This could result in over-fitting if there are only a few observations, possibly in very different scales, scattered across the space. To illustrate this phenomenon, a tensor spline basis, where the basis functions have compact support, was used. If there is no observation in the area covered by the k-th basis function $\phi_k(x, y)$, the corresponding basis coefficient $\beta_{k,t|t}$ (i.e. the k-th element of vector $\boldsymbol{\beta}_{t|t}$) is supposed to be more or less similar to its temporal neighbours, $\beta_{k,t-1|t-1}$ and $\beta_{k,t+1|t+1}$. Whereas for the observed areas, the coefficients are more likely to be governed by the data. Note that since there are overlaps in the compact support of different basis functions, the data in observed areas can affect more than one basis coefficients, so the connection between $\beta_{k,t|t-1}$ and $\beta_{k,t-1|t-1}$, $\beta_{k,t+1|t+1}$ is not always clear. Nonetheless, information can be obtained from this comparison.

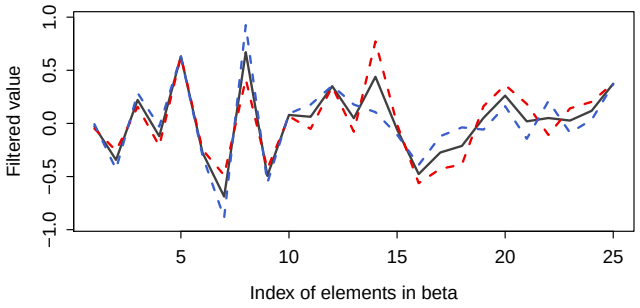

FIGURE 4.1: Example of the filtered states $\boldsymbol{\beta}_{t-1|t-1}$ (red dashed), $\boldsymbol{\beta}_{t|t}$ (black solid) and $\boldsymbol{\beta}_{t+1|t+1}$ (blue dashed), where the data at $t = 10$ are completely missing. The horizontal axis represents the index of the elements in the filtered state vector.

Two examples are presented here. Figure 4.1 shows an extreme case where data at time point $t = 10$ are completely missing. Without a doubt, the filtered $\boldsymbol{\beta}_{t|t}$ (black curve) follows its neighbours $\boldsymbol{\beta}_{t-1|t-1}, \boldsymbol{\beta}_{t+1|t+1}$ (red and blue dashed curves) closely, for the only information available for time point t is $\boldsymbol{Z}_{t-1}$ and $\boldsymbol{Z}_{t+1}$. Situations are different when there are a few

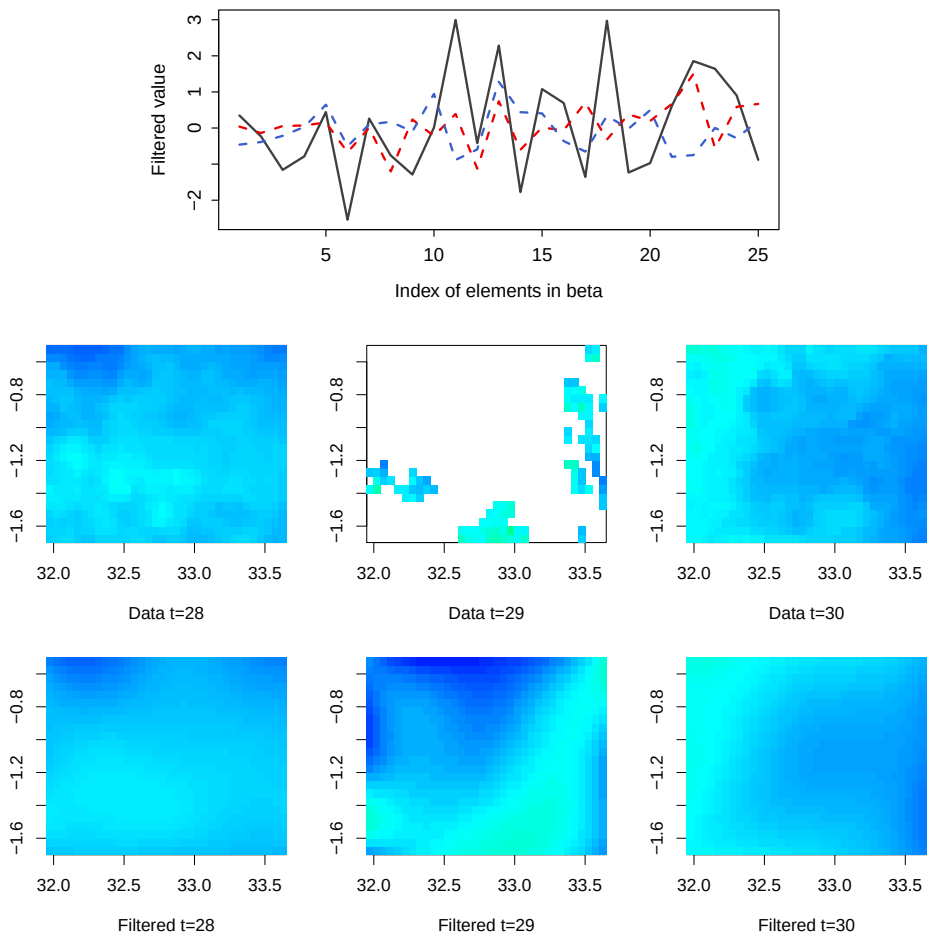

FIGURE 4.2: (Top) Example of the filtered states $\boldsymbol{\beta}_{t-1|t-1}$ (red dashed), $\boldsymbol{\beta}_{t|t}$ (black solid) and $\boldsymbol{\beta}_{t+1|t+1}$ (blue dashed), where the majority of data at $t = 29$ are missing. (Bottom) Images of the data $\boldsymbol{Z}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{Z}_{t+1}$ and their filtered counterparts. The horizontal and vertical axes of the images are longitude and latitude respectively.

observations scattered over space, as illustrated in Figure 4.2. The top panel shows the filtered states $\boldsymbol{\beta}_{t-1|t-1}, \boldsymbol{\beta}_{t|t}$ and $\boldsymbol{\beta}_{t+1|t+1}$, where the majority of the data at $t = 29$ are missing. The middle three panels represent the data $\boldsymbol{Z}_{t-1}, \boldsymbol{Z}_t, \boldsymbol{Z}_{t+1}$ and the bottom three panels show their filtered counterparts. There are clear discrepancies in some parts of the curves, which is most likely induced by the sparse observations. Meanwhile, the filtered $\widehat{\boldsymbol{Z}}_t$ displays a large contrast between the southeast and the northwest, which appears too dramatic considering the high missing percentage and the smoother patterns in its neighbouring images.

The over-fitting of very sparse images could be a problem in data imputation. In remote-sensing, extremely sparse images may be associated with very high uncertainty in data retrievals. Therefore, a wiggly interpolation based on only a few anomalies may be less attractive than a smooth imputation reflecting the more average situation. One solution to

this problem is to introduce a threshold to the filtering algorithm based on the percentage of missing data per image. If the missing percentage surpasses the threshold, omit the filtering step and use local smoothing to estimate $\boldsymbol{\beta}_{t|t}$ and $\boldsymbol{B}_{t|t}$. There are various approaches to local smoothing, such as simple averaging of $\boldsymbol{\beta}_{t-1|t-1}$ and $\boldsymbol{\beta}_{t+1|t+1}$, prediction based on the system transition equation $\boldsymbol{\beta}_{t|t} = \boldsymbol{M}\boldsymbol{\beta}_{t-1|t-1}$ and the n-step backward smoothing. In this thesis, a filtering algorithm with threshold on the missing percentage, combined with the n-step backward smoothing, was proposed. The algorithm consists of the following steps.

(a) Compute the missing percentage of the data at time point t, denoted as $p_t\%$.

(b) If $p_t\%$ is greater than the threshold $\chi\%$, apply the filtering equations in (4.19) as if no observation is available for time point t.

(c) Go to time $t+1$ and repeat steps (a) and (b).

(d) Continue repeating steps (a) to (c) until there is an image with $p_{t^*}\% \leq \chi\%$ and apply the standard filtering equations in (4.8) and then the n-step ($n = t^* - t$) backward smoothing.

(e) Go to the next time point $t^* + 1$ and repeat steps (a) to (d).

4.2.3 Simulation study on the Kalman filter with threshold

An important issue with respect to the above algorithm is the selection of the threshold and its effect. A simulation study was carried out to investigate this problem.

Part 1: simulation design

- Data are simulated on a regular 30×30 grid, where images are recorded at 120 time points to mimic a 10-year observing period.

- The same data generating function (3.18) as in section 3.2.3 is used. However, this simulation study only considers one spatial variation scenario for the Gaussian random field (GRF), which has $d = 1$. Three noise levels are considered, controlled by the nugget effect parameter σ^2_{ng} of the GRF,

$$\sigma^2_{ng} = \{0.01, 0.05, 0.1\}\,.$$

- Only one missing condition is considered: the high missing percentage with spatial patterns scenario. A new design of extremely sparse images is added to the missing pattern design in section 3.2.3. To create the extremely sparse images, 12 out of 120 simulated images are selected. These images are then divided into two categories, 6 images with observations gathered in a small area and 6 images with observations scattered over space. The missing percentage of the images with gathered observations is set to be 99%; the missing percentage of the images with scattered observations is 92%. These two missing percentages are used to create images with high levels of missingness, so that the high filtering thresholds (i.e. 95% and 100%) can take effect on these images. With these extremely sparse images, the total missing percentage in the simulated data set is about 48%.

- The filtering thresholds investigated in this simulation study are

$$\chi\% = \{70\%, 80\%, 90\%, 95\%, 100\%\} \ ,$$

  where $\chi\% = 100\%$ means filtering every single image in the data set.

- This gives 15 simulation scenarios in total.

The following statistics are recorded in this simulation study.

(a) The filtered $\{\boldsymbol{\beta}_{t|t}\}$ and the smoothed $\{\boldsymbol{\beta}_{t|T}\}$ from the KF/KS with five different thresholds, the estimated $\widehat{\boldsymbol{G}}_t$ (diagonal elements only) and $\widehat{\boldsymbol{H}}_t$ matrices.

(b) The RSS of both the sparse observations (denoted as RSS1) and the complete observations (denoted as RSS2), for the entire data set and the images with more than 70%, 80% and 90% of data missing. This helps to assess the influence of the thresholding on the spatial interpolation.

For each scenario, 200 replicates are run. To speed up the computation, the EM algorithm is only run for the $\chi\% = 100\%$ threshold for each replicate. The converged estimations of $\widehat{\boldsymbol{G}}_t$ and $\widehat{\boldsymbol{H}}_t$ are carried into the Kalman filtering with thresholds $\chi\% = \{70\%, 80\%, 90\%, 95\%\}$. To initialize the EM algorithm, the variance of the data σ_z^2 is used as the starting point. For the first replicate, set $\boldsymbol{G}_t^{(0)} = \sigma_z^2 I$ and $\boldsymbol{H}_t^{(0)} = \kappa\sigma_z^2 I$, where κ is a scaling factor. For the remaining 199 replicates, the converged results from the first replicate, $\widehat{\boldsymbol{G}}_t^{1*}$ and $\widehat{\boldsymbol{H}}_t^{1*}$, are used to initialize the parameters as

$$\boldsymbol{G}_t^{(0)} = \sigma_g^2 \boldsymbol{I}, \quad \sigma_g^2 = \frac{1}{n}\sum_{i=1}^{n} \text{diag}\{\widehat{\boldsymbol{G}}_t^{1*}\} \ ,$$
$$\boldsymbol{H}_t^{(0)} = \sigma_h^2 \boldsymbol{I}, \quad \sigma_h^2 = \frac{1}{K}\sum_{k=1}^{K} \text{diag}\{\widehat{\boldsymbol{H}}_t^{1*}\} \ .$$

Part 2: simulation results To illustrate some interesting features of the Kalman filter with a threshold, RSS1 and RSS2 with respect to two different groups of images are compared, one consisting of all 120 simulated images and the other including 42 images which have more than 70% data missing. Tables 4.1 and 4.2 present the mean and the 95% confidence intervals (obtained using the quantiles of the simulation estimates) of RSS1 and RSS2 from different filters, based on 120 and 42 images. In general, it is better to filter the image than to leave them to the smoother. This is reflected by the decreasing RSS1 and RSS2 values with the increasing thresholds in all scenarios. The decrease in RSS1 and RSS2 values from a 70% threshold to a 80% threshold is distinctive, but the decrease from a 95% threshold to a 100% threshold is almost negligible. The changes are much more distinctive in the results computed using 42 images. These decreasing patterns can be seen clearer in Figure 4.3, which shows the boxplots of the RSS1 and RSS2 from different filters, based on all 120 images. These results suggest that the relative changes of RSS, computed as $\text{RSS}_{\chi_1\%}/\text{RSS}_{\chi_2\%} - 1$, can be used as a criterion to select the filtering threshold. For example, if the relative changes in RSS1 values based on all the images are assessed, then a criterion of $\leq 5\%$ would suggest $\chi\% = 95\%$ for the small noise scenario, $\chi\% = 90\%$ for the medium noise scenario and $\chi\% = 80\%$ for the large noise scenario.

In addition, there is also evidence that, sometimes filtering without thresholding produces larger RSS2 values. This is illustrated by the plots of the differences between the RSS2 based on all 120 images from the filter with the 95% threshold and the filter without threshold ($\text{RSS2}_{95\%} - \text{RSS2}_{100\%}$) in Figure 4.4. The differences from 200 replicates under the small, medium and large noise scenarios are displayed in three panels. The plots also show that the occurrence of the negative differences becomes higher as the noise level increases. In this case, occurrences in three different scenarios are 22, 53 and 67 (out of 200) respectively.

In conclusion, this simulation study reveals some features of the Kalman filter with an additional threshold based on missing percentages. The results suggest that it is usually better to use a relatively high threshold than a low threshold, e.g. 90% versus 70%. However, if there is concern of over-fitting due to high uncertainties in the observations, then it would

TABLE 4.1: The mean and 95% confidence interval of RSS1 computed using all 120 images and 42 images (with more than 70% observations missing), from the small, medium and large noise scenarios, with Kalman filters of increasing filtering threshold χ%.

$\chi\% = 70\%$	$\chi\% = 80\%$	$\chi\% = 90\%$	$\chi\% = 95\%$	$\chi\% = 100\%$
120 images				
		Small		
0.0487	0.0208	0.0162	0.0139	0.0136
(0.0406, 0.0577)	(0.0182, 0.0242)	(0.0150, 0.0179)	(0.0133, 0.0145)	(0.0131, 0.0141)
		Medium		
0.0882	0.0598	0.0552	0.0528	0.0525
(0.0807, 0.0976)	(0.0571, 0.0642)	(0.0538, 0.0570)	(0.0519, 0.0537)	(0.0515, 0.0533)
		Large		
0.1369	0.1086	0.1041	0.1018	0.1015
(0.1301, 0.1451)	(0.1055, 0.1116)	(0.1023, 0.1060)	(0.1003, 0.1032)	(0.1000, 0.1029)
42 images ($\geq$ 70% miss)				
		Small		
0.3142	0.0746	0.0349	0.0148	0.0120
(0.2452, 0.3910)	(0.0505, 0.1031)	(0.0251, 0.0491)	(0.0131, 0.0178)	(0.0113, 0.0128)
		Medium		
0.3575	0.1134	0.0731	0.0530	0.0503
(0.2941, 0.4382)	(0.0905, 0.1508)	(0.0633, 0.0864)	(0.0505, 0.0559)	(0.0484, 0.0523)
		Large		
0.4040	0.1596	0.1209	0.1013	0.0986
(0.3431, 0.4699)	(0.1374, 0.1831)	(0.1113, 0.1322)	(0.096y, 0.1054)	(0.0943, 0.1022)

TABLE 4.2: The mean and 95% confidence interval of RSS2 computed using all 120 images and 42 images (with more than 70% observations missing), from the small, medium and large noise scenarios, with Kalman filters of increasing filtering threshold χ%.

$\chi\% = 70\%$	$\chi\% = 80\%$	$\chi\% = 90\%$	$\chi\% = 95\%$	$\chi\% = 100\%$
120 images				
		Small		
0.1241	0.0600	0.0485	0.0387	0.0356
(0.1051, 0.1441)	(0.0506, 0.0692)	(0.0403, 0.0573)	(0.0328, 0.0460)	(0.0305, 0.0423)
		Medium		
0.1659	0.1022	0.0913	0.0819	0.0803
(0.1483, 0.1866)	(0.0924, 0.1151)	(0.0833, 0.1013)	(0.0760, 0.0891)	(0.0742, 0.0879)
		Large		
0.2149	0.1536	0.1433	0.1349	0.1340
(0.1987, 0.2322)	(0.1454, 0.1632)	(0.1365, 0.1514)	(0.1296, 0.1430)	(0.1276, 0.1418)
42 images ($\geq$ 70% miss)				
		Small		
0.3232	0.1400	0.1073	0.0793	0.0705
(0.269, 0.3806)	(0.1131, 0.1654)	(0.0851, 0.1307)	(0.0626, 0.1007)	(0.0563, 0.0889)
		Medium		
0.3665	0.1844	0.1533	0.1267	0.1219
(0.3166, 0.4247)	(0.1568, 0.2204)	(0.1310, 0.1825)	(0.1090, 0.1466)	(0.1051, 0.1443)
		Large		
0.4113	0.2361	0.2066	0.1827	0.1801
(0.3653, 0.4630)	(0.2136, 0.2642)	(0.1873, 0.2285)	(0.1672, 0.2057)	(0.1626, 0.2017)

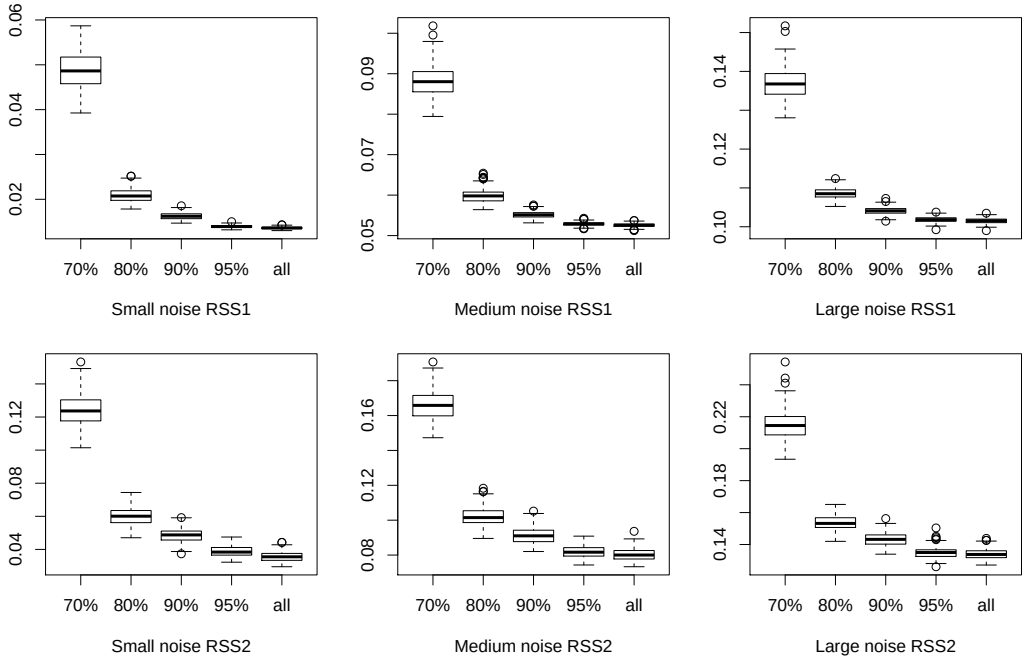

FIGURE 4.3: The boxplot of RSS1 (top) and RSS2 (bottom) based on all 120 images using Kalman filters with the increasing thresholds. In each row, from left to right are plots from the small, medium and large noise scenarios. The horizontal axis represents the threshold $\chi\% = 70\%, 80\%, 90\%, 95\%$ and 100%.

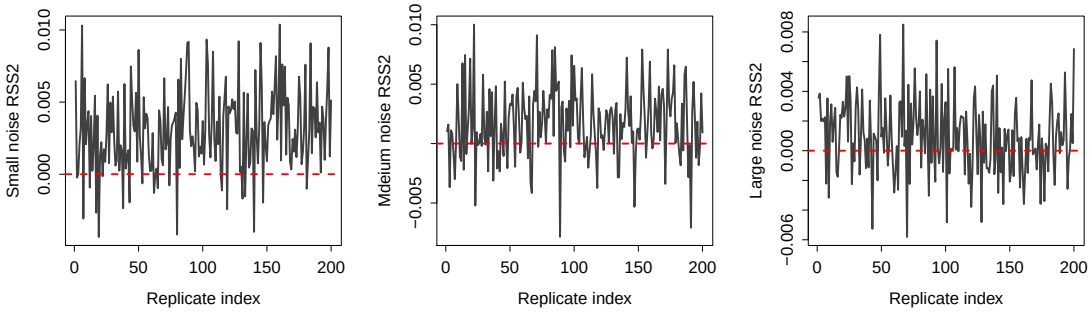

FIGURE 4.4: The difference in RSS2 (based on 120 images) between the filter with a 95% and a 100% threshold. The three panels show results from the small (left), medium (middle) and large (right) noise scenarios. The dashed horizontal line indicates $\text{RSS}_{95\%} - \text{RSS}_{100\%} = 0$.

be beneficial to avoid filtering the extremely sparse data at certain time points. In addition, the selection of the threshold can be made using the relative change in the RSS values from filters with different thresholds [i].

[i] During the PhD viva, the examiners suggested that, as the best fit usually comes from the filtering with a 100% threshold, it might be helpful to introduce certain types of penalty into the selection criterion of the threshold. The penalty could be based on the desired smoothness of the very sparse images.

4.3 Spatio-temporal model development

4.3.1 Preliminaries on parameterization & estimation

The above sections have laid the foundations of the development of the spatio-temporal model. In the following paragraphs, two more specific aspects in terms of the DSTM framework are presented, including the widely-used approaches to parameterize the data/process model and a specific algorithm used to estimate a spatio-temporal random effect (STRE) model. These two aspects provide details that are directly linked to the proposal of the spatio-temporal model for sparse remote-sensing image time series.

Preliminary 1: parameterizing the data/process models Spatio-temporal processes are usually of high dimensionality, probably also with missing observations across space and time. As a result, estimation of model components can be problematic, especially if the model parameters are unconstrained. A helpful solution to this problem is to parameterize the model components based on '*prior scientific knowledge and/or common spatial models*' (Xu & Wikle, 2007). Some frequently used parameterizations of the data model and process model covariance matrices are summarised in Xu & Wikle (2007).

(a) Assume the residuals of the data model are i.i.d. random noises, i.e. $\boldsymbol{G} = \sigma_\epsilon^2 \boldsymbol{I}$.

(b) Use an empirical orthogonal function (EOF) expansion to parameterize the data model residual covariance matrix, i.e. $\boldsymbol{G} = \sigma^2 \boldsymbol{I} + \sum_{p=a+1}^{P} \lambda_p \boldsymbol{\xi}_p \boldsymbol{\xi}_p^\top$, where $\boldsymbol{\xi}_p$ are the EOFs and λ_p are corresponding eigenvalues.

(c) Specify an exponential covariance function for the process model covariance matrix, i.e. $\boldsymbol{H} = \sigma^2 \boldsymbol{V}(h; d)$, where the elements in matrix $\boldsymbol{V}(h; d)$ is determined by correlation function $\rho(h) = \exp(-h/d)$ with d being the parameter.

(d) Use a conditional auto-regressive (CAR) model for the residuals of the process model, i.e. $u(s_i)|u(s_j) \sim \mathcal{N}\left(b \sum_{j \neq i} c_{ij} u(s_j),\ \sigma_i^2\right)$, where b is the CAR model parameter, c_{ij} describes the adjacency of $u(s_i)$ and $u(s_j)$.

Note that the last parameterization only applies to a process model describing the evolution of the actual spatial process. It does not apply to the dimension-reduced state space model as defined in equation (4.1) and (4.3).

Among the four parameterizations above, the EOF expansion is the most interesting for the aims of this thesis. The concept of EOF in atmospheric and meteorological science is similar to the concept of principal component in statistics. Therefore, parameterizing the residual covariance matrix $\boldsymbol{G}$ using EOFs is associated with the PCA based on the covariance matrix $\boldsymbol{G}$. Furthermore, if a basis representation is used for the data $\boldsymbol{Z}_t$ and the EOFs, then this parameterization could be linked to a FPCA in that

$$\begin{aligned} \boldsymbol{Z}_t &= \sum_{p=1}^{a} \boldsymbol{\xi}_p \alpha_{pt} + \boldsymbol{\epsilon}_t \\ \boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G}) \,, \quad \boldsymbol{G} &= \sigma^2 \boldsymbol{I} + \sum_{p=a+1}^{P} \lambda_p \boldsymbol{\xi}_p \boldsymbol{\xi}_p^\top \,, \end{aligned} \tag{4.23}$$

where λ_p is the eigenvalue and $\boldsymbol{\xi}_p$ is the vector of the evaluated eigenfunction $\xi_p(\centerdot)$. Both of them can be extracted from the FPCA. Unfortunately, the MLE of σ^2 has no analytical solution, so numerical methods are required for the estimation. Specifically, it is done by numerically equating the score function of σ^2 to zero. (Xu & Wikle, 2007) derived the score function,

$$\begin{aligned} F(\sigma^2) = \frac{Tn}{\sigma^2} + T \sum_{p=a+1}^{P} \left(\frac{1}{\sigma^2 + \lambda_p} - \frac{1}{\sigma^2} \right) \mathrm{tr} \left\{ \boldsymbol{\xi}_p \boldsymbol{\xi}_p^\top \right\} \\ - \frac{\mathrm{tr}\{\boldsymbol{A}\}}{\sigma^4} - \sum_{p=a+1}^{P} \left[\frac{1}{(\sigma^2 + \lambda_p)^2} - \frac{1}{\sigma^4} \right] \mathrm{tr} \left\{ \boldsymbol{\xi}_p \boldsymbol{\xi}_p^\top \boldsymbol{A} \right\} \,, \end{aligned} \tag{4.24}$$

where $\boldsymbol{A} = \sum_{t=1}^{T} \left[\boldsymbol{\Phi}_t \boldsymbol{B}_{t|T} \boldsymbol{\Phi}_t^\top + \left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t \boldsymbol{\beta}_{t|T} \right) \left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t \boldsymbol{\beta}_{t|T} \right)^\top \right]$. The `R` function `uniroot` can be used to solve the score function.

The parameterization of the residual covariance matrix $\boldsymbol{G}$ and the estimation method of the parameter σ^2 given the eigenvalues and eigenfunctions are important to the development and estimation of the spatio-temporal model in this thesis. The disadvantage of this method, however, is that the EOFs are estimated before fitting the state space model, i.e. the estimation is based on potentially correlated data. Whereas ideally, the estimation should use independent data. Therefore, this approach does not solve the problem put forward at the end of Chapter 3. There are also criticisms that the leading EOFs may not be adequate to explain the dominant system dynamics, despite their power in describing the variation in the data (Cressie & Wikle, 2011). On the contrary, the dynamics might be governed by a component that accounts for only a small proportion of the variance.

Preliminary 2: the STRE model and FRF One type of STRE model for very large spatio-temporal data sets has a data model for dimension reduction,

$$\begin{aligned} Z(\boldsymbol{s};t) &= Y(\boldsymbol{s};t) + \epsilon(\boldsymbol{s};t) \\ Y(\boldsymbol{s};t) &= \mu_t(\boldsymbol{s}) + \boldsymbol{S}_t(\boldsymbol{s})^\top \boldsymbol{\beta}_t + \zeta(s;t) \ . \end{aligned} \tag{4.25}$$

Here $Z(\boldsymbol{s};t)$ is the observation and $Y(\boldsymbol{s};t)$ is the true underlying process, which is further decomposed into a spatial mean function $\mu_t(s)$, a spatio-temporal dynamic component $\boldsymbol{S}_t(\boldsymbol{s})^\top \boldsymbol{\beta}_t$ and an additional random component $\zeta(\boldsymbol{s};t)$. Dimension reduction comes with the basis representation of the dynamic component as a spatial basis $\boldsymbol{S}_t(\boldsymbol{s})$ multiplied by the time-varying basis coefficient vector $\boldsymbol{\beta}_t$. The process model of $\boldsymbol{\beta}_t$ is specified as

$$\boldsymbol{\beta}_t = \boldsymbol{M}_t \boldsymbol{\beta}_{t-1} + \boldsymbol{u}_t \ ,$$

with propagator matrix $\boldsymbol{M}_t$ and residual $\boldsymbol{u}_t$. It is assumed that the series $\{\boldsymbol{\zeta}_t\}_{t=1}^T$, where $\boldsymbol{\zeta}_t = (\zeta(\boldsymbol{s}_1;t), \cdots, \zeta(\boldsymbol{s}_n;t))^\top$, is not temporally correlated and only depends on the observations at time t. It is also assumed that the series $\{\boldsymbol{\zeta}_t\}_{t=1}^T$ is independent of the series $\{\boldsymbol{\beta}_t\}_{t=1}^T$. Both $\{\boldsymbol{\beta}_t\}_{t=1}^T$ and $\{\boldsymbol{\zeta}_t\}_{t=1}^T$ are independent from the measurement error process $\{\boldsymbol{\epsilon}_t\}_{t=1}^T$.

The estimation of model components $\boldsymbol{\beta}_t$ and $\boldsymbol{\zeta}_t$ uses the fixed rank filtering (FRF), where 'rank' refers to the dimension of the basis matrix $\boldsymbol{S}_t(\boldsymbol{s})$. It was proposed by Cressie *et al.* (2010), based on the fixed rank kriging method (Cressie & Johannesson, 2008) by incorporating the temporal component through a process model estimated using the Kalman filter/smoother. One thing worth pointing out is that, although $\boldsymbol{\zeta}_t$ is independent of $\boldsymbol{\beta}_t$, its estimation is accomplished by a filter based on the conditional distribution of $(\boldsymbol{\zeta}_t, \boldsymbol{\beta}_t)$ given data $\boldsymbol{Z}_{1:T}$ as

$$\boldsymbol{\zeta}_{t|t} = \boldsymbol{C}_t^\top \left(\boldsymbol{S}_t \boldsymbol{B}_{t|t-1} \boldsymbol{S}_t^\top + \boldsymbol{D}_t \right)^{-1} \left(\boldsymbol{Z}_t - \boldsymbol{\mu}_t - \boldsymbol{S}_t \boldsymbol{\beta}_{t|t-1} \right) \ . \tag{4.26}$$

Here $\boldsymbol{C}_t = \mathbf{Cov}[\boldsymbol{Z}_t, \boldsymbol{\zeta}_t]$ is the covariance matrix, $\boldsymbol{D}_t = \sigma_\zeta^2 \boldsymbol{I}_t + \sigma_\epsilon^2 \boldsymbol{W}_t$ is the covariance matrix of $\boldsymbol{\zeta}_t + \boldsymbol{\epsilon}_t$ and $\boldsymbol{\beta}_{t|t-1}$, $\boldsymbol{B}_{t|t-1}$ come from the Kalman filtering of $\boldsymbol{\beta}_t$. This suggests that, $\boldsymbol{\zeta}_t$ and $\boldsymbol{\beta}_t$ are no longer independent after conditioning on the data $\boldsymbol{Z}_{1:T}$.

What is enlightening about this method are the dependence/independence assumptions on the model components and its estimation using FRF embedded in an EM algorithm (Katzfuss & Cressie, 2011). To some extent, the conditional dependence of the two random components $\boldsymbol{\beta}_t$ and $\boldsymbol{\zeta}_t$ is crucial in terms of model estimation. However, the random component $\boldsymbol{\zeta}_t$, while accounting for the variation not covered by the system dynamic, cannot provide a conclusive

summary of the spatial variation. Unlike the eigenfunction and scores from a FPCA, $\boldsymbol{\zeta}_t$ can hardly be used as a measure of the spatial variation patterns in the data or their evolution. Therefore, it is not the optimal solution for the analysis in this thesis, where the spatial variation patterns are also of interest.

4.3.2 The proposed state space FPCA model (SS-FPCA)

Based on the above two preliminaries and all the basic elements introduced in previous sections, a spatio-temporal model with a system dynamic component and a FPCA component was proposed. The same notation as in Chapter 3 is used here, with subscript t indicating the time point and (x, y) indicating the spatial coordinates. The same hierarchies as in the STRE model (4.25) are used here, giving the following three levels.

(a) At the top level is a data model, which involves a dimension reduction of the underlying process through a basis representation,

$$\begin{aligned} Z_t(x,y) &= Y_t(x,y) + \epsilon_t(x,y) \\ Y_t(x,y) &= \mu_t(x,y) + \Phi(x,y)\boldsymbol{\beta}_t + \sum_{p=1}^{P} \Phi(x,y)\boldsymbol{\theta}_p\alpha_{pt} \\ &= \mu_t(x,y) + \Phi(x,y)\boldsymbol{\beta}_t + \Phi(x,y)\boldsymbol{\Theta}\boldsymbol{\alpha}_t \ , \end{aligned}$$

where $\mu_t(x,y)$ is a fixed mean component, $\Phi(x,y)\boldsymbol{\beta}_t$ is the system dynamic component (also referred to as the state space, or SS component) and $\Phi(x,y)\boldsymbol{\Theta}\boldsymbol{\alpha}_t$ is a K-L expansion of order P with orthonormal $\Phi(x,y)\boldsymbol{\Theta}$ (referred to as a FPCA component), accounting for the remaining spatial variations in the data.

(b) In the middle level is a process model, which assumes a random walk (or local level model) for the system dynamic,

$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{u}_t \, .$$

The motivation is, after appropriate detrending, this first order dependence structure would be adequate for most of the remote-sensing image time series considered in this thesis. Recall the exploratory analysis in section 2.1, where it suggested that an AR(1) structure is appropriate for the majority of the LSWT time series after accounting for the seasonal structure. In addition, even though the above model means that the

element of $\boldsymbol{\beta}_t$ follows separate temporal evolution, the spatio-temporal dependence can be incorporated by the covariance structure of $\boldsymbol{u}_t$. More details on this issue are given in the paragraphs explaining the model assumptions.

(c) At the bottom level, the following distributions are assigned to the data and process model components. The measurement errors $\epsilon_t(x, y)$ are assumed to be i.i.d. normally distributed as $\mathcal{N}(0, \sigma^2)$. The residuals of the process model $\boldsymbol{u}_t$ are assumed to be normally distributed as $\mathcal{N}(\mathbf{0}, \boldsymbol{H})$, where $\boldsymbol{H}$ is symmetric, positive definite, but not necessarily diagonal. Finally, random coefficient vector $\boldsymbol{\alpha}_t$ is required to satisfy the assumptions of the PC scores as defined in Chapter 3. That is, $\boldsymbol{\alpha}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda})$ with $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \cdots, \lambda_P\}$. Particularly, λ_p, $p = 1, \cdots, P$, are arranged in decreasing order.

Putting (a), (b) and (c) together, the proposed model (using matrix notation) is

$$\begin{aligned} \boldsymbol{Z}_t &= \boldsymbol{\mu}_t + \boldsymbol{\Phi}_t\boldsymbol{\beta}_t + \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t \\ \boldsymbol{\beta}_t &= \boldsymbol{\beta}_{t-1} + \boldsymbol{u}_t \end{aligned} \tag{4.27}$$

where

$$\begin{aligned} &\boldsymbol{\Phi}^\top\boldsymbol{\Phi} = \boldsymbol{I}, \quad \boldsymbol{\Theta}^\top\boldsymbol{\Theta} = \boldsymbol{I} \\ &\boldsymbol{\alpha}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}) \,, \qquad \boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \cdots, \lambda_P\} \\ &\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{I}) \\ &\boldsymbol{u}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{H}) \,. \end{aligned}$$

In model (4.27), $\boldsymbol{Z}_t$ is the data vector, $\boldsymbol{\Phi}_t$ is a (bivariate) basis matrix and $\boldsymbol{\epsilon}_t$ and $\boldsymbol{u}_t$ are residual vectors of the data and process models. Model (4.27) is referred to as the state space functional principal component analysis and is abbreviated as the SS-FPCA model. Note that the subscript t in $\boldsymbol{\Phi}_t$ is used to reflect the influence of the missing data at time point t. The same notation was used in Chapter 3 for the MM-FPCA. In the following sections, the subscript t is dropped only when it is referred to the fitted results or a general case without emphasizing on the sparsity.

The SS-FPCA model extends the MM-FPCA in James *et al.* (2000) by incorporating the temporal dependence through a hierarchical design. The time invariant mean function $\boldsymbol{\Phi}_t\boldsymbol{\beta}$ in the MM-FPCA is replaced by a time dependent mean function $\boldsymbol{\Phi}_t\boldsymbol{\beta}_t$. The dynamic of this function is governed by a first order random walk process in a lower hierarchy. With the

system dynamic component accounting for the temporal correlation, the FPCA component would be estimated based on (nearly) temporally independent data.

The SS-FPCA model also modifies the STRE model in Cressie *et al.* (2010) by allowing more than one non-dynamic random component ($\boldsymbol{\zeta}_t$ in model (4.25)). In addition, the SS-FPCA imposes structures on these non-dynamic random components so that they can provide a summary of the spatial variation patterns. As the constraints follow the assumptions of the MM-FPCA, the resulting random component would consist of spatial variation patterns of the corresponding PCs. In consequence, the random components in the SS-FPCA model would be more informative than their counterpart in the STRE model (4.25) and would fit the problem in this thesis better. Finally, dimension reduction is achieved through the functional representation of the random components and the truncation of the number of functional PCs. The mixed effect nature of the model suggests that the missing observations can be accommodated in a straightforward way. Both are desirable properties in terms of the application to high dimensional, sparse remote-sensing data.

The details of the model specifications are listed below.

(a) It should be pointed out that using a local level model for the system transition equation in model (4.27) is out of concern for computational simplicity. It is possible to assume $\boldsymbol{\beta}_t = \boldsymbol{M}\boldsymbol{\beta}_{t-1} + \boldsymbol{u}_t$ for $\boldsymbol{M} \neq \boldsymbol{I}$ (Cressie *et al.*, 2010, Katzfuss & Cressie, 2011), such as $\boldsymbol{M} = \text{diag}\{m_1, \cdots, m_K\}$. However, estimating such a propagator matrix $\boldsymbol{M}$ can be difficult and computationally intensive. It usually requires prior information to get a suitable design of $\boldsymbol{M}$ and sensible estimation result (Cressie & Wikle, 2011). This can be hard for image time series, especially when $\boldsymbol{\beta}_t$ is a vector of basis coefficient. On the other hand, the local level model assumption, though being non-stationary, can be appropriate for the remote-sensing environmental measurements, as many of them are indeed non-stationary in reality.

(b) No special structure is imposed on the residual covariance matrix in the process model, $\boldsymbol{H}$. The only requirement is it being positive definite. It is possible to parameterize the $\boldsymbol{H}$ matrix, as suggested in section 4.3.1. This typically involves imposing certain spatial structure on the $\boldsymbol{H}$ matrix, such as a covariogram model and a CAR model. In this way, the spatio-temporal dynamic of the process can be modelled. Specifically, imposing a diagonal structure on $\boldsymbol{H}$ would suggest separate evolution of the elements in $\boldsymbol{\beta}_t$. It can significantly simplify the estimation, but is often unrealistic in practice. This is

because most of the basis functions are not spatially 'separable', in the sense that their compact supports often overlap in space. Due to this overlapping, the elements in the basis coefficient vector $\boldsymbol{\beta}_t$ would not be independent. There might be some cases where the diagonal assumption is adequate, but this relies on a strong assumption of space-time separability and a spatially non-overlapping basis. To avoid setting too many impractical constraints, the $\boldsymbol{H}$ matrix is left unstructured for the SS-FPCA model, so that the residual process $\boldsymbol{u}_t$ can be used to account for the (unknown) spatio-temporal dependence.

(c) It is required that $\Phi(x, y)$ is a orthonormal basis and $\boldsymbol{\Theta}$ is a column orthonormal matrix. This is to ensure that the estimated results are valid eigenvalues and eigenfunctions from a FPCA. The rationals for these assumptions have already been explained in Chapter 3. Depending on the estimation methods, a final orthonormalization might be applied to the estimated $\widehat{\boldsymbol{\Theta}}$ as in James *et al.* (2000).

(d) To ensure the identifiability of the model, further assumptions are made on the random components $\boldsymbol{\beta}_t$, $\boldsymbol{\alpha}_t$ and model residuals $\boldsymbol{\epsilon}_t$, $\boldsymbol{u}_t$. It is assumed that $\{\boldsymbol{\beta}_t\}_{t=1}^T$ and $\{\boldsymbol{\alpha}_t\}_{t=1}^T$ are independent; $\{\boldsymbol{\beta}_t\}_{t=1}^T$ is independent of $\{\boldsymbol{\epsilon}_t\}_{t=1}^T$; $\{\boldsymbol{\alpha}_t\}_{t=1}^T$ is independent of $\{\boldsymbol{u}_t\}_{t=1}^T$ and $\{\boldsymbol{\epsilon}_t\}_{t=1}^T$. In addition, it is assumed that the estimation of $\boldsymbol{\beta}_t$ at time point t relies on information from all the observed data $\{\boldsymbol{Z}_t\}_{t=1}^T$. Whereas the prediction of $\boldsymbol{\alpha}_t$ at time point t requires only the information from $\boldsymbol{Z}_t$ as in a FPCA. This assumption is similar to that of $\boldsymbol{\zeta}_t$ in the STRE model (4.25) in Cressie *et al.* (2010). The difference between the two models is that, while $\boldsymbol{\beta}_t$ and $\boldsymbol{\zeta}_t$ are independent but not conditionally independent given $\boldsymbol{Z}_{1:T}$ in the STRE model (4.25), $\boldsymbol{\beta}_t$ and $\boldsymbol{\alpha}_t$ are assumed to be also conditionally independent given $\boldsymbol{Z}_{1:T}$ in the SS-FPCA.

It should be acknowledged at this stage that, it is always better to take into account the conditional dependence of $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$, wherever possible. However, the conditional independence assumption could be justified through the fact that $\boldsymbol{\alpha}_t$ are essentially PC scores. In a FPCA computed using matrix decomposition, the PC scores are obtained after the eigen-decomposition of the covariance matrix. In other words, they are not directly related to the extraction of the eigenfunctions and eigenvalues. Although the estimation of the SS-FPCA model would inevitably involve iterative steps, the conditional correlation between $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ is not presumed to have a large influence on model estimation if the algorithm is designed sensibly. Moreover, the evaluation of the conditional distribution for $\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t | \boldsymbol{Z}_{1:T}$ is extremely difficult due to the different temporal dependence structures of $\boldsymbol{\beta}_t$ and $\boldsymbol{\alpha}_t$. To be specific, $\boldsymbol{\beta}_t$ is governed by a

first-order Markov structure through $\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1}$, which means the distribution of $f(\boldsymbol{\beta}_t)$ for each time point t cannot be separated from the joint distribution of $f(\boldsymbol{\beta}_{1:T})$ due to the dependence. Whereas $\boldsymbol{\alpha}_t$ does not depend on its temporal neighbours and relies solely on the information at time t. Considering the complexity in determining the joint distribution $f(\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t|\boldsymbol{Z}_{1:T})$, this thesis assumes that $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ are conditionally independent.

(e) As mentioned in section 4.1, it is sometimes sensible to use different bases for the state space and the FPCA component, such as

$$Z_t(x,y) = \mu_t(x,y) + \Phi_\beta(x,y)\boldsymbol{\beta}_t + \Phi_\xi(x,y)\boldsymbol{\Theta}\boldsymbol{\alpha}_t + \epsilon_t(x,y)\,.$$

This would offer more flexibility in describing the spatial/temporal variations. For example, basis $\Phi_\beta(x,y)$ may be designed to capture the large scale temporal variation; whereas $\Phi_\xi(x,y)$ is intended to explain the smaller scale spatial variation via the FPCA. However, this could complicate the estimation of the model, as some simplifications (e.g. the matrix identity used in inverting high-dimensional matrix) may not be plausible if two different bases are used. As far as the problem in this thesis is concerned, the gain from specifying two different bases may not compensate the loss in the computational cost. Therefore, it is assumed that $\Phi_\beta(x,y) = \Phi_\xi(x,y)$.

4.4 Spatio-temporal model estimation

The SS-FPCA model is a mixed effect model with fixed effect component $\boldsymbol{\mu}_t$ and random effect components $\boldsymbol{\Phi}_t\boldsymbol{\beta}_t$ and $\boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t$. The fixed effect $\boldsymbol{\mu}_t$ can be estimated as a constant or as an overall mean function $\boldsymbol{X}\boldsymbol{b}$ using the generalized least squares as in Cressie *et al.* (2010). Without loss of generality, it is assumed that $\boldsymbol{\mu}_t = 0$ in the following content. Under this setting, the observed information of the model is $\{\boldsymbol{Z}_t\}_{t=1}^T$ and the unobserved information is $\{\boldsymbol{\beta}_t\}_{t=1}^T$ and $\{\boldsymbol{\alpha}_t\}_{t=1}^T$. Then the parameter set of the model becomes $\Psi = \{\boldsymbol{H}, \boldsymbol{\Theta}, \boldsymbol{\Lambda}, \sigma^2\}$. Fitting the SS-FPCA model involves both the estimation of Ψ and the prediction of $\{\boldsymbol{\beta}_t\}_{t=1}^T$, $\{\boldsymbol{\alpha}_t\}_{t=1}^T$ based on the observed data $\{\boldsymbol{Z}_t\}_{t=1}^T$. Despite the simplification of distributional assumptions brought by the hierarchical design, the observed data and the complete data log-likelihood functions of the model are still non-trivial. This brings computational challenges to model estimation.

The first challenge is associated with the high dimensionality (i.e. large volume) of the data. For a time series of remote-sensing images, even with a low-dimensional representation, the estimation of some model components can be computationally intensive. Hence, when it comes to the choice between the EHM (empirical hierarchical modelling) and the BHM (Bayesian hierarchical modelling) frameworks (refer to section 4.1), the problems brought by the data dimension must be taken into account. In general, it is believed that the EHM using an EM-based algorithm would require less computational cost than the BHM approach. Implementation using the BHM could encounter difficulties, e.g. sampling from high-dimensional posterior distributions and monitoring convergence. Although it is slightly restricted in terms of the types of model it can handle, the EHM approach is more computational friendly. As far as the SS-FPCA model is concerned, implementation using the EHM can be done with analytical solutions or low-dimensional numerical optimizations. It also maintains consistency with the estimation method used in the MM-FPCA, which has been shown to be reliable in literature such as James *et al.* (2000), Peng & Paul (2009). Therefore, in this thesis, the EHM approach is adopted.

The second challenge is the identifiability of the model components. The SS-FPCA model consists of two random effect components. Despite the distributional assumptions on $\boldsymbol{\Phi}_t\beta_t$, $\boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t$ and $\boldsymbol{\epsilon}_t$ to help distinguish them from each other, there is still ambiguity in some measures used in the estimation. For example, the covariance matrix of the data model,

$$\mathbf{Cov}[\boldsymbol{Z}_t] = \boldsymbol{\Phi}_t\boldsymbol{B}_t\boldsymbol{\Phi}_t^\top + \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top + \sigma^2\boldsymbol{I}\,,$$

where $\boldsymbol{B}_t = \mathbf{Cov}[\boldsymbol{\beta}_t]$, involves both random components and measurement errors. It is hard to identify the influence of each component on $\mathbf{Cov}[\boldsymbol{Z}_t]$. Therefore, it is ideal to exploit an algorithm that can avoid the (frequent) use of such identification.

The above two challenges lead to the problem of the design of the EM-type algorithm. It is preferable to have analytical solutions in both the E-step and M-step estimations. This is not always possible for a complex model. However, by augmenting the data appropriately, that is, by purposefully defining the 'missing information', the computation of the E-step and M-step can be simplified. Even if an analytical solution is not available, there could be a gain in computational efficiency by transforming a high-dimensional optimization problem to a low-dimensional one. Therefore, a flexible algorithm with a suitable data augmentation scheme is required.

4.4.1 The proposed estimation framework: AECM

One algorithm satisfying the requirements is the Alternating Expectation - Conditional Maximization (AECM) algorithm. This algorithm was first proposed by Meng & Van Dyk (1997) and was developed based on various extensions of the classic EM algorithm formalized by Dempster *et al.* (1977). It makes use of both data augmentation and model reduction to create a more efficient algorithm for models with complex structures. Under this framework, the estimation of the SS-FPCA model can be done using only analytical solutions, or simple 1-dimensional numerical optimizations.

Two concepts inspiring the development of the EM-type algorithms are data augmentation and model reduction. Specifically, data augmentation refers to the '*methods for constructing iterative optimization or sampling algorithms via the introduction of unobserved data or latent variables*' (Van Dyk & Meng, 2001). Model reduction refers to '*using a set of conditional distributions in a computation method designed to learn about the corresponding joint distribution*' (Van Dyk & Meng, 2010). Both data augmentation and model reduction, when appropriately applied, lead to an improved algorithm. Various methods are developed based on these two concepts. A diagram showing the development of the EM-type algorithms is given in Figure 4.5. The diagram was originally created by Van Dyk & Meng (2010), which categorizes the extended algorithms based on different modelling techniques.

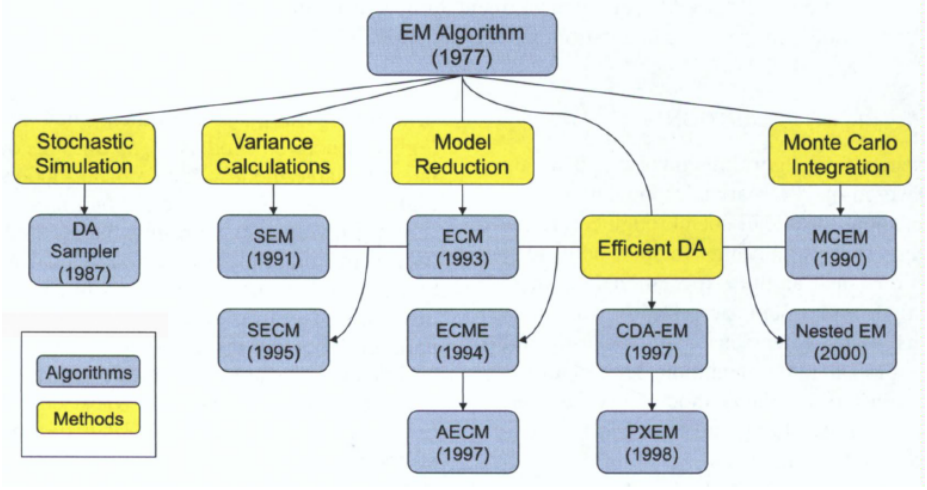

FIGURE 4.5: The family tree of the EM-type method. This figure is originally created by Van Dyk & Meng (2010).

The category associated with the AECM algorithm is shown in the middle branch, labeled as 'model reduction'. It begins with the Expectation - Conditional Maximization (ECM)

algorithm. This algorithm divides the M-step into several conditional maximization (CM) steps by partitioning the parameter space, to ease the computational burden in the M-step (Meng & Rubin, 1993). The Expectation - Conditional Maximization Either (ECME) algorithm goes a step further. It allows the maximization of either the actual data likelihood function $\mathcal{L}(\Psi; \boldsymbol{Z}_{obs})$, or the target function $\mathcal{Q}(\Psi; \Psi^{(it)})$, depending on different CM-steps (Liu & Rubin, 1994). The Space Alternating Generalized EM (SAGE) algorithm, though not shown in the diagram, is another important step in the development of the middle branch. The algorithm was proposed at the same time as ECME, but takes a different route by using different target function $\mathcal{Q}_s(\Psi; \Psi^{(it)})$, $s = 1, \cdots, S$, according to the design of the parameter subspaces and their corresponding hidden data spaces (Fessler & Hero, 1994). Finally, the AECM algorithm merges the ECME and the SAGE into a more general method (Meng & Van Dyk, 1997). It consists of C ($C \geq 1$) cycles within each iteration. Each cycle corresponds to one type of data augmentation and is paired with S_c ($S_c \geq 1$) CM-steps. The subscript c of S indicates that the number of CM-steps is allowed to vary with cycles (Meng & Van Dyk, 1997), giving full flexibility to the design of the algorithm.

Specifically, omitting the iteration index (it), the target function in the E-step of the $(c+1)$-th cycle of the AECM algorithm can be written as (Meng & Van Dyk, 1997)

$$\mathcal{Q}^{[c+1]}\left(\Psi; \Psi^{[c]}\right) = \mathbf{E}\left[\mathcal{L}\left(\Psi; \boldsymbol{Z}_{aug}^{[c+1]}\right) \middle| \boldsymbol{Z}_{obs}, \Psi^{[c]}\right]; \tag{4.28}$$

then the s-th CM-step in cycle $c+1$ calculates $\Psi^{[c+\frac{s}{S_{c+1}}]}$ such that

$$\begin{aligned} &\mathcal{Q}^{[c+1]}\left(\Psi^{[c+\frac{s}{S_{c+1}}]}; \Psi^{[c]}\right) \geq \mathcal{Q}^{[c+1]}\left(\Psi; \Psi^{[c]}\right) \\ &\forall\ \Psi \in \mathcal{W}_s^{[c+1]} \equiv \left\{\Psi \in \mathcal{W} : g_s^{[c+1]}(\Psi) = g_s^{[c+1]}\left(\Psi^{[c+\frac{s}{S_{c+1}}]}\right)\right\} \end{aligned} \tag{4.29}$$

where $g_s^{[c+1]}(\Psi)$ is the constraint function of the s-th CM-step in cycle $c+1$. Due to its flexibility, the AECM algorithm has seen wide applications. Examples include the estimation of mixture models (McLachlan *et al.*, 2003, McNicholas & Murphy, 2008), fitting mixed models with non-exponential family distributions (Ho & Lin, 2010), etc.

4.4.2 The 2-cycle AECM algorithm for the SS-FPCA model

The AECM algorithm for the SS-FPCA model consists of two cycles, each with its own data augmentation scheme. See the flow chart in Figure 4.6. Specifically, the first cycle estimates the parameter $\boldsymbol{H}$ and random coefficient $\boldsymbol{\beta}_t$ for the dynamic component; the second cycle

estimates the parameters and random effects associated with the FPCA component, $\boldsymbol{\Theta}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\alpha}_t$. The residual variance σ^2 can be estimated in both cycles; the preference here is to estimate it in the second cycle. In terms of notation, the subscripts *obs, mis* and *aug* indicate the observed, missing and augmented data respectively. The superscript [1], [2] are the cycle indexes and superscript (it) is the iteration index. The subscript $1:t$ refers to the time series from time point 1 to t, e.g. $\boldsymbol{Z}_{1:T} = \{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_T\} = \{\boldsymbol{Z}_t\}_{t=1}^T$.

AECM - CYCLE 1

- The observed data in this cycle are $\boldsymbol{Z}_{obs} = \{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_T\}$; the missing data are $\boldsymbol{Z}_{mis} = \{\boldsymbol{\beta}_0, \cdots, \boldsymbol{\beta}_T\}$. Combining the observed and missing data generates the augmented data as $\boldsymbol{Z}_{aug} = \{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_T\,;\, \boldsymbol{\beta}_0, \cdots, \boldsymbol{\beta}_T\}$, which is denoted as $\boldsymbol{Z}^{[1]}$.
- The parameter set to be updated in this cycle is $\Psi^{[1]} = \{\boldsymbol{H}\}$; the parameter set fixed at current value in this cycle is $\widetilde{\Psi}^{[1]} = \{\boldsymbol{\Theta}, \boldsymbol{\Lambda}, \sigma^2\}$
- In the it-th iteration, the current estimates of the parameters in this cycle are denoted as $\Psi^{(it-1)} = \{\boldsymbol{H}^{(it-1)}, \boldsymbol{\Theta}^{(it-1)}, \boldsymbol{\Lambda}^{(it-1)}, \sigma^{2(it-1)}\}$
- The estimation of $\{\boldsymbol{\beta}_t\}_{t=1}^T$ uses the Kalman filter/smoother (with threshold).
- The estimation of $\{\boldsymbol{\alpha}_t\}_{t=1}^T$ is not considered in this cycle. The component $\boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t$ is treated as part of the model residual, with its influence reflected in the covariance matrix $\mathbf{Cov}[\boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t]$
- Based on the state-process evolution of the model and the idea of 'sequential implementation' (i.e. to update the previous filtering distribution each time new data become available) in Cressie & Wikle (2011), the complete data distribution in this cycle is

$$\begin{aligned} f(\boldsymbol{Z}_{1:T}, \boldsymbol{\beta}_{0:T}; \Psi) &= f\left(\boldsymbol{Z}_{1:T} | \boldsymbol{\beta}_{0:T};\, \Psi\right) f\left(\boldsymbol{\beta}_{0:T};\, \Psi\right) \\ &= \prod_{t=1}^{T} f\left(\boldsymbol{Z}_t | \boldsymbol{\beta}_t;\, \Psi\right) f\left(\boldsymbol{\beta}_t | \boldsymbol{\beta}_{t-1};\, \Psi\right) f\left(\boldsymbol{\beta}_0;\, \Psi\right) . \end{aligned} \tag{4.30}$$

The $\mathcal{Q}$-function (i.e. the target function) from the E-Step in this cycle is

$$\begin{aligned} \mathcal{Q}^{[1]}\left(\Psi^{[1]}\,;\, \Psi^{(it-1)}\right) &= \mathbf{E}\left[-2\mathcal{L}\left(\Psi^{[1]}; \boldsymbol{Z}^{[1]}, \widetilde{\Psi}^{[1]}\right) \middle|\, \boldsymbol{Z}_{1:T}, \Psi^{(it-1)}\right] \\ &= \mathbf{E}\left[-2\log f\left(\boldsymbol{Z}_{1:T}, \boldsymbol{\beta}_{0:T}\,;\, \Psi^{[1]}, \widetilde{\Psi}^{[1]}\right) \middle|\, \boldsymbol{Z}_{1:T}, \Psi^{(it-1)}\right] \\ &= \mathbf{E}\left[-2\log f\left(\boldsymbol{Z}_{1:T}, \boldsymbol{\beta}_{0:T}\,;\, \boldsymbol{H}, \boldsymbol{\Theta}^{(it-1)}, \boldsymbol{\Lambda}^{(it-1)}, \sigma^{2(it-1)}\right) \middle|\, \boldsymbol{Z}_{1:T}, \Psi^{(it-1)}\right] . \end{aligned} \tag{4.31}$$

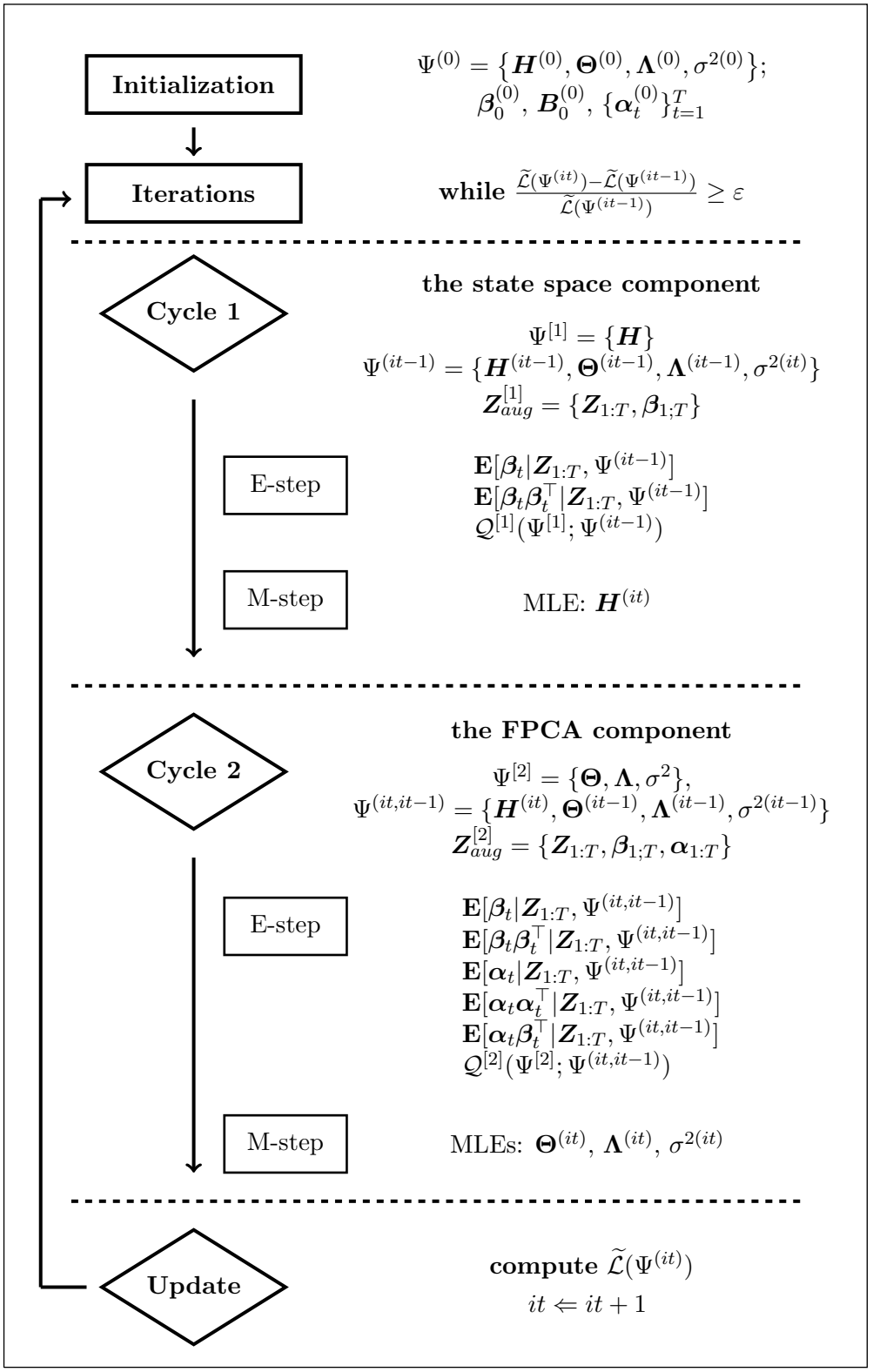

FIGURE 4.6: A diagram showing the 2-cycle AECM algorithm for the SS-FPCA model. It features the basic settings of the two cycles, the conditional expectations and MLEs to be computed in each cycle.

The detailed expression of the $\mathcal{Q}^{[1]}$ function follows the same form as the E-step equations (4.12) with the propagator matrix being $\boldsymbol{M} = \boldsymbol{I}$, which is

$$\begin{aligned}\mathcal{Q}^{[1]} = &\sum_{t=1}^{T}\left\{\log(|\boldsymbol{G}_t|) + \mathbf{E}\left[(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_t)^\top \boldsymbol{G}_t^{-1}(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_t)\middle|\boldsymbol{Z}_{1:T}, \Psi^{(it-1)}\right]\right\} \\ &+ \sum_{t=1}^{T}\left\{\log(|\boldsymbol{H}|) + \mathbf{E}\left[(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^\top \boldsymbol{H}^{-1}(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})\middle|\boldsymbol{Z}_{1:T}, \Psi^{(it-1)}\right]\right\} \\ &+ \left\{\log(|\boldsymbol{B}_0|) + \mathbf{E}\left[(\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \boldsymbol{B}_0^{-1}(\boldsymbol{\beta}_0 - \boldsymbol{\beta})\middle|\boldsymbol{Z}_{1:T}, \Psi^{(it-1)}\right]\right\} + \text{constant}\,.\end{aligned} \tag{4.32}$$

Specifically, the covariance matrix $\boldsymbol{G}_t$ in the it-th iteration is parameterized as

$$\boldsymbol{G}_t^{(it)} = \mathbf{Cov}\left[\boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t\middle|\Psi^{(it-1)}\right] = \boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)}\boldsymbol{\Theta}^{(it-1)\top}\boldsymbol{\Phi}_t^\top + \sigma^{2(it-1)}\boldsymbol{I}\,,$$

which is implied by the fact that the entire $\boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t$ is treated as the model residual. The computational form of the $\mathcal{Q}^{[1]}$ function is the same as equation (4.13) in section 4.2.1, with $\{\boldsymbol{\beta}_{t|T}\}_{t=1}^T$ obtained using the KF/KS under the current parameter estimates $\Psi^{(it-1)}$. Details can be found in Appendix B.1.

In the M-Step, the MLE of $\boldsymbol{H}$ can be shown to follow the expression

$$\boldsymbol{H}^{(it)} = \frac{1}{T}\left(\boldsymbol{V}_{11} - 2\boldsymbol{V}_{10} + \boldsymbol{V}_{00}^\top\right)\,, \tag{4.33}$$

where $\boldsymbol{V}_{11}$, $\boldsymbol{V}_{10}$ and $\boldsymbol{V}_{00}$ are defined in the same way as in equation (4.13). This is essentially the same as the MLE derived in Shumway & Stoffer (2006), with $\boldsymbol{M} = \boldsymbol{I}$ plugged in.

AECM - CYCLE 2

- The observed data in this cycle are $\boldsymbol{Z}_{obs} = \{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_T\}$; the missing data are $\boldsymbol{Z}_{mis} = \{\boldsymbol{\beta}_0, \cdots, \boldsymbol{\beta}_T\,; \boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_T\}$. Hence, the augmented data become $\boldsymbol{Z}_{aug} = \{\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_T\,; \boldsymbol{\beta}_0, \cdots, \boldsymbol{\beta}_T\,; \boldsymbol{\alpha}_1, \cdots, \alpha_T\}$, which is denoted as $\boldsymbol{Z}^{[2]}$.

- The parameter set to be updated in this cycle is $\Psi^{[2]} = \{\boldsymbol{\Theta}, \boldsymbol{\Lambda}, \sigma^2\}$; the parameter set fixed at current value in this cycle is $\widetilde{\Psi}^{[2]} = \{\boldsymbol{H}\}$

- In the it-th iteration, the current estimates of the parameters in this cycle are denoted as $\Psi^{(it,it-1)} = \{\boldsymbol{H}^{(it)}, \boldsymbol{\Theta}^{(it-1)}, \boldsymbol{\Lambda}^{(it-1)}, \sigma^{2(it-1)}\}$. Note that the superscript for parameter $\boldsymbol{H}$ is (it) instead of $(it-1)$, for it has been updated in cycle 1

- The estimation of $\{\boldsymbol{\beta}_t\}_{t=1}^T$ proceeds using the KF/KS (with threshold), with the up-to-date estimate of $\boldsymbol{H}^{(it)}$

- The estimation of $\{\boldsymbol{\alpha}_t\}_{t=1}^T$ is carried out using the FPCA algorithm, but some details are subject to changes due to the presence of $\{\boldsymbol{\beta}_t\}_{t=1}^T$.

- Under the assumption that $\{\boldsymbol{\alpha}_t\}_{t=1}^T$ and $\{\boldsymbol{\beta}_t\}_{t=1}^T$ are independent, the complete data distribution in this cycle can be written as

$$\begin{aligned} f(\boldsymbol{Z}_{1:T}, \boldsymbol{\beta}_{0:T}, \boldsymbol{\alpha}_{1:T}; \Psi) &= f\left(\boldsymbol{Z}_{1:T} | \boldsymbol{\beta}_{0:T}, \boldsymbol{\alpha}_{1:T};\, \Psi\right) f\left(\boldsymbol{\alpha}_{1:T}, \boldsymbol{\beta}_{0:T};\, \Psi\right) \\ &= \prod_{t=1}^{T} f\left(\boldsymbol{Z}_t | \boldsymbol{\beta}_t, \boldsymbol{\alpha}_t;\, \Psi\right) f\left(\boldsymbol{\beta}_t | \boldsymbol{\beta}_{t-1};\, \Psi\right) f\left(\boldsymbol{\alpha}_t;\, \Psi\right) f\left(\boldsymbol{\beta}_0;\, \Psi\right) \end{aligned} \tag{4.34}$$

The $\mathcal{Q}$-function from the E-Step in this cycle is

$$\begin{aligned} \mathcal{Q}^{[2]}\left(\Psi^{[2]};\, \Psi^{(it,it-1)}\right) &= \mathbf{E}\left[-2\mathcal{L}\left(\Psi^{[2]}; \boldsymbol{Z}^{[2]}, \widetilde{\Psi}^{[2]}\right) \middle| \boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}\right] \\ &= \mathbf{E}\left[-2\log f\left(\boldsymbol{Z}_{1:T}, \boldsymbol{\beta}_{0:T}, \boldsymbol{\alpha}_{1:T};\, \Psi^{[2]}, \widetilde{\Psi}^{[2]}\right) \middle| \boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}\right] \\ &= \mathbf{E}\left[-2\log f\left(\boldsymbol{Z}_{1:T}, \boldsymbol{\beta}_{0:T}, \boldsymbol{\alpha}_{1:T};\, \boldsymbol{\Theta}, \boldsymbol{\Lambda}, \sigma^2, \boldsymbol{H}^{(it)}\right) \middle| \boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}\right], \end{aligned} \tag{4.35}$$

which can be written explicitly as

$$\begin{aligned} &\sum_{t=1}^{T}\left\{ n_t \log\left(\sigma^2\right) + \mathbf{E}\left[\frac{1}{\sigma^2}\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_t - \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t\right)^\top\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_t - \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t\right) \middle| \boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}\right]\right\} \\ &+ \sum_{t=1}^{T}\left\{ \log(|\boldsymbol{H}|) + \mathbf{E}\left[\left(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}\right)^\top \boldsymbol{H}^{-1}\left(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}\right) \middle| \boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}\right]\right\} \\ &+ \log(|\boldsymbol{B}_0|) + \mathbf{E}\left[\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right)^\top \boldsymbol{B}_0^{-1}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right) \middle| \boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}\right] \\ &+ \sum_{t=1}^{T}\left\{ \log(|\boldsymbol{\Lambda}|) + \mathbf{E}\left[\boldsymbol{\alpha}_t^\top \boldsymbol{\Lambda}^{-1}\boldsymbol{\alpha}_t \middle| \boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}\right]\right\} + \text{constant}\,, \end{aligned} \tag{4.36}$$

where n_t is the number of observations at time t. The computational form of function (4.36) requires the conditional expectations of $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t$ and their cross products. In particular, the expectations of $\boldsymbol{\beta}_t|\boldsymbol{Z}_{1:T}$ is obtained using the standard KF/KS under the current parameter estimate $\Psi^{(it,it-1)}$, without considering the estimation of $\boldsymbol{\alpha}_t$. The expectation of $\boldsymbol{\alpha}_t|\boldsymbol{Z}_{1:T}$ can be computed using a method similar to the one used in the MM-FPCA described in section 3.1.2. The difference is that $\boldsymbol{\beta}_t$ in the SS-FPCA is no longer the fixed effect coefficient vector, but the random coefficient of the state space component. Hence its influence needs to be accounted for in a different way. First of all, the covariance matrix of $\boldsymbol{Z}_t$, which is used in

calculating the distribution of $\boldsymbol{\alpha}_t|\boldsymbol{Z}_{1:T}$, becomes

$$\mathbf{Cov}[\boldsymbol{Z}_t] = \boldsymbol{\Phi}_t\mathbf{Cov}[\boldsymbol{\beta}_t]\boldsymbol{\Phi}_t^\top + \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top + \sigma^2\boldsymbol{I}\ .$$

It turns out that $\mathbf{E}[\boldsymbol{\beta}_t] = \mathbf{E}[\boldsymbol{\beta}_{t|T}]$ and hence $\mathbf{Cov}[\boldsymbol{\beta}_t] = \mathbf{E}[\boldsymbol{B}_{t|T}]$ from applying the property of double expectation. Therefore, $\boldsymbol{B}_{t|T}$ could be used as an estimator of $\mathbf{Cov}[\boldsymbol{\beta}_t]$ in the evaluation of $\mathbf{Cov}[\boldsymbol{Z}_t]$. Similarly, $\boldsymbol{\beta}_{t|T}$ can be used as an estimator of $\mathbf{E}[\boldsymbol{\beta}_t]$. Secondly, the expectation $\mathbf{E}[\boldsymbol{\alpha}_t\boldsymbol{\beta}_t^\top|\boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}]$ is required. Recall from section 4.3.2 that $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ are assumed to be independent given all the observed data $\boldsymbol{Z}_{1:T}$, i.e. $\mathbf{Cov}[\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t|\boldsymbol{Z}_{1:T}] = 0$. As a result, the expectation can be simplified to

$$\mathbf{E}\left[\boldsymbol{\alpha}_t\boldsymbol{\beta}_t^\top\middle|\boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}\right] = \mathbf{E}\left[\boldsymbol{\alpha}_t\middle|\boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}\right]\mathbf{E}\left[\boldsymbol{\beta}_t\middle|\boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}\right]^\top\ .$$

Finally, knowing $\boldsymbol{Z}_{1:T}$ and $\Psi^{(it,it-1)}$ essentially means that $\boldsymbol{\beta}_{t|T}$ and $\boldsymbol{B}_{t|T}$ are also known. Using these results, the following E-step prediction equations can be derived,

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}_t &= \mathbf{E}\left[\boldsymbol{\alpha}_t\middle|\boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}\right] \\
&= \left(\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)}\right)^\top\left(\boldsymbol{\Sigma}^{(it)}\right)^{-1}\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_{t|T}\right) \\
&= \frac{1}{\sigma^{2(it-1)}}\left(\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)}\right)^\top\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_{t|T}\right) \\
&\quad - \frac{1}{\sigma^{2(it-1)}}\left(\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)}\right)^\top\boldsymbol{\Phi}_t\left(\boldsymbol{R}^{(it)} + \boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t\right)^{-1}\boldsymbol{\Phi}_t^\top\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_{t|T}\right)\ ,
\end{aligned} \tag{4.37}$$

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^\top} &= \mathbf{E}\left[\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^\top\middle|\boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}\right] \\
&= \hat{\boldsymbol{\alpha}}_t\hat{\boldsymbol{\alpha}}_t^\top + \boldsymbol{\Lambda}^{(it-1)} - \left(\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)}\right)^\top\left(\boldsymbol{\Sigma}^{(it)}\right)^{-1}\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)} \\
&= \hat{\boldsymbol{\alpha}}_t\hat{\boldsymbol{\alpha}}_t^\top + \boldsymbol{\Lambda}^{(it-1)} - \frac{1}{\sigma^{2(it-1)}}\left(\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)}\right)^\top\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)} \\
&\quad + \frac{1}{\sigma^{2(it-1)}}\left(\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)}\right)^\top\boldsymbol{\Phi}_t\left(\boldsymbol{R}^{(it)} + \boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t\right)^{-1}\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)}\ ,
\end{aligned} \tag{4.38}$$

$$\widehat{\boldsymbol{\alpha}_t\boldsymbol{\beta}_t^\top} = \mathbf{E}\left[\boldsymbol{\alpha}_t\boldsymbol{\beta}_t^\top\middle|\boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}\right] = \hat{\boldsymbol{\alpha}}_t\boldsymbol{\beta}_{t|T}^\top\ . \tag{4.39}$$

In equation (4.37) and (4.38), $\boldsymbol{\Sigma}^{(it)} = \boldsymbol{\Phi}_t\boldsymbol{B}_{t|T}\boldsymbol{\Phi}_t^\top + \boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)}\boldsymbol{\Theta}^{(it-1)\top}\boldsymbol{\Phi}_t^\top + \sigma^{2(it-1)}\boldsymbol{I}$ and $\boldsymbol{R}^{(it)} = \sigma^{2(it-1)}\left(\boldsymbol{B}_{t|T} + \boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)}\boldsymbol{\Theta}^{(it-1)\top}\right)^{-1}$, which comes from applying the Woodbury identity to reduce the dimension of the matrix inversion. Further simplification can be made if $\boldsymbol{\Phi}$ is an orthonormal basis matrix, i.e. $\boldsymbol{\Phi}^\top\boldsymbol{\Phi} = \boldsymbol{I}$. However, in situations where there

are missing observations, even if $\mathbf{\Phi}$ is orthonormal by design, the evaluated basis $\mathbf{\Phi}_t$ for data at time t would not be orthonormal. Eventually, the computational form of function $\mathcal{Q}^{[2]}$ can be obtained by putting all the results from (4.36) to (4.39) together. More details of the above derivations are given in Appendix B.1.

The M-step functions for the MLEs are obtained from the $\mathcal{Q}^{[2]}$ function as

$$\begin{aligned}\sigma^{2(it)} &= \frac{1}{N}\sum_{t=1}^{T}\mathbf{E}\left[\boldsymbol{\epsilon}_t^\top\boldsymbol{\epsilon}_t\middle|\boldsymbol{Z}_{1:T},\Psi^{(it,it-1)}\right] \\ &= \frac{1}{N}\sum_{t=1}^{T}\mathbf{E}\left[\left(\boldsymbol{Z}_t-\mathbf{\Phi}_t\boldsymbol{\beta}_t-\mathbf{\Phi}_t\mathbf{\Theta}\boldsymbol{\alpha}_t\right)^\top\left(\boldsymbol{Z}_t-\mathbf{\Phi}_t\boldsymbol{\beta}_t-\mathbf{\Phi}_t\mathbf{\Theta}\boldsymbol{\alpha}_t\right)\middle|\boldsymbol{Z}_{1:T},\Psi^{(it,it-1)}\right] \\ &= \frac{1}{N}\sum_{t=1}^{T}\left[\operatorname{tr}\left\{\mathbf{\Phi}_t\boldsymbol{B}_{t|T}\mathbf{\Phi}_t^\top+\left(\boldsymbol{Z}_t-\mathbf{\Phi}_t\boldsymbol{\beta}_{t|T}\right)\left(\boldsymbol{Z}_t-\mathbf{\Phi}_t\boldsymbol{\beta}_{t|T}\right)^\top\right\}+\operatorname{tr}\left\{\mathbf{\Phi}_t\mathbf{\Theta}\widehat{\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^\top}\mathbf{\Theta}^\top\mathbf{\Phi}_t^\top\right\}\right. \\ &\quad\left.-2\operatorname{tr}\left\{\mathbf{\Phi}_t\mathbf{\Theta}\hat{\boldsymbol{\alpha}}_t\boldsymbol{Z}_t^\top-\mathbf{\Phi}_t\mathbf{\Theta}\widehat{\boldsymbol{\alpha}_t\boldsymbol{\beta}_t^\top}\mathbf{\Phi}_t^\top\right\}\right]\end{aligned} \tag{4.40}$$

where $N=\sum_{t=1}^{T}n_t$ is the sum of the number of observations at each time point t, and for $p=1,\cdots,P$,

$$\lambda_p^{(it)}=\frac{1}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^\top}_{(p,p)}, \tag{4.41}$$

$$\boldsymbol{\theta}_p^{(it)}=\left[\sum_{t=1}^{T}\widehat{\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^\top}_{(p,p)}\mathbf{\Phi}_t^\top\mathbf{\Phi}_t\right]^{-1}\sum_{t=1}^{T}\mathbf{\Phi}_t^\top\left[\hat{\boldsymbol{\alpha}}_{t(p)}\boldsymbol{Z}_t-\mathbf{\Phi}_t\left(\widehat{\boldsymbol{\alpha}_t\boldsymbol{\beta}_t^\top}_{(p,\cdot)}\right)^\top-\sum_{j\neq p}\widehat{\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^\top}_{(p,j)}\mathbf{\Phi}_t\hat{\boldsymbol{\theta}}_j\right], \tag{4.42}$$

where $\hat{\boldsymbol{\alpha}}_{t(p)}$ is the p-th element in vector $\hat{\boldsymbol{\alpha}}_t$, $\widehat{\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^\top}_{(p,j)}$ is the (p,j)-th element in $\widehat{\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^\top}$ and $\widehat{\boldsymbol{\alpha}_t\boldsymbol{\beta}_t^\top}_{(p,\cdot)}$ represents the p-th row of $\widehat{\boldsymbol{\alpha}_t\boldsymbol{\beta}_t^\top}$. Note that $\boldsymbol{\theta}_p^{(it)}$ is updated sequentially with $\hat{\boldsymbol{\theta}}_j=\boldsymbol{\theta}_j^{(it)}$ for $j<p$ and $\hat{\boldsymbol{\theta}}_j=\boldsymbol{\theta}_j^{(it-1)}$ for $j>p$.

AECM - evaluate convergence After running through cycle 1 and cycle 2, the parameter set is updated to $\Psi^{(it)}=\{\boldsymbol{H}^{(it)},\mathbf{\Theta}^{(it)},\mathbf{\Lambda}^{(it)},\sigma^{2(it)}\}$, completing one iteration of the AECM algorithm. The current estimate of $\Psi^{(it)}$ and the predictions of $\{\boldsymbol{\beta}_t\}$, $\{\boldsymbol{\alpha}_t\}$ are then used to evaluate the log-likelihood of the model, giving

$$\begin{aligned}&\widetilde{\mathcal{L}}\left(\Psi^{(it)};\boldsymbol{Z}_{1:T},\boldsymbol{\beta}_{1:T},\boldsymbol{\alpha}_{1:T}\right) \\ =&-\frac{1}{2}\sum_{t=1}^{T}\left\{n_t\log\left(\sigma^{2(it)}\right)+\frac{1}{\sigma^{2(it)}}\left(\boldsymbol{Z}_t-\mathbf{\Phi}_t\boldsymbol{\beta}_{t|T}-\mathbf{\Phi}_t\mathbf{\Theta}^{(it)}\hat{\boldsymbol{\alpha}}_t\right)^\top\left(\boldsymbol{Z}_t-\mathbf{\Phi}_t\boldsymbol{\beta}_{t|T}-\mathbf{\Phi}_t\mathbf{\Theta}^{(it)}\hat{\boldsymbol{\alpha}}_t\right)\right\} \\ &-\frac{1}{2}\sum_{t=1}^{T}\left\{\log\left(\left|\boldsymbol{H}^{(it)}\right|\right)+\left(\boldsymbol{\beta}_{t|T}-\boldsymbol{\beta}_{t-1|T}\right)^\top\left(\boldsymbol{H}^{(it)}\right)^{-1}\left(\boldsymbol{\beta}_{t|T}-\boldsymbol{\beta}_{t-1|T}\right)\right\}\end{aligned} \tag{4.43}$$

$$
\begin{aligned}
&-\frac{1}{2}\left\{\log\left(|\boldsymbol{B}_0|\right)+\left(\boldsymbol{\beta}_{0|T}-\boldsymbol{\beta}\right)^{\top}\boldsymbol{B}_0^{-1}\left(\boldsymbol{\beta}_{0|T}-\boldsymbol{\beta}\right)\right\}\\
&-\frac{1}{2}\sum_{t=1}^{T}\left\{\log\left(\left|\boldsymbol{\Lambda}^{(it)}\right|\right)+\hat{\boldsymbol{\alpha}}_t{}^{\top}\left(\boldsymbol{\Lambda}^{(it)}\right)^{-1}\hat{\boldsymbol{\alpha}}_t\right\}+\text{constant}\,.
\end{aligned}
$$

Function (4.43) can be used to evaluate the convergence of the 2-cycle AECM algorithm. The strategy here is to stop the iterations when the relative change of the $\widetilde{\mathcal{L}}\left(\Psi^{(it)};\cdots\right)$ values from two successive iterations is smaller than a threshold,

$$
\frac{\widetilde{\mathcal{L}}\left(\Psi^{(it)};\cdots\right)-\widetilde{\mathcal{L}}\left(\Psi^{(it-1)};\cdots\right)}{\widetilde{\mathcal{L}}\left(\Psi^{(it-1)};\cdots\right)}\leq\varepsilon\,,
$$

where ε is a pre-determined small number. Alternatively, the updated target function $\mathcal{Q}^{[2]}$ and/or some crucial parameters can be used to evaluate the convergence of the algorithm.

Estimate σ^2 in cycle 1 In the above design of the AECM algorithm, the residual variance σ^2 is estimated in cycle 2 as part of the FPCA component. The same parameter can be estimated in cycle 1, without changing the data augmentation scheme. However, the estimation equation for σ^2 changes and, unfortunately, no analytical solution is available. In particular, the MLE of σ^2 can be obtained using the same method as in Xu & Wikle (2007), which involves a 1-dimensional numerical optimization of the score function (4.24).

The algorithm can be implemented using `R` with code developed especially for it. Some simplifications with respect to the matrix inversion using the Sherman-Morrison-Woodbury identity are presented in Appendix B.2.

4.4.3 Initialization and finalization of the algorithm

The initialization of the state space component adopts the procedure in section 4.2.1.

(a) The initial state $\boldsymbol{\beta}_0$ follows a normal distribution, $\boldsymbol{\beta}_0\sim\mathcal{N}\left(\mathbf{0},\ \tau^2\boldsymbol{I}\right)$, where τ^2 is set to a relatively large number to reflect the lack of knowledge of the initial situation, e.g. $\tau^2=100$. The initial value of $\boldsymbol{\beta}_0$ is set to zero.

(b) The initial value of the covariance matrix of the state transition equation $\boldsymbol{H}$ is initialized as $\boldsymbol{H}^{(0)}=\sigma_h^2\boldsymbol{I}$, where $\sigma_h^2=\frac{1}{n}\sum_{i=1}^{n}\mathbf{Var}[Z_t(x_i,y_i)]$. Some other values of σ_h^2 may be used depending on the features of the data.

To initialize the FPCA component, the same procedure as used in the MM-FPCA is used.

(a) Before the launch of the 2-cycle AECM algorithm, no information on the state space component $\boldsymbol{\Phi}_t\boldsymbol{\beta}_t$ is available. In order to get a mean function, a fixed mean component $\boldsymbol{\Phi}_t\boldsymbol{\beta}$ is produced to represent the time-varying mean in this stage. The initial value of $\boldsymbol{\beta}$ is computed through fitting the linear regression model $\boldsymbol{Z} = \boldsymbol{\Phi}\boldsymbol{\beta}$ using vectorized data, $\boldsymbol{Z} = \mathbf{vec}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_T)$, and is denoted as $\boldsymbol{\beta}^{(0)}$.

(b) The sum of the residuals and the random effects are then calculated as $\hat{\boldsymbol{r}}_t = \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t = \boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}^{(0)}$. Rewriting $\boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t$ as $\boldsymbol{\Phi}_t\boldsymbol{\eta}_t$ and fitting the model $\hat{\boldsymbol{r}}_t = \boldsymbol{\Phi}_t\boldsymbol{\eta}_t + \boldsymbol{\epsilon}_t$ gives the least square estimate $\hat{\boldsymbol{\eta}}_t = (\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t)^{-1}\boldsymbol{\Phi}_t^\top\hat{\boldsymbol{r}}_t$. Apply the eigenvalue decomposition $\mathbf{Cov}[\hat{\boldsymbol{\eta}}_t] = \boldsymbol{U}\boldsymbol{\Sigma}_\alpha\boldsymbol{U}^\top$. The initial value of $\boldsymbol{\Theta}$ can be obtained as $\boldsymbol{\Theta}^{(0)} = \boldsymbol{U}$.

(c) Set the initial value of $\boldsymbol{\Lambda}$ as $\boldsymbol{\Lambda}^{(0)} = \boldsymbol{\Sigma}_\alpha$.

For the initial value of σ^2, two different approaches can be considered

(a) Initializing as $\sigma^{2(0)} = \frac{1}{N}\sum_{t=1}^{T}\hat{\boldsymbol{r}}_t^\top\hat{\boldsymbol{r}}_t$, where $\hat{\boldsymbol{r}}_t$ is obtained from the initialization procedure of the FPCA component and $N = \sum_t n_t$.

(b) Initializing as $\sigma^{2(0)} = \mathbf{Var}[Z_t(x, y)]$, which is a rough guage of the variance of the data.

It is widely recognized that the EM-type algorithm may be sensitive to initial values. So the influence of the above initialization methods on model estimation needs to be investigated. This is carried out later in Chapter 5.

Similar to the MM-FPCA in Chapter 3, the eigenfunctions estimated from the 2-cycle AECM algorithm are not necessarily orthonormal, which is the key assumption of the K-L expansion. Therefore, a final orthonormalization step is added to the converged AECM estimations, to transform $\boldsymbol{\Phi}\boldsymbol{\Theta}^*$ into an orthonormal matrix. The approach used here is the same as the final orthogonalization of the MM-FPCA given in section 3.1.2, which involves the eigen-decomposition $\boldsymbol{\Theta}^*\boldsymbol{\Lambda}^*\boldsymbol{\Theta}^{*\top} = \boldsymbol{\Theta}^{(new)}\boldsymbol{\Lambda}^{(new)}\boldsymbol{\Theta}^{(new)\top}$. Finally, the column vectors of $\boldsymbol{\Theta}^{(new)}$ are reported as the coefficients of the eigenfunctions and the diagonal elements of $\boldsymbol{\Lambda}^{(new)}$ as the variances of the principal components.

4.4.4 Selecting 'smoothing' parameters

Strictly speaking, the three parameters controlling the smoothness of the SS-FPCA model, namely, the degrees of freedom of the basis K, the K-L expansion order P and the Kalman

filter threshold $\chi\%$, are not the actual smoothing parameters as in a penalized regression model. However, the term 'smoothing' parameter is adopted here, with the quotation mark indicating the difference.

Considering the computational burdens of a cross-validation approach or a penalized approach, a method similar to the two-stage approach of the MM-FPCA is proposed to select the degrees of freedom of the basis K and the expansion order P. Specifically, two options can be considered for selecting K.

(a) Choosing K based on the MM-FPCA. In this approach, the degrees of freedom are considered as the level of smoothness associated with the functional data analysis, hence the main focus is on the FPCA component, $\boldsymbol{\Phi\Theta\alpha}_t$. The state space component, $\boldsymbol{\Phi\beta}_t$, though fitted using the same basis, is considered as the counterpart of $\boldsymbol{\Phi\beta}$ in the MM-FPCA and hence its influence is not considered in this selection.

(b) Choosing K based on the SS-FPCA. That is, both the levels of smoothness in the state space and the FPCA components are considered, which appears to be more sensible than method (a). It also introduces the flexibility to the selection when different bases for different model components are considered. However, the computation time of this method is much longer and therefore might not be practical in some situations.

The decision on K can be made using information criteria such as AIC, BIC or their modified versions, which intend to correct the potential bias, such as the conditional AIC (Greven & Kneib, 2010) and the adaptive AIC (Zhang *et al.*, 2012). Alternatively, the choice can be made based on the background of the application, if relevant information is available. The selection of the expansion order P follows the choice of the basis dimension K. Similar to the method used in the MM-FPCA, there are also two approaches to this problem.

(a) Fit a series of models with different expansion orders, then choose the optimal expansion order based on information criteria.

(b) Fit a high rank or full rank model, then choose the expansion order based on the magnitudes of the variances of the PCs, which is similar to the selection based on the percentage of variation explained by the leading PCs.

Note that due to the inclusion of the state space component, part of the variation in the data would be accounted for by this component. What remains to be explained by the FPCA

component is not the total variation in the data. Therefore, the concept of 'percentage of variance explained by the PCs' in the SS-FPCA model is different from that in the MM-FPCA or the PCA. The proportions of variance described by the state space component and the FPCA component thus need to be handled carefully. Further explanation is given in Chapter 5.

Finally, the filtering threshold $\chi\%$ needs to be selected. It has previously been noted that the Kalman filter would sometimes give a very wiggly interpolation of an image, when there are only a few observations scattered around the grid and they happened to be in very different scales. So the filtering threshold is included to impose some restrictions on the filtering of the extremely sparse images in the series. The relative change/ratio of the RSS of the fitted models with different filtering thresholds can be used as a selection criterion. In general, the RSS value would keep on decreasing as the filtering threshold increases, but there may be a point from which the increase of the threshold ceases to make a big difference on the RSS values. Such a point can be taken as the filtering threshold, because a higher threshold is less likely to improve the fit of the data significantly and is more likely to over-interpolate some extremely sparse images.

4.5 Summary

So far, all elements contributing to the development and estimation of the spatio-temporal model, the SS-FPCA, have been introduced. The model provides a way of addressing the three challenges of modelling sparse remote-sensing image time series, dimension reduction, missing data imputation and analysing the spatial/temporal dependence. It also offers an answer to the question raised at the end of Chapter 3 to improve the MM-FPCA by accounting for temporal dependence between remote-sensing images. The SS-FPCA inherits the advantages of both the MM-FPCA and the STRE models and is able to describe the spatio-temporal dependence in a flexible way. The 2-cycle AECM algorithm is designed so that analytical solutions are available for all model parameters. `R` functions have been developed to implement this method. In the next chapter, features of the SS-FPCA and the asymptotic results of the estimation algorithm are investigated, along with two applications using the Lake Victoria data, to complete the 'portrait' of this new model.

Chapter 5

The SS-FPCA model investigation

... there is no need to ask the question 'Is the model true?'. If 'truth' is to be the 'whole truth' the answer must be 'No'. The only question of interest is 'Is the model illuminating and useful?'

George Box (1978)

This chapter carries out the investigation of the proposed SS-FPCA model, including a study of the influence of initial values and model degrees of freedom, a simulation study on the 2-cycle AECM algorithm using 1-dimensional data and an exploratory analysis on the variance components. Specific asymptotic results with respect to the algorithm and model estimates are presented, offering a method to assess the performance of the SS-FPCA. Applications to the Lake Victoria data are given at the end of the chapter.

5.1 Investigation of initial values and 'smoothing' parameters

5.1.1 Model sensitivity with respect to initial values

One criticism of the EM-type algorithm is that the results can be sensitive to the initial values. To check if this is a problem for the SS-FPCA model estimated using the 2-cycle AECM algorithm, an assessment of the sensitivity of the model parameter estimates in terms of the initial values was carried out. In particular, the influences of the initial values of σ_h^2

and σ^2 were assessed. The investigation on σ_h^2 was carried out due to the fact that the initial value of $\boldsymbol{H} = \sigma_h^2 \boldsymbol{I}$ is selected in a somewhat arbitrarily way. The motivation of the investigation on σ^2 was that two different methods are proposed in section 4.4.3 to initialize σ^2, but their impacts are yet to be assessed. The initial values of the rest of the parameters are not examined because their initialization methods are the same as those used in the MM-FPCA, which have already been justified in literature. For this investigation, the 'LSWT section' data set, first introduced at the end of Chapter 1, was used. This is a subset of the Lake Victoria LSWT data set and is of dimension $34 \times 24 \times 202$. As usual, a monthly mean was removed first. A tensor spline basis with degrees of freedom $K = 5 \times 5 = 25$ was used and an orthogonalization was applied to obtain the basis matrix $\boldsymbol{\Phi}$. The order of the K-L expansion was set to be $P = 6$.

To assess the impact of the initial values of $\boldsymbol{H} = \sigma_h^2 \boldsymbol{I}$, a sequence of increasing $\sigma_h^{2(0)}$ values was considered. The lower and upper bounds of the sequence are the 5-th and 95-th quantiles of $\mathbf{Var}[Z_t(x_i, y_i)]$, $i = 1, \cdots, n$, which are 0.114 and 0.231 in this case. A series of 10 values was then take from this interval and the 2-cycle AECM algorithm was run 10 times, each with a different $\sigma_h^{2(0)}$. Under the convergence criterion of $\varepsilon \leq 0.001$, the approximated joint log-likelihood values are within the range of 69741 to 70640. These may not be regarded as converged under the threshold of $\varepsilon \leq 0.001$, but the estimated model parameters show small discrepancies. The estimated $\hat{\sigma}^2$ differs from the third decimal place. The estimated $\hat{\lambda}_p$, $\hat{\boldsymbol{\theta}}_p$, $p = 1, \cdots, 6$, (before final orthogonalization) and the diagonal elements of $\widehat{\boldsymbol{H}}$ are also similarly close. The results are plotted in Figure 5.1 and 5.2. In some panels, the 10 curves representing the estimates from 10 runs are almost identical. It suggests that the parameter estimates from the 2-cycle AECM algorithm with different initial values of $\sigma_h^{2(0)}$ are relatively robust, i.e. $\sigma_h^{2(0)}$ does not have a substantial influence on model fitting.

To assess the impact of two different versions of $\sigma^{2(0)}$ introduced in section 4.4.3, two runs of the AECM algorithms were processed. The first one used $\sigma^{2(0)} = \frac{1}{\sum_t n_t} \sum_{t=1}^{T} \hat{\boldsymbol{r}}_t^\top \hat{\boldsymbol{r}}_t$, denoted as the FPCA version; the second one used $\sigma^{2(0)} = \mathbf{Var}[Z_t(x, y)]$, denoted as the data version. In this case, the initial value from the FPCA version is $\sigma^{2(0)} = 0.0862$ and that from the data version is $\sigma^{2(0)} = 0.1725$. The log-likelihood values and the MLEs of the parameters after convergence ($\varepsilon \leq 0.001$) are shown in Table 5.1. The two log-likelihood values differ by a factor smaller than the convergence threshold 0.001; the discrepancy in the estimated $\hat{\sigma}^2$, $\hat{\lambda}_p$ and $\hat{\boldsymbol{\theta}}_p$ (not presented in the table), $p = 1, \cdots, 6$, are also negligible. Therefore, it may be concluded that despite the difference in initializing method, the proposed 2-cycle AECM algorithm is capable of generating robust estimates of the model parameters.

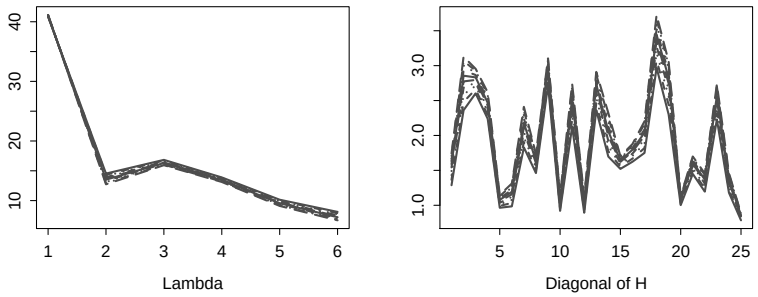

FIGURE 5.1: (Left) The MLEs of $\hat{\lambda}_p$, $p = 1, \cdots, 6$, from 10 runs of the AECM algorithm shown as 10 curves. (Right) The diagonal elements of the MLEs of $\widehat{\boldsymbol{H}}$, $\hat{h}_k$, $k = 1, \cdots, 25$ from 10 runs of the algorithm shown as 10 curves. The indexes on the horizontal axis of the left panel are $p = 1, \cdots, 6$ and those of the right panel are $k = 1, \cdots, 25$.

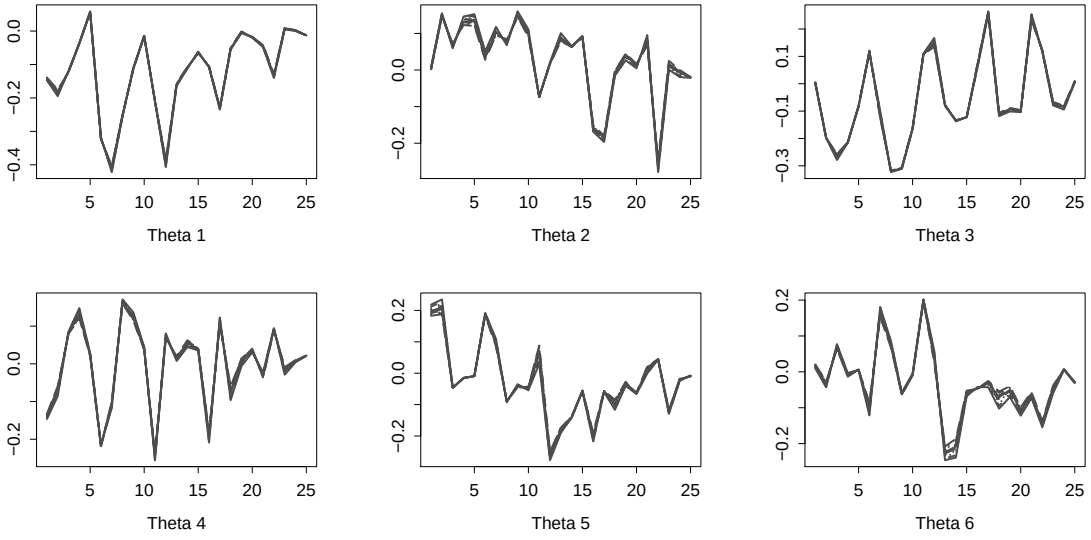

FIGURE 5.2: (Left to right, top to bottom) The MLEs of the basis coefficient vectors $\hat{\boldsymbol{\theta}}_1, \cdots, \hat{\boldsymbol{\theta}}_6$ from the 10 runs of the AECM algorithm shown as 10 curves. The horizontal axis represents the index of the 25 elements in vector $\hat{\boldsymbol{\theta}}_p$.

TABLE 5.1: The log-likelihood and MLEs of σ^2 and λ_p, $p = 1, \cdots, 6$, from 2 runs of the AECM algorithm with different initial values of σ^2.

	loglike	$\hat{\sigma}^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$	$\hat{\lambda}_6$
FPCA version	70174	0.0822	31.24	13.63	4.90	2.00	1.74	1.47
Data version	70167	0.0822	31.30	13.65	4.94	2.03	1.76	1.50

5.1.2 Basis dimension, expansion order and filtering threshold

Choosing the appropriate level of smoothness is an important task for almost all functional data analysis problems. The selection methods, though not a main theme of this thesis, are worth considering. The approaches to the selection of basis dimension K, expansion order P and filtering threshold $\chi\%$ introduced in section 4.4.4 are relatively easy to implement. However, as discussed in Chapter 3 for the MM-FPCA, the process and the criteria should not be treated too rigidly. The final decision might benefit from other relevant information, such as the scientific background of the application, practical knowledge and the purpose of the analysis. To illustrate this point, an example of selecting the 'smoothing' parameters of the SS-FPCA model applied to the 'LSWT section' data set is presented here.

The selection of the degrees of freedom of the basis K was investigated initially. Both methods based on the MM-FPCA and the SS-FPCA were implemented. The number of knots considered in this study ranges from 1 to 4 knots each along the longitude and latitude coordinates. This, combined with a spline basis of order 4, gives a sequence of the candidate bases with degrees of freedom from 5×5 to 8×8. The knots are placed evenly along the coordinates. Every time, one knot was added to one dimension (longitude first, then latitude). The initial expansion order used at this stage was $P = 15$ and the filtering threshold was not considered in this investigation. The log-likelihood, AIC and BIC values from the selection using the MM-FPCA are shown in Figure 5.3; those from the process using SS-FPCA model are shown in Figure 5.4. Note that the index on the horizontal axis represents the index of the basis with dimensions from 5×5 to 8×8[i]. Chances are the information may not always be clear based on different criteria. Therefore, it is better not to make the decision based one single measure. Rather, several different aspects of the model need to be considered, such as the application background and the modelling purpose. For example, since dimension reduction is a priority in this analysis, the 6×6 basis, as suggested by the BIC plot in the right panel of Figure 5.3, may be an appropriate choice.

Next, the investigation of the selection of P was carried out using the 6×6 basis selected from the previous stage. Both the information criteria approach and the variance proportion approach were applied. For the selection using information criteria, a sequence of expansion order ranging from 2 to 10 was considered. Figure 5.5 presents the log-likelihood, AIC and BIC values from fitting the SS-FPCA model using $P = 2, \cdots, 10$. For the selection using

[i]The index system is: 1 for the 5×5 basis, 2 for the 6×5 basis, 3 for the 5×6 basis, 4 for the 6×6 basis, so on and so forth until 10 for the 8×8 basis

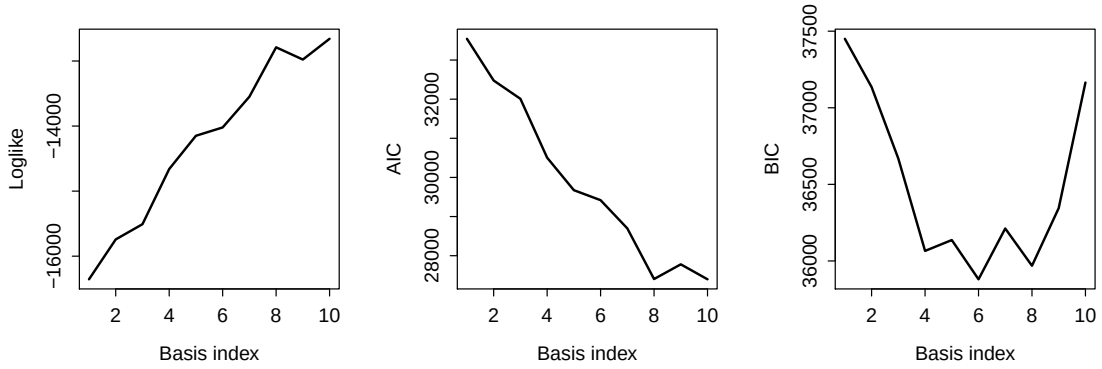

FIGURE 5.3: An example of selecting basis dimension using the MM-FPCA. The three panels are the log-likelihood (left), AIC (middle) and BIC (right) against the index of basis of increasing degrees of freedom, from 5×5 to 8×8.

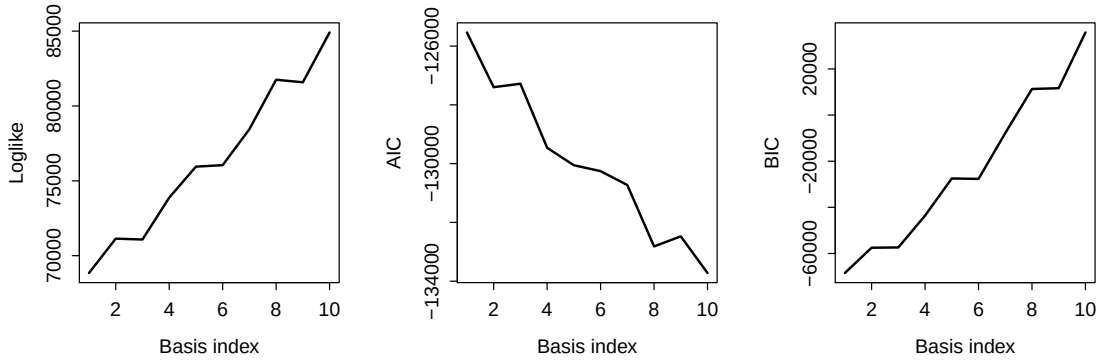

FIGURE 5.4: An example of selecting basis dimension using the SS-FPCA. The three panels are the log-likelihood (left), AIC (middle) and BIC (right) against the index of basis of increasing degrees of freedom, from 5×5 to 8×8.

variance proportion criteria, a high rank SS-FPCA model with $P = 20$ was fitted first. Then the accumulated variances were computed and an appropriate value of P was chosen based on a specified threshold $\delta\%$. In this illustration, the results in Figure 5.5 did not provide much information in terms of the choice of the optimal expansion order, as a boundary solution was not necessarily appropriate. As for the variance proportion criterion, a 80% threshold would suggest $P = 6$ and a 90% threshold would indicate $P = 9$. Similar to the selection of the basis dimension, the final decision need to be made by considering the test results and other related metrics. In this case, $P = 6$ could be a reasonable choice.

Sometimes, there may be an interval within which all choices seem to make sense. In these situations, a decision could be made based on how sensitive the final results are to the change of P in such an interval. If the differences are not substantial, then a parsimonious choice could be appropriate. As an illustration, two SS-FPCA models using the same 6×6 basis,

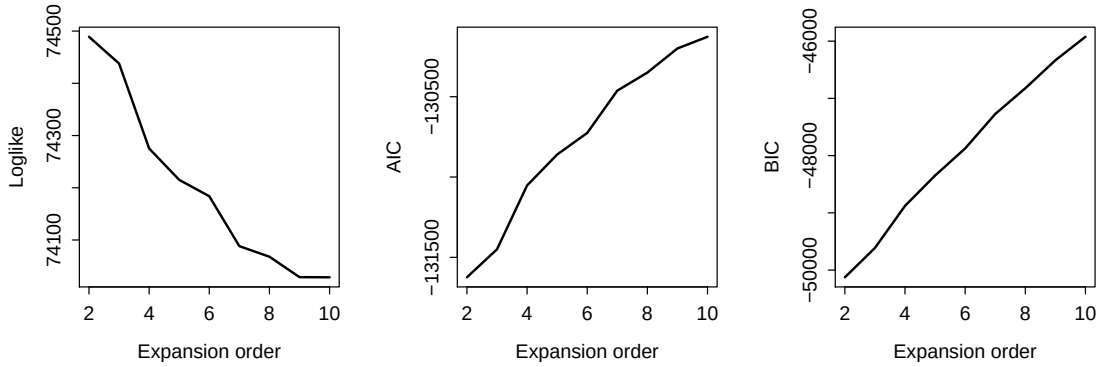

FIGURE 5.5: An example of selecting the K-L expansion order through fitting a series of SS-FPCA models with increasing expansion order P. The three panels are the log-likelihood (left), AIC (middle) and BIC (right) against $P = 2, \cdots, 10$.

but different expansion orders, one with $P = 6$ (for $\geq 80\%$ variance) and the other with $P = 9$ (for $\geq 90\%$ variance), were fitted. The estimated $\hat{\sigma}^2$, λ_p, $p = 1, \cdots, 6$, and the RSS values were recorded in Table 5.2. The majority of the estimates from the two models appear to be similar, suggesting a minor change moving from a lower rank model to a higher rank model. In addition, the log-likelihood value from the $P = 6$ model is larger (when using the same convergence criterion), making it a better choice, especially when dimension reduction and data imputation are among the priorities of this analysis.

TABLE 5.2: The sensitivities of final results to the change of expansion orders of the SS-FPCA models fitted to the 'LSWT section' data set.

	loglike	$\hat{\sigma}^2$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$	$\hat{\lambda}_6$
P = 6	79750	0.0709	34.84	13.23	5.88	2.23	1.02	0.66
P = 9	79504	0.0730	34.29	14.21	6.54	3.67	2.06	1.43

Finally, the selection of the filtering threshold $\chi\%$ was processed, through testing a sequence of increasing filtering thresholds. For an illustration, a series of thresholds, 90%, 92.5%, 95%, 97.5% and 100%, were used in the selection. The log-likelihood and the RSS of the models using five different thresholds were recorded. Results were plotted in two panels of Figure 5.6. Both plots appear to have a 'jump' from the $\chi\% = 92.5\%$ to the $\chi\% = 95\%$; whereas the curves before and after these two points are relatively flat. Therefore, it seems to be appropriate to take $\chi\% = 95\%$ in this case.

The above illustration highlighted a potential issue of the selection procedure proposed in section 4.4.3. Sometimes, there might not be straightforward choices of the parameters K and P from the two-stage method, or of $\chi\%$ from the test of a sequence of increasing

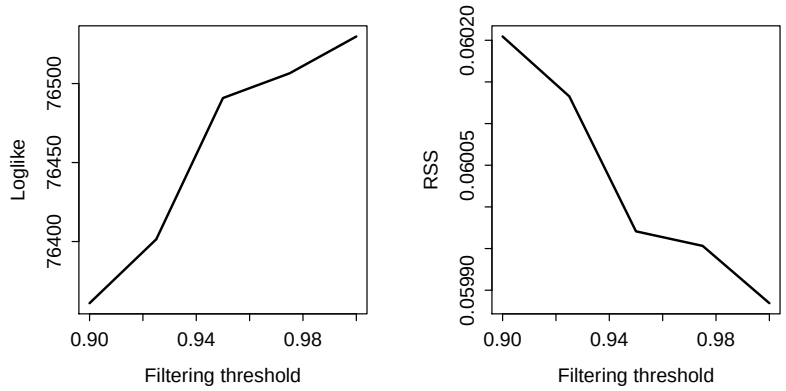

FIGURE 5.6: An example of selecting the filtering threshold $\chi\%$ through fitting a series of SS-FPCA models with increasing thresholds. The two panels are the log-likelihood (left) and the RSS (right) against $\chi\% = 90\%, 92.5\%, 95\%, 97.5\%$ and 100%.

thresholds. However, a justifiable decision can usually be reached with the consideration of the application background and/or the priorities of the problem under study. Alternatively, an investigation on the sensitivity of the model results to different 'smoothing' parameters may be used to select values from candidate intervals.

5.2 Investigation on the performance of the SS-FPCA

This section explores several aspects of the SS-FPCA model to see whether this complex model is capable of achieving what it is intended to do. A simulation study and a comparison of the SS-FPCA to other two models were carried out to asses the model performance. An investigation on the variances of the model components was also conducted to highlight some interesting features of the model.

5.2.1 Simulation study on 1-dimensional data

The purpose of this simulation study is to test if the SS-FPCA model estimated using the 2-cycle AECM algorithm can identify the temporal and spatial structure in the data. For computational efficiency and ease of interpretation, this simulation study was conducted on 1-dimensional data. This, though different from the application, would not result in significant loss of generality as the model assumptions and the estimation method remain the same. A similar investigation on a STRE model of the form (4.25) has been carried out in Katzfuss & Cressie (2011), where data defined on a 1-dimensional space was used in their simulation

study. Simulation studies on models equivalent to the MM-FPCA using 1-dimensional data can be found in Gervini (2009), James *et al.* (2000), Peng & Paul (2009).

Part 1: simulation design The SS-FPCA for 1-dimensional data can be written as

$$Z_t(x) = \Phi(x)\boldsymbol{\beta}_t + \sum_{p=1}^{P} \Phi(x)\boldsymbol{\theta}_p \alpha_{tp} + \epsilon_t(x) \tag{5.1}$$
$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{u}_t \,,$$

with model assumptions

$$\begin{aligned}
&\boldsymbol{\alpha}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}), \quad \boldsymbol{\Lambda} = \text{diag}\{\lambda_1, \cdots, \lambda_P\} \\
&\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}) \\
&\boldsymbol{u}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{H}), \quad \boldsymbol{H} = \text{diag}\{h_1, \cdots, h_K\} \,.
\end{aligned}$$

It is also assumed that functions $\xi_p(x) = \Phi(x)\boldsymbol{\theta}_p$, $p = 1, \cdots, P$, are orthonormal eigenfunctions, satisfying

$$\int \xi_p(x)\xi_q(x)\, dx = \begin{cases} 1, & p = q \\ 0, & p \neq q \end{cases} \,.$$

Model (5.1) is used as the data generating function in this simulation. Specifically, the FPCA component $\sum_{p=1}^{P} \Phi(x)\boldsymbol{\theta}_p \alpha_{pt}$ has $P = 2$ and functions $\xi_p(x) = \Phi(x)\boldsymbol{\theta}_p$ are generated using the eigenfunctions from the MM-FPCA applied to a subset of the Lake Victoria Chlorophyll data. This subset is defined on a stripe in the Chl image, with longitude fixed at 33.0957°E and latitude stretching from 1.1521°S to 1.5661°S. The Chl data are used here because its missing percentage is substantially smaller than the LSWT data, which helps to provide a better estimation of the eigenfuntions.

A few key aspects of the simulation design are listed below.

- The dimension of the simulated data is 50×100, where $n = 50$ is the number of observations (indexed by i) at each time point in the 1-dimensional space $\mathcal{D} = [1.1521, 1.5661]$ and $T = 100$ is the total number of time points (indexed by t). The function argument x represents the spatial location in the 1-dimensional space. This means, the data are $Z_t(x_i)$, $x_i \in \mathcal{D}$, for $i = 1, \cdots, 50$ and $t = 1, \cdots, 100$.

- A cubic B-spline basis with 3 equally spaced interior knots is used as $\Phi(x)$. This gives the basis dimension of $K = 3 + 3 + 1 = 7$.

- The basis coefficient vector series $\{\boldsymbol{\beta}_t = (\beta_{1t} \cdots \beta_{Kt})^\top\}_{t=1}^T$ are generated using $K = 7$ random walk processes $\{u_{kt}\}_{t=1}^T$, $k = 1, \cdots, 7$, each with distribution $u_{kt} \sim \mathcal{N}(0, h_k)$, and a random zero mean starting point. In this case

$$\{h_1, \cdots, h_7\} = \{0.33,\ 0.25,\ 0.42,\ 0.25,\ 0.27,\ 0.62,\ 0.28\}\ .$$

  This gives the dynamic component (using matrix notation) $\boldsymbol{Z}_t^{(d)} = \boldsymbol{\Phi}\boldsymbol{\beta}_t$.

- To obtain the FPCA component, first use eigenfunctions, $\xi_1(x)$, $\xi_2(x)$ and eigenvalues, $\lambda_1 = 9.64$, $\lambda_2 = 1.80$, from the MM-FPCA applied to the subset of the Lake Victoria Chl data, to obtain an approximation of the data covariance matrix $\boldsymbol{\Sigma}_{chl}$. Then generate a sequence of random realizations $\boldsymbol{Y}_t = (Y_{t1}, \cdots, Y_{tn})^\top$ from the $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ distribution and transform them to get $\boldsymbol{Z}_t^{(s)} = \boldsymbol{\Sigma}_{chl}^{\frac{1}{2}} \boldsymbol{Y}_t$. Finally, multiply $\boldsymbol{Z}_t^{(s)}$ with a factor κ ($\kappa \geq 1$) to control the strength of the spatial signal. In this simulation study, $\kappa = 1.25$ and $\kappa = 1.5$ are considered.

- The residual component $\boldsymbol{\epsilon}_t$ is generated from the normal distribution $\mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. In this simulation study $\sigma^2 = 0.01$ and $\sigma^2 = 0.25$ are considered.

- The dynamic, FPCA and residual components are then combined to obtain the simulated data as $\boldsymbol{Z}_t = \boldsymbol{Z}_t^{(d)} + \kappa\boldsymbol{Z}_t^{(s)} + \boldsymbol{\epsilon}_t$.

Note that the square root of the inverse $\boldsymbol{\Sigma}_{chl}^{\frac{1}{2}}$ is obtained through a singular value decomposition of the matrix $\boldsymbol{\Xi}\boldsymbol{\Lambda}^{\frac{1}{2}}$, where $\boldsymbol{\Xi}$ is the matrix of the eigenfunctions and $\boldsymbol{\Lambda}^{\frac{1}{2}}$ is the square root of the eigenvalue matrix $\boldsymbol{\Lambda}$. In addition, to mimic the sparsity in the real life data, the following procedure is used to create the missing patterns.

- First generate a series of 100 missing proportions p_t, $t = 1, \cdots, 100$, from the uniform $\mathcal{U}(0, 1)$ distribution

- Then for each t, generate 50 binomial random variables from distribution $\mathcal{B}(1, p_t)$. Assign the locations corresponding to 1 with observations and regard the locations corresponding to 0 as missing points.

Two factors of interest in this simulation study are the strength of the spatial signal and the initialization method of the 2-cycle AECM. The first factor consists of two levels, weak signal

$\kappa = 1.25$ and strong signal $\kappa = 1.5$. The second factor considers two initialization methods, the standard method as described in section 4.4.3 and a separate methods which uses $\boldsymbol{Z}_t^{(d)} + \boldsymbol{\epsilon}_t$ to initialize the dynamic component and $\boldsymbol{Z}_t^{(s)} + \boldsymbol{\epsilon}_t$ to initialize the FPCA component. The separate method is supposed to provide initial values with higher precision. These two factors are paired to create three combinations, weak + standard, weak + separate and strong + standard. For each combination, four different situations based on noise levels (small or large) and missing conditions (complete or missing) are created, generating 12 scenarios in total. The diagram in Figure 5.7 presents the factors and the 12 scenarios, denoted as S1 to S12 respectively. An example of simulated data with weak spatial signal $\kappa = 1.25$, small noise $\sigma^2 = 0.01$ and missing observations is given in Figure 5.8.

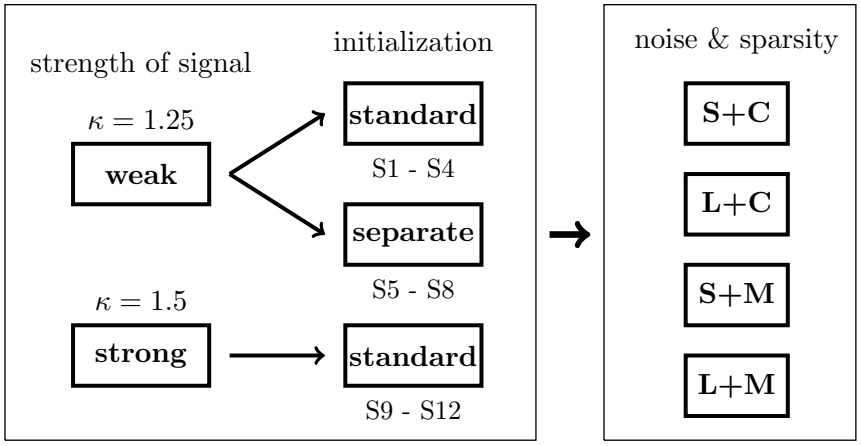

FIGURE 5.7: A diagram showing the settings of 12 simulation scenarios. The abbreviations used in the diagram are S for small noise, L for large noise, C for complete data and M for missing data.

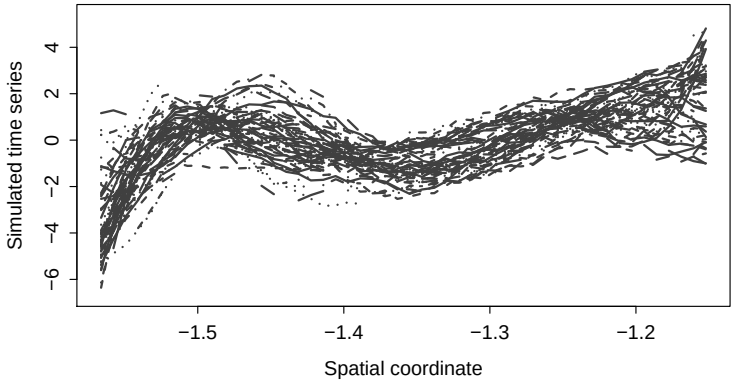

FIGURE 5.8: The time series of curves from one simulated replicate. These curves are generated using weak spatial signal $\kappa = 1.25$, small noise $\sigma^2 = 0.01$ with missing patterns. Each curve consists of observations at one time point.

Part 2: simulation results 500 replicates were run for each scenario. The computation times varies, depending on the number of iterations involved (from $\leq$ 10 to 500) in each replicate. On average, one iteration took 0.25 - 0.5 seconds. Some details of the simulation results are presented below.

First of all, the fitted models are capable of recovering the patterns of the dynamic (i.e. state space) and the FPCA components. The top three panels in Figure 5.9 represent the estimated first eigenfunction $\xi_1(x)$ from scenarios S4, S8 and S12, with the true eigenfunction plotted as the red curve. S4, S8 and S12 represent three simulation scenarios labeled as weak + standard, weak + separate, strong + standard, each paired with large noises and missing observations. Clearly, the pattern in $\xi_1(x)$ is very well identified. All 500 replicates produced curves bearing the feature of the true eigenfunction. The situation with the second eigenfunction is slightly worse, with occasional miss of the target (see bottom three panels in Figure 5.9). However, considering that the first PC is dominant and the magnitude of the variance of the second PC is less than one-fifth of the variance of the first PC, it is not surprising that the pattern in $\xi_2(x)$ is harder to capture.

The patterns in the time series of the coefficient vector were also captured by the smoothed series $\{\boldsymbol{\beta}_{t|T}\}_{t=1}^{T}$. Figure 5.10 gives an example of the smoothed series of each component of $\boldsymbol{\beta}_t$, β_{kt}, $k = 1, \cdots, 7$, taken from scenario S1 (the weak + standard, paired with small noises and complete data scenario). It is straightforward to see that the smoothed series (grey curves) track the true simulated series (red curves) in the majority of the cases. As the plots are produced using the same scale, the variation of the estimates of each component can be compared easily. There seems to be a relatively large difference in the variations, with the 3rd component having the largest variation and the 7th component varying the least. This result could be attributed to the feature of the data. One explanation would be that the variation is larger in the range of support of the third basis function $\phi_3(x)$ in $\Phi(x) = (\phi_1(x), \cdots, \phi_7(x))^\top$. In terms of the 12 simulation scenarios, the variations of the estimates appear to be determined mainly by noise levels; the influence from sparsity and initialization method seems to be much smaller.

However, the estimation of the variances of three model components appears to be more difficult. The boxplots in Figure 5.11 show the distribution of the estimated eigenvalues λ_1 and λ_2 (subject to the signal adjusting factor κ) from three scenarios, S4, S8 and S12, with the red dots indicating the true values. These boxplots show an underestimation of λ_1 and λ_2, especially for λ_2. The increase in the strength of the spatial signal and the more precise

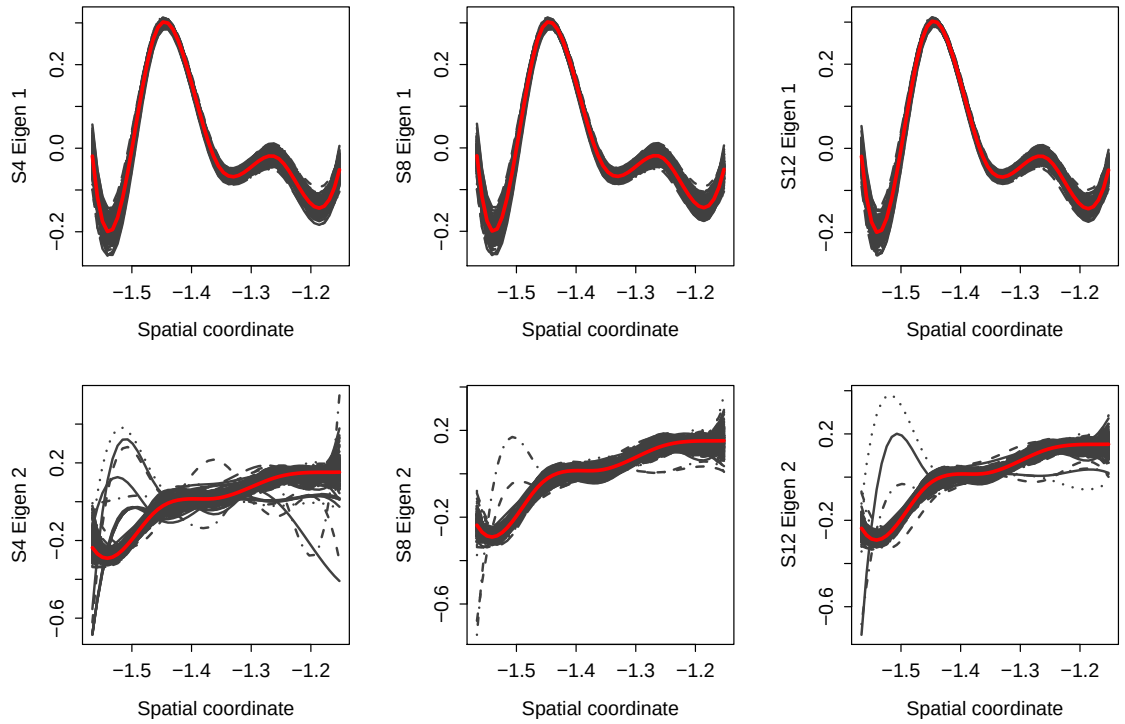

FIGURE 5.9: (Top) The estimated (black curves) and the true (red curve) values of eigenfunction of PC1, from scenario S4, S8 and S12. (Bottom) The estimated (black curves) and the true (red curve) values of eigenfunction of PC2, from scenarios S4, S8 and S12.

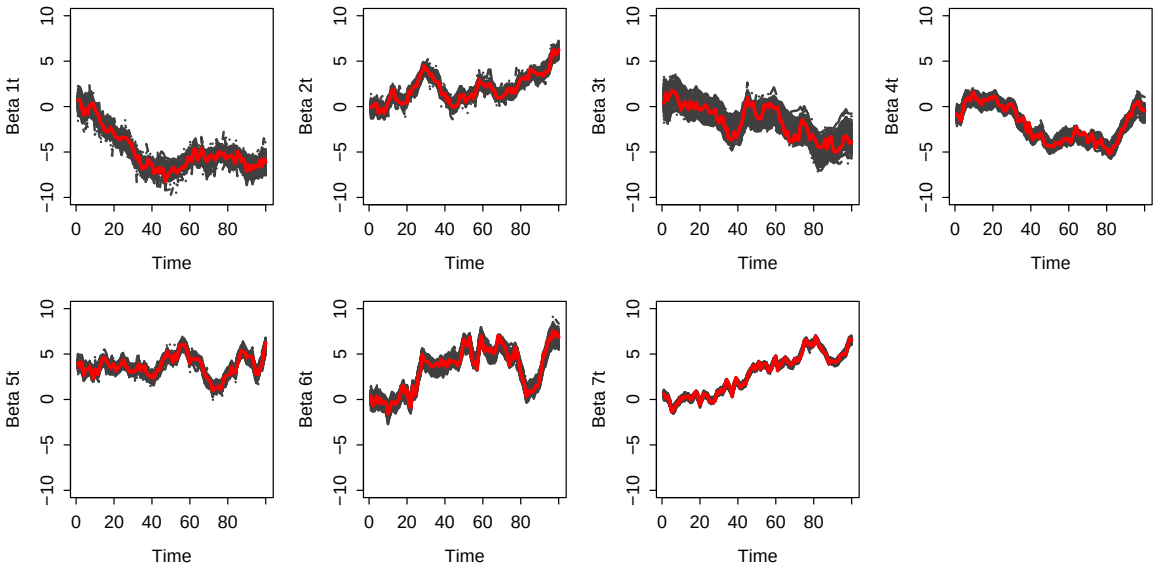

FIGURE 5.10: The Kalman smoothed $\{\boldsymbol{\beta}_{t|T}\}_{t=1}^{T}$ from scenario S1. From left to right, top to bottom are the smoothed $\beta_{kt|T}$, $k = 1, \cdots, 7$ curves (black) and the true curves (red).

separate initialization method did not seem to improve the estimation, which can be seen from the middle and right panels.

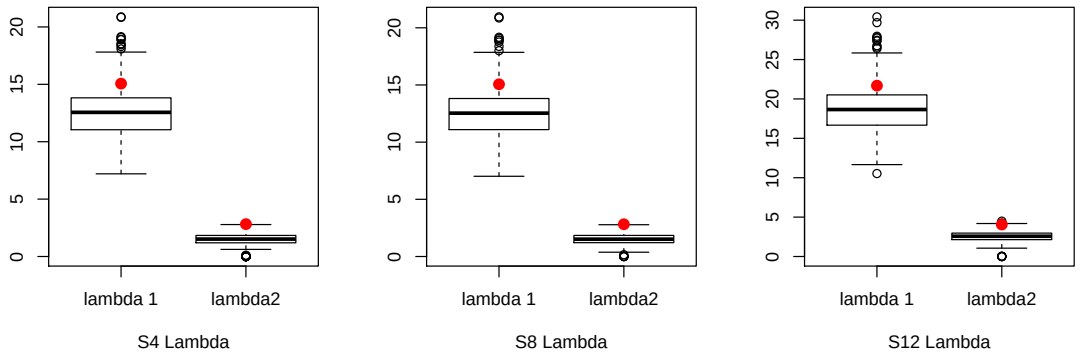

FIGURE 5.11: The boxplots of the estimated eigenvalues λ_1 and λ_2 from 500 repetitions, with the red dots representing the true values of the two eigenvalues. The three panels represent scenario S4, S8 and S12 respectively.

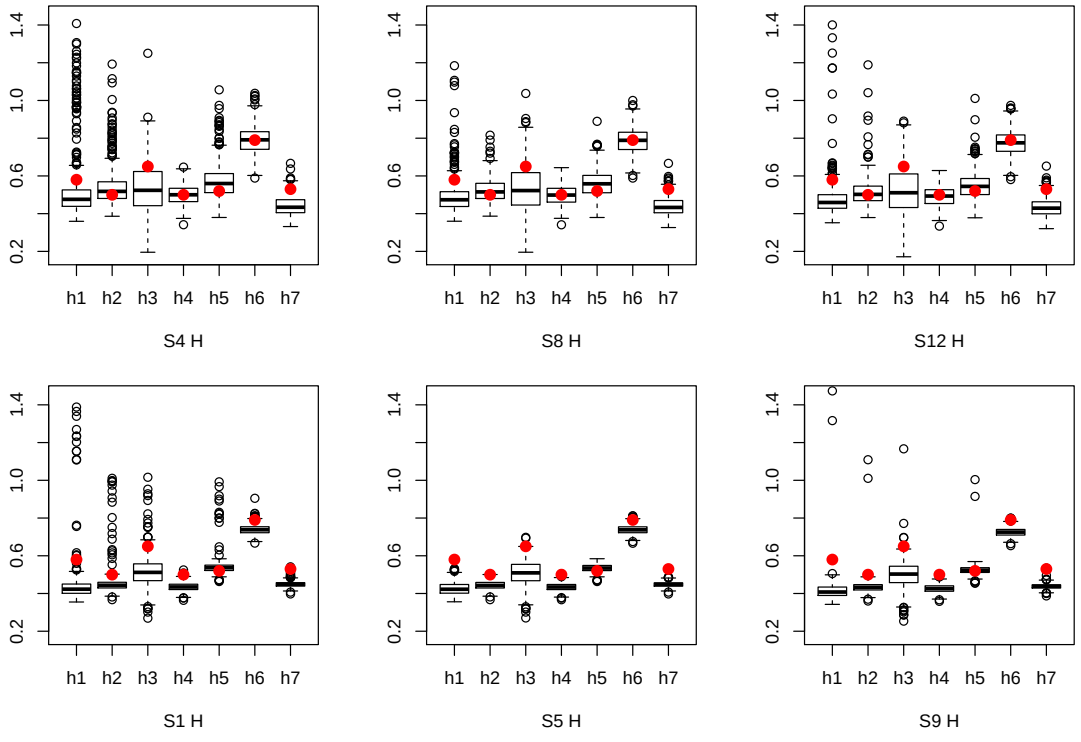

FIGURE 5.12: The boxplots of the estimated $h_1, \cdots, h_7$ from scenario S4, S8 and S12 (top) and S1, S5 and S9 (bottom), with the red dots representing the true values.

Figure 5.12 presents the boxplots of the estimated h_k, $k = 1, \cdots, 7$, from scenarios S4, S8, S12 (top panels) and S1, S5, S9 (bottom panels). S1, S5 and S9 represent the scenarios labeled as weak + standard, weak + separate, strong + standard, all paired with small noises and complete data. The boxplots show that the patterns in h_k, $k = 1, \cdots, 7$, were captured by

the model. Yet the fitted models appear to underestimate some of the h_k. Again, making the spatial signal stronger did not help to distinguish the temporal signal, neither did the separate initialization method. However, the separate initialization method seems to have the potential of avoiding extreme estimates, as highlighted by the bottom middle panel.

As for the residual variances, there appears to be an overestimation of σ^2 in all 12 scenarios (see Table 5.3). The true values are not included in the 95% confidence intervals estimated from the 500 replicates. Nevertheless, the RSS values of the reconstructions are small and they seem to be consistent with the true variance of the residuals. In addition, introducing sparsity to the data did not make a big difference in the RSS values.

TABLE 5.3: The means and the 95% confidence intervals of the MLEs of σ^2 and the means of RSS (numbers in italics) based on the 500 replicates from 12 simulation scenarios.

		small $\sigma^2 = 0.01$		**large** $\sigma^2 = 0.25$	
		complete	**sparse**	**complete**	**sparse**
weak + standard	MLE	0.0690	0.0708	0.3114	0.3136
	CI	(0.0589, 0.0782)	(0.0600, 0.0808)	(0.2956, 0.3281)	(0.2933, 0.3320)
	RSS	*0.0132*	*0.0139*	*0.2365*	*0.2323*
weak + separate	MLE	0.0689	0.0709	0.3117	0.3140
	CI	(0.0591, 0.0782)	(0.0606, 0.0806)	(0.2965, 0.3281)	(0.2963, 0.3330)
	RSS	*0.0132*	*0.0139*	*0.2366*	*0.2323*
strong + standard	MLE	0.0816	0.0838	0.3250	0.3276
	CI	(0.0700, 0.0928)	(0.0716, 0.0957)	(0.3061, 0.3429)	(0.3063, 0.3488)
	RSS	*0.0136*	*0.0145*	*0.2367*	*0.2322*

To summarize, this simulation study illustrated some interesting properties of the SS-FPCA model estimated using the 2-cycle AECM algorithm. First of all, the study show that the SS-FPCA model is capable of identifying the temporal and spatial patterns in the data. Particularly, two PCs were used in the data generating function and the model identified the eigenfunction of the dominant PC in almost all replicates. The RSS of the model appears to be in a comparable scale to the noise level specified in the corresponding scenario. The SS-FPCA model appears to lack precision in estimating the variance components. Strengthening the spatial signal and changing the initialization method did not improve the results in this case. This can be explained by the identifiability of the model components. As the dynamic component $\boldsymbol{\Phi\beta}_t$ describes a space-time non-separable process, it could be spatially confounded with the FPCA component $\boldsymbol{\Phi\Theta\alpha}_t$, which is a linear combination of the orthogonal spatial patterns. A similar situation has been discussed in section 2.2.3 for the spline regression model with spatially correlated residuals. The confounding problem is common to spatial or spatio-temporal models. Discussion and solutions can be found in Hodges & Reich

(2010), Hughes & Haran (2013), Paciorek (2010), Wakefield (2007), etc. Solutions to the problem of the SS-FPCA model might be more complicated than those proposed for spatial models in existing literature. Nonetheless, it helps to examine the scales and features of the variation of the data. It is also important to bear in mind that the truth is always unknown in practice. Hence any effort made to improve identifiability and any interpretation with respect to this issue need to be carefully considered.

5.2.2 A comparison of three models

In this investigation, the SS-FPCA model was compared to the MM-FPCA introduced in Chapter 3 and the dimension-reduced state space model introduced in Chapter 4. As the SS-FPCA model can be regarded as the combination of these two models, it is straightforward to propose this comparison as an approach to the model investigation. For clarity, the MM-FPCA is referred to as the 'FPCA-only' method; the dimension-reduced state space model is denoted as the 'SS-only' method. The investigation used the 'synthetic section' data set, which is a subset of the reconstructed LSWT data of Lake Victoria, with imposed missing patterns from the 'LSWT section' data set (see Chapter 1). Since the data set provides the 'true values' for the imposed missing observations, RSS can be calculated in both the observed and unobserved areas to examine the fit of the models from different aspects. In particular, three different types of RSS were considered, (a) the RSS of the observed part of the data, denoted as RSS_o, (b) the RSS of the unobserved part of the data, denoted as RSS_u, and (c) the overall RSS, which considers the entire data set, denoted as RSS_a.

The data set was first centered by a monthly mean. The three models were then applied to the centered data, using the same basis and the same K-L expansion order. The degrees of freedom of the basis K and the expansion order P were selected using the methods described in section 4.4.4. With some practical concern on dimension reduction and computational cost, the choice was taken to be $K = 5 \times 6 = 30$ and $P = 5$. A filtering threshold of $\chi\% = 95\%$ was used and the convergence criteria for the EM and AECM algorithm was set to be $\varepsilon \leq 0.0005$. Table 5.4 presents some information and summary statistics from the three fitted models. In this case, the SS-FPCA method took the longest time to estimate due to the complexity of the AECM algorithm. The FPCA-only method is the fastest among the three, as it does not involve the Kalman filter and smoother, which took up the majority of computation time in the AECM iterations. The SS-FPCA model also took the largest number of iterations to converge.

In terms of three different RSS measures, the FPCA-only method produced the largest RSS_o, RSS_u and RSS_a. The SS-FPCA model generated much smaller RSS measures than those from the FPCA-only method, suggesting the improvement as a result of taking into account the temporal structure. The SS-only method outperformed the SS-FPCA model in RSS_o (the RSS from the observed part) by a tiny margin. However, the SS-FPCA model appears to have some advantage in interpolating the unobserved part of the data, which is shown by the values of RSS_u and RSS_a in Table 5.4. This phenomenon can be explained by the Kalman filter/smoother's tendency to over-fit the extremely sparse data as described in section 4.2.2 and 4.2.3. On the contrary, the SS-FPCA model overcomes this problem through the balance of the state space and the FPCA components. This is illustrated in Figure 5.13 by the images of the sparse/complete data and three reconstructed images using the FPCA-only, SS-only and SS-FPCA model respectively. The figure provides an example of over-fitting of the SS-only model (shown as the 4th panel in each row). In practice, when a large amount of data are missing, a relatively smoother interpolation may be preferred.

TABLE 5.4: The comparison of the FPCA-only, SS-only and the SS-FPCA models applied to the 'synthetic section' LSWT data

	Iterations	$\hat{\sigma}^2$	RSS_o	RSS_u	RSS_a
FPCA-only	7	0.0057	0.0030	0.0125	0.0095
SS-only	12	0.0026	0.0015	0.0110	0.0095
SS-FPCA	19	0.0051	0.0016	0.0084	0.0068

To complete the investigation, the estimated state transition equation residual covariance matrix $\widehat{\boldsymbol{H}}$ from the SS-FPCA model and the SS-only method are presented in Figure 5.14. The estimated eigenvalues $\hat{\lambda}_p$ and coefficient vectors $\hat{\boldsymbol{\theta}}_p$ from the SS-FPCA model and the FPCA-only method are given in Table 5.5 and Figure 5.15. In this example, the estimated covariance matrix $\widehat{\boldsymbol{H}}$ in the SS-FPCA and the SS-only models display different patterns; whereas the estimated FPCA components in the SS-FPCA and the FPCA-only models share more similarities. The variances of the functional PCs $\hat{\lambda}_p$ from the SS-FPCA model are smaller than those from the FPCA-only method, because part of the variation in the data has been accounted for by the state space component in the SS-FPCA model.

In general, it is difficult to tell whether the estimated model components are telling the 'truth' of the spatio-temporal structure in the data. The best solution is probably to examine the variation explained by different model components and see if it matches the science behind

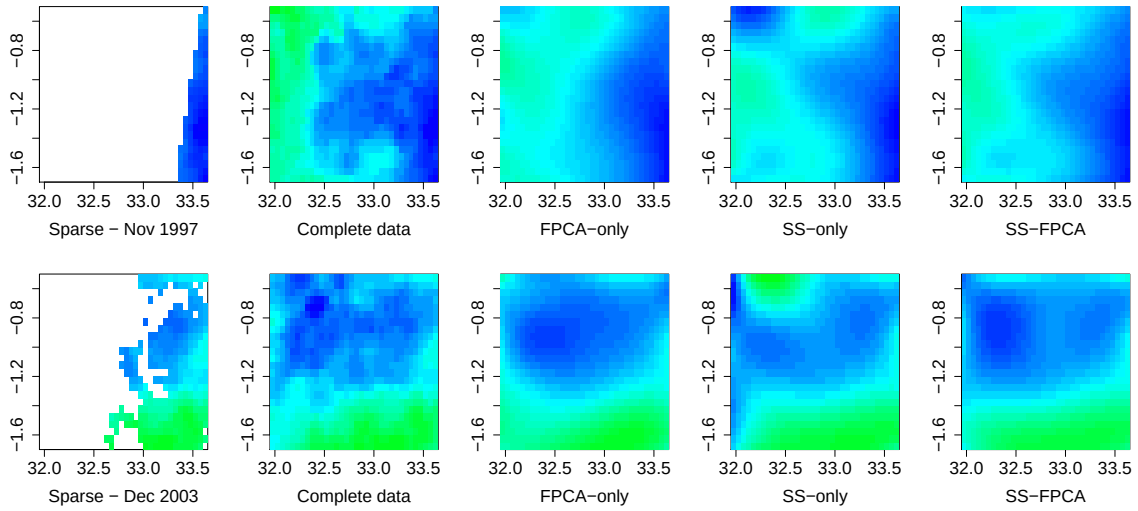

FIGURE 5.13: Examples of the data and the fitted images from November 1997 (top) and December 2003 (bottom), using three different models. In each row, from left to right are the sparse data, the complete data, the reconstructed images from the FPCA-only method, the SS-only method and the SS-FPCA model. The horizontal and vertical axes are latitude and longitude respectively.

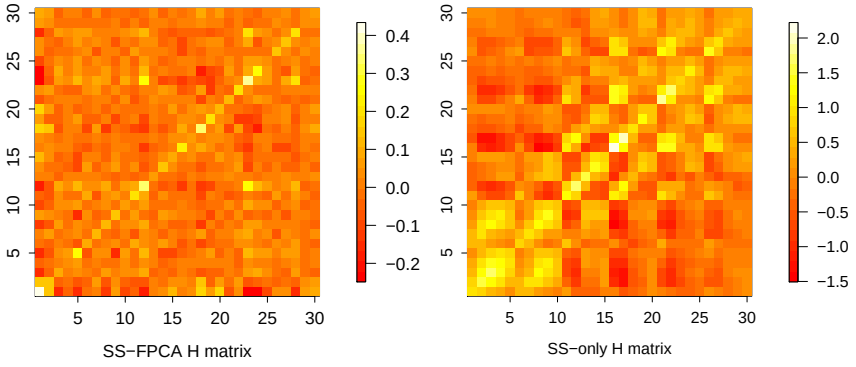

FIGURE 5.14: The estimated $\widehat{\boldsymbol{H}}$ matrix from the SS-FPCA model (left) and the SS-only method (right). The horizontal and vertical axes represent the index of the elements in matrix $\widehat{\boldsymbol{H}}$, $k = 1, \cdots, 30$.

TABLE 5.5: The comparison of the estimated $\hat{\lambda}_p$, $p = 1, \cdots, 5$, from the FPCA-only model and the SS-FPCA model

	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$
FPCA-only	10.1815	4.3488	3.3909	3.0826	1.8347
SS-FPCA	9.3476	3.6101	2.3287	1.5567	0.5149

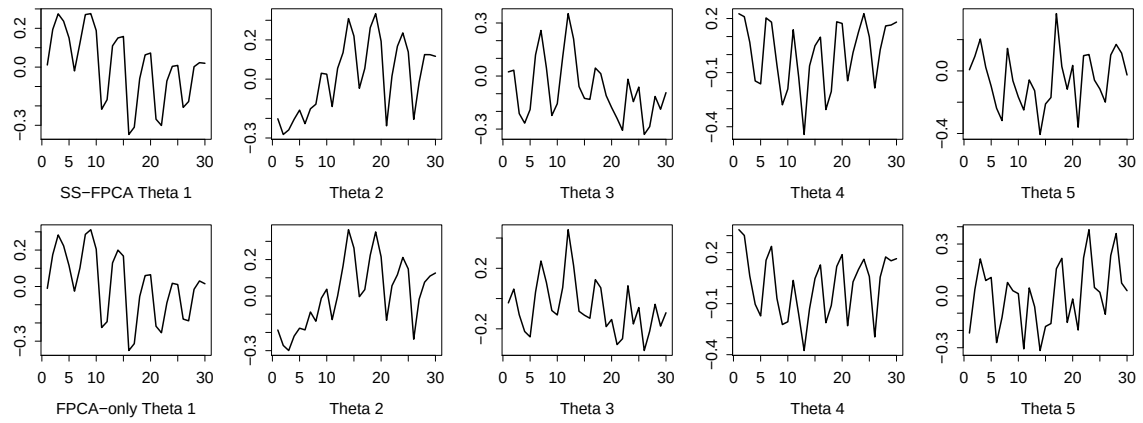

FIGURE 5.15: The estimated $\hat{\boldsymbol{\theta}}_p$ vectors, $p = 1, \cdots, 5$, from the SS-FPCA model (top) and the FPCA-only model (bottom). The horizontal axis shows the index of the elements in vector $\hat{\boldsymbol{\theta}}_p$, $k = 1, \cdots, 30$.

the problem. An investigation to explore the variations of the model components is carried out in the next section.

5.2.3 An investigation of model variance components

In the SS-FPCA model (4.27), the variation of the data $Z_t(x, y)$ comes from three different parts, the dynamic component, the FPCA component and the residual, i.e.

$$\mathbf{Cov}[\boldsymbol{Z}_t] = \boldsymbol{\Phi}_t \mathbf{Cov}[\boldsymbol{\beta}_t] \boldsymbol{\Phi}_t^\top + \boldsymbol{\Phi}_t \boldsymbol{\Theta} \boldsymbol{\Lambda} \boldsymbol{\Theta}^\top \boldsymbol{\Phi}_t^\top + \sigma^2 \boldsymbol{I}\,.$$

In this section, the estimated variance components $\boldsymbol{\Phi} \boldsymbol{B}_{t|T} \boldsymbol{\Phi}^\top$, $\boldsymbol{\Phi} \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\Lambda}} \widehat{\boldsymbol{\Theta}}^\top \boldsymbol{\Phi}^\top$ and $\hat{\sigma}^2 \boldsymbol{I}$ are investigated. Specifically, three different residuals are computed. They are the model residuals $\boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}_t^m$, the residuals when only the state space component (the dynamic) is considered $\boldsymbol{\epsilon}_t^d$ and the residuals when only the FPCA component (the spatial pattern) is included $\boldsymbol{\epsilon}_t^s$. The three types of residuals are computed as

$$\begin{aligned} \boldsymbol{\epsilon}_t^m &= \boldsymbol{Z}_t - \left(\boldsymbol{\Phi} \boldsymbol{\beta}_{t|T} + \boldsymbol{\Phi} \widehat{\boldsymbol{\Theta}} \hat{\boldsymbol{\alpha}}_t \right) \\ \boldsymbol{\epsilon}_t^d &= \boldsymbol{Z}_t - \boldsymbol{\Phi} \boldsymbol{\beta}_{t|T} \\ \boldsymbol{\epsilon}_t^s &= \boldsymbol{Z}_t - \boldsymbol{\Phi} \widehat{\boldsymbol{\Theta}} \hat{\boldsymbol{\alpha}}_t\,. \end{aligned} \tag{5.2}$$

Their corresponding RSS can be regarded as a measure of the contribution of the two model components, $\boldsymbol{\Phi} \boldsymbol{\beta}_t$ and $\boldsymbol{\Phi} \boldsymbol{\Theta} \boldsymbol{\alpha}_t$, to the fit of the model. Alternatively, the diagonal elements of

$\boldsymbol{\Phi}\boldsymbol{B}_{t|T}\boldsymbol{\Phi}^\top$ and $\boldsymbol{\Phi}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Theta}}^\top\boldsymbol{\Phi}^\top$ and the estimated residual variance $\hat{\sigma}^2$ can be used to evaluate the contribution of different components to the total variation.

The investigation used the 'Chl section' data set. It is a subset of the spatially aggregated (using the average of the values in a 3×3 grid) Lake Victoria Chlorophyll data and is of dimension $36 \times 36 \times 119$. The Chl data were used here due to their different features from the LSWT data in both space and time. By applying two centering methods, one using a monthly mean and the other using a simple overall mean, two data sets with very different total variations can be created. The SS-FPCA model was fitted to both data sets, using a basis of dimension $K = 5 \times 5 = 25$ and expansion orders $P = 6$ and $P = 2$ respectively (both explaining 95% of the variance of the FPCA component). The filtering threshold was set to 95% and the convergence criterion was $\varepsilon \leq 0.0005$.

TABLE 5.6: The comparison of the RSS computed from three different types of residuals, $\boldsymbol{\epsilon}^m$, $\boldsymbol{\epsilon}^d$ and $\boldsymbol{\epsilon}^s$, from the SS-FPCA model applied to the 'Chl section' data set with different centering methods.

	Variation in data	$\hat{\sigma}^2$	RSS of $\boldsymbol{\epsilon}^m$	RSS of $\boldsymbol{\epsilon}^d$	RSS of $\boldsymbol{\epsilon}^s$
Monthly mean	0.0829	0.0314	0.0226	0.0654	0.0314
Overall mean	0.5733	0.0483	0.0228	0.4603	0.0706

Table 5.6 shows the total variation of the data, the estimated residual variance and the three types of RSS from the fitted models. Despite the difference in the total variation, the discrepancy in the estimated residual variance $\hat{\sigma}^2$ between two models is relatively small, so are the RSS of $\boldsymbol{\epsilon}^m$ and $\boldsymbol{\epsilon}^s$. However, a large gap is found in the RSS of $\boldsymbol{\epsilon}^d$, which are the residuals of the imputation using only the state space component. In the model centered by a monthly mean, the RSS of $\boldsymbol{\epsilon}^d$ is 0.0654; whereas in the model centered by an overall mean, it is 0.4603 and is about 6 times larger. Given that the RSS of $\boldsymbol{\epsilon}^m$ and $\boldsymbol{\epsilon}^s$ are similar in two models, this suggests that a large amount of the variation was explained by the FPCA component in the model centered by an overall mean. This can also be seen in Figure 5.16, which plots the diagonal elements of $\boldsymbol{\Phi}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Theta}}^\top\boldsymbol{\Phi}^\top$. The left panel, representing the model using a monthly mean, has the majority of the pixels coloured red, corresponding to values at about 0.1. The right panel, representing the model using an overall mean, has the majority of pixels coloured amber and yellow, corresponding to values of 0.5 - 0.6.

As the majority of the variation in the Chl data centered by an overall mean was explained by the FPCA components, the expansion order P may have a relatively large impact on the model. Table 5.7 presents the MLEs of $\hat{\sigma}^2$ and the RSS of different types of model

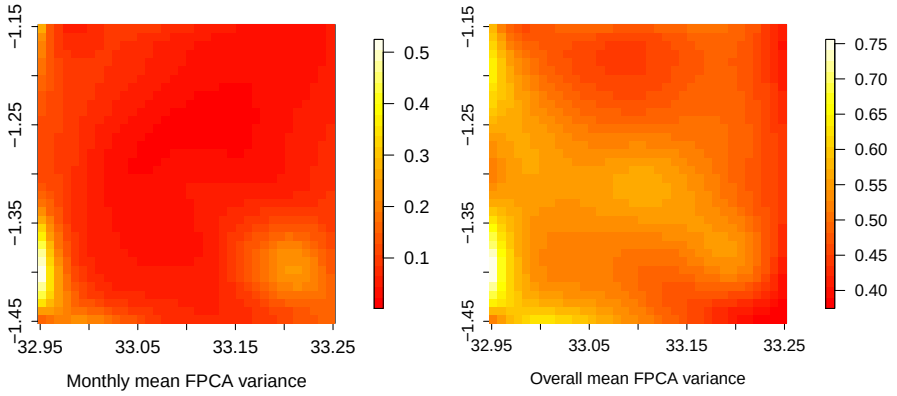

FIGURE 5.16: The diagonal elements in matrix $\boldsymbol{\Phi}\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Theta}}^\top\boldsymbol{\Phi}^\top$ plotted as images, from the SS-FPCA applied to the 'Chl section' data centered by monthly mean (left) and by overall mean (right). The horizontal and vertical axes are latitude and longitude respectively.

residuals from two SS-FPCA models fitted with expansion order $P = 2$ and 4 respectively. The difference in expansion order does appear to introduce discrepancies in $\hat{\sigma}^2$ and the RSS values, especially for $\boldsymbol{\epsilon}^s$. This suggest that the selection of expansion order P in this case may require more attention. On the contrary, if the dynamic component is more influential, then the selection of P may be less crucial.

TABLE 5.7: Example of changing expansion orders and the corresponding RSS computed using different model components of the SS-FPCA model applied to the 'Chl section' data.

	$\hat{\sigma}^2$	RSS of $\boldsymbol{\epsilon}^m$	RSS of $\boldsymbol{\epsilon}^d$	RSS of $\boldsymbol{\epsilon}^s$
$\mathbf{P = 2}$	0.0483	0.0228	0.4603	0.0706
$\mathbf{P = 4}$	0.0525	0.0230	0.4792	0.0504

To summarise, the above investigation provided insight into the characteristics of the state space component and the FPCA components in the SS-FPCA model. The investigation using the 'Chl section' data centered by two different mean functions shows that pre-processing of the data, such as centering and de-trending, can play an important part in the final results. The scales of the variance of different model components may also provide some information on the selection of expansion order.

5.3 Convergence properties of the SS-FPCA

The above investigation may be helpful in providing measures on the general fit of the SS-FPCA model. However, the standard errors of the estimates or the convergence of the

algorithm cannot be obtained from this type of analysis. Fortunately, there are several useful asymptotic properties associated with EM-type algorithms which have been developed over the past 40 years (McLachlan & Krishnan, 1997). These asymptotic results could be used to approximate the standard errors of the MLEs of the SS-FPCA model.

5.3.1 Convergence properties of the AECM algorithm

Given the complexity of the SS-FPCA model, the evaluation of the parameter estimates and the convergence of the algorithm is often difficult. Approaches such as cross validation, bootstrap and simulation study can be computationally challenging for very large data set. However, through the empirical estimates of the theoretical convergence measures, indications of the asymptotic behaviors of the estimated model might be extracted.

The monotonic and convergence properties of the EM-type algorithm have been studied in detail ever since the earlier years of the development of the algorithm. Dempster *et al.* (1977) showed that the (incomplete data) likelihood function does not decrease after each EM iteration. McLachlan & Krishnan (1997) explained in their book that, *'in the case where the likelihood function* $\mathcal{L}(\Psi)$ *is unimodal in* $\mathbf{\Omega}$ *(and a certain differentiability condition is satisfied), any EM sequence converges to the unique MLE, irrespective of its starting point* $\Psi^{(0)}$*'* and *'if* $\mathcal{L}(\Psi)$ *has several stationary points, convergence of the EM sequence to either type (local or global maximizers, saddle points) depends on the choice of starting point'*[ii]. The two authors also introduced a series of convergence theorems with respect to a generalized EM (GEM) algorithm, some of which can be traced back to Wu (1983). For extensions, such as ECM, SAGE and AECM, related convergence properties have also been derived and evaluated in various literatures (Fessler & Hero, 1994, Liu & Rubin, 1994, Meng & Rubin, 1993, Meng & Van Dyk, 1997).

To present the convergence of a GEM algorithm (EM, ECM, AECM are all in this family), a set of notation is introduced. These notations follow mainly McLachlan & Krishnan (1997), with a few changes to maintain the consistency with the rest of the thesis.

- Denote $\boldsymbol{Z}$ as the observed data and $\boldsymbol{Z}_c$ as the complete data (or the augmented data), where the subscript c stands for 'complete'.
- Denote Ψ as the set of parameters and $\Psi \in \mathcal{W}$. The observed data log-likelihood is $\mathcal{L}(\Psi)$ and the complete data log-likelihood is $\mathcal{L}_c(\Psi)$. This is slightly different from McLachlan

[ii] The notation $\mathbf{\Omega}$ used in the quotation is the same as $\mathcal{W}$ in this thesis, referring to the parameter space.

& Krishnan (1997) where $\mathcal{L}(.)$ represents the likelihood function. Also define $\Psi^{(it)}$ as the estimated values after the it-th GEM iteration and Ψ^* as the stationary point of the log-likelihood functions $\mathcal{L}(\Psi)$ and $\mathcal{L}_c(\Psi)$.

- The density functions for the observed and the complete data are denoted as $f(\boldsymbol{Z};\Psi)$ and $f_c(\boldsymbol{Z}_c;\Psi)$ respectively. The conditional density of the complete data given the observed data is $f(\boldsymbol{Z}_c|\boldsymbol{Z};\Psi) = f_c(\boldsymbol{Z}_c;\Psi)/f(\boldsymbol{Z};\Psi)$.

- The target function computed in the E-step and used in the M-step is

$$\mathcal{Q}\left(\Psi;\Psi^{(it)}\right) = \mathbf{E}\left[\mathcal{L}_c(\Psi) \,\middle|\, \boldsymbol{Z};\Psi^{(it)}\right] .$$

  A function associated with the conditional density of $f(\boldsymbol{Z}_c|\boldsymbol{Z};\Psi)$ is

$$\mathcal{H}\left(\Psi;\Psi^{(it)}\right) = \mathbf{E}\left[\log f(\boldsymbol{Z}_c|\boldsymbol{Z};\Psi) \,\middle|\, \boldsymbol{Z};\Psi^{(it)}\right] .$$

- The observed/complete data score statistics, which are the first derivative of the observed/complete data log-likelihood with respect to Ψ, are denoted as $F(\boldsymbol{Z};\Psi)$ and $F_c(\boldsymbol{Z}_c;\Psi)$.

- The observed/complete data Fisher information matrices, which are the expectations of the negative second derivatives of the observed/complete data log-likelihood, are denoted as $\mathcal{I}(\Psi,\boldsymbol{Z})$ and $\mathcal{I}(\Psi;\boldsymbol{Z}_c)$.

All the results below are associated with the GEM algorithm, in which the M-step is to choose a value of $\Psi^{(it+1)}$ such that $\mathcal{Q}(\Psi^{(it+1)};\Psi^{(it)}) \geq \mathcal{Q}(\Psi^{(it)};\Psi^{(it)})$. Only a brief summary is presented here; whereas detailed proofs can be found in McLachlan & Krishnan (1997). The first key statement is, the observed data log-likelihood is not decreased after a GEM iteration. To see this, write the observed data log-likelihood as

$$\mathcal{L}(\Psi) = \mathcal{Q}\left(\Psi;\Psi^{(it)}\right) - \mathcal{H}\left(\Psi;\Psi^{(it)}\right) ,$$

which is from the result $f(\boldsymbol{Z};\Psi) = f_c(\boldsymbol{Z}_c;\Psi)/f(\boldsymbol{Z}_c|\boldsymbol{Z};\Psi)$. Then the difference between the log-likelihood from two iterations would be

$$\begin{aligned}
&\mathcal{L}\left(\Psi^{(it+1)}\right) - \mathcal{L}\left(\Psi^{(it)}\right) \\
&= \left\{\mathcal{Q}\left(\Psi;\Psi^{(it+1)}\right) - \mathcal{Q}\left(\Psi;\Psi^{(it)}\right)\right\} - \left\{\mathcal{H}\left(\Psi;\Psi^{(it+1)}\right) - \mathcal{H}\left(\Psi;\Psi^{(it)}\right)\right\} .
\end{aligned}$$

The first curly bracket in the above equation is non-negative by definition of the GEM; the second curly bracket is non-positive by Jensen's inequality (McLachlan & Krishnan, 1997). Hence the observed data log-likelihood is non-decreasing.

Then there is the theorem with respect to the convergence of a GEM sequence to a stationary point. It is based on the regularity conditions of Wu (1983), which are

(a) $\mathcal{W}$ is a subset of the d-dimensional Euclidean space $\mathbb{R}^d$, where d is the dimension of the parameter set;

(b) $\mathcal{W}_{\Psi_0} = \{\Psi \in \mathcal{W} : \mathcal{L}(\Psi) \geq \mathcal{L}(\Psi_0)\}$ is a compact set for any $\mathcal{L}(\Psi_0) > -\infty$;

(c) $\mathcal{L}(\Psi)$ is continuous in $\mathcal{W}$ and differentiable in the interior of $\mathcal{W}$;

(d) each $\Psi^{(it)}$ is in the interior of $\mathcal{W}$, i.e. $\Psi^{(it+1)}$ is the solution of $\partial \mathcal{Q}(\Psi; \Psi^{(it)})/\partial \Psi = 0$.

Theorem 5.1. Let $\{\Psi^{(it)}\}$ be an instance of a GEM algorithm generated by $\Psi^{(it+1)} \in \mathcal{F}(\Psi^{(it)})$. Suppose that the mapping $\mathcal{F}(\Psi^{(it)})$ is closed over the complement of $\mathcal{A}$, which is the set of stationary points in the interior of $\mathcal{W}$, and

$$\mathcal{L}\left(\Psi^{(it+1)}\right) > \mathcal{L}\left(\Psi^{(it)}\right), \ \forall\, \Psi^{(it)} \notin \mathcal{A}\,.$$

Then all the limit points of $\{\Psi^{(it)}\}$ are stationary points and $\mathcal{L}(\Psi^{(it)})$ converges monotonically to $\mathcal{L}^* = \mathcal{L}(\Psi^*)$ for some stationary point $\Psi^* \in \mathcal{A}$ (McLachlan & Krishnan, 1997).

Following the above theorem, the convergence of a GEM sequence of iterates $\{\Psi^{(it)}\}$ to a stationary point Ψ^* can also be established (McLachlan & Krishnan, 1997). The additional conditions required are either

$$\mathcal{A}(\mathcal{L}^*) = \{\Psi^*\}\ ,$$

i.e. set $\mathcal{A}(\mathcal{L}^*)$ consists of single point Ψ^*, or

$$\left\|\Psi^{(it+1)} - \Psi^{(it)}\right\| \to 0, \text{ as } it \to \infty\ ,$$

if set $\mathcal{A}(\mathcal{L}^*)$ consists of multiple elements and they are discrete.

Since there are more than one cycles of the E-step and M-step iteration in the AECM algorithm, an additional condition is required to ensure the convergence of the algorithm. This condition is called 'space-filling' and was first put forward by Meng & Rubin (1993) for

the CM-steps in the ECM algorithm. Its extension to the AECM algorithm was made in Meng & Van Dyk (1997).

The space-filling condition. Let $c = 1, \cdots, C$ be the index of the cycles in the AECM algorithm and $s = 1, \cdots, S_c$ be the index of the CM-steps in the c-th cycle within one iteration. Denote $\{g_s^{[c]}(\Psi),\ s = 1, \cdots, S_c\}$ as a collection of pre-selected constraint functions for the CM-steps in the c-th cycle. The updating criterion in the s-th CM-step of the $(c+1)$-th cycle in the it-th iteration can be written, by omitting the iteration index (it), as

$$\mathcal{Q}^{[c+1]}\left(\Psi^{[c+\frac{s}{S_c}]}; \Psi^{[c]}\right) \geq \mathcal{Q}^{[c+1]}\left(\Psi; \Psi^{[c]}\right) \ , \ \forall\ \Psi \in \mathcal{W}_s\left(\Psi^{[c+1]}\right) \ ,$$

where

$$\mathcal{W}_s\left(\Psi^{[c+1]}\right) \equiv \left\{\Psi \in \mathcal{W} :\ g_s^{[c+1]}(\Psi) = g_s^{[c+1]}\left(\Psi^{[c+\frac{s-1}{S_c}]}\right)\right\}$$

is the parameter space for the s-th CM step of the $(c+1)$-th cycle determined by the constraint function $g_s^{[c+1]}(\Psi)$. Then the space-filling condition can be written as

$$\bigcap_{c=1}^{C}\bigcap_{s=1}^{S_c} \mathcal{G}_s^{[c]}(\Psi) = \{\mathbf{0}\} \ , \tag{5.3}$$

where

$$\mathcal{G}_s^{[c]}(\Psi) \equiv \left\{\bigtriangledown g_s^{[c]}(\Psi)\boldsymbol{a} \ :\ \boldsymbol{a} \in \mathbb{R}^{d_s^{[c]}}\right\}$$

is the column space of the gradient vectors $\bigtriangledown g_s^{[c]}(\Psi)$. The space-filling condition should be fulfilled after all the cycles are completed (Meng & Van Dyk, 1997).

McLachlan & Krishnan (1997) interpreted the space-filling condition for an ECM algorithm in terms of its complement. That is '*the convex hull of all feasible directions determined by the constraint spaces ... is the whole Euclidean space* $\mathbb{R}^d$'. Similar interpretation can be made on the AECM algorithm. In fact, the directions defined in $\mathcal{G}_s^{[c]}(\Psi)$ are directions restricted to the search; whereas its compliment contains the feasible directions. Therefore, the compliments of (5.3), $\bigcup_{c=1}^{C}\bigcup_{s=1}^{S_c}\overline{\mathcal{G}_s^{[c]}(\Psi)} = \mathbb{R}^d \backslash \{\mathbf{0}\}$, inclines that the search of optimal solutions in the CM-steps within different cycles can be carried out in the entire parameter space. Based on this condition, the convergence theorems of the AECM algorithm were established (Meng & Van Dyk, 1997). As they have little difference from those of the GEM algorithm, the theorems are not listed here to avoid redundancy. Readers are referred to Appendix C.1 for more details.

Liu & Rubin (1994) described an example using the partition of the parameter space in the CM-steps to verify the space-filling condition. In that case, it is straightforward to show that $g_s(\Psi) = \{\psi_1, \cdots, \psi_{s-1}, \psi_{s+1}, \cdots, \psi_S\}$ and $\bigcap_{s=1}^{S} \mathcal{G}_s(\Psi) = \{\mathbf{0}\}$. The fact that the CM-steps in the 2-cycle AECM algorithm for the SS-FPCA model also uses a partition of the parameter space suggests that the space-filling condition is supposed to hold.

The results presented in this subsection may be giving the impression of being hard to verify in practice and hence are somewhat redundant. However, in some situations, it helps even just to demonstrate that some essential conditions ensuring the convergence of the algorithm are satisfied, because monitoring convergence can be extremely difficult for a complex EM-type algorithm. Meanwhile, these results offer a way of investigating if the design of the new algorithm (the 2-cycle AECM for the SS-FPCA model in this case) is sensible.

5.3.2 Approximation of the standard errors of the MLEs

Among the various measures of the convergence of the EM-type algorithm, the standard error of the estimated parameter is one of the most interesting in application. Unfortunately, there is no direct solution from the implementation of an EM-type algorithm. It is possible to bootstrap the standard errors, but the computation can be challenging for a large data set or a complicated model. Alternatively, the inverse of the observed information matrix $\mathcal{I}(\Psi)^{-1}$ can be used as an approximation to the standard errors of Ψ. To obtain the observed information matrix, the expectation of the observed log-likelihood $\mathcal{L}(\Psi)$ is required. This is no easy task and is the reason why the EM-type algorithm is used in the first place. As a result, approximations to $\mathcal{I}(\Psi)^{-1}$ need to be considered.

The reason the observed information matrix $\mathcal{I}(\Psi)$ is favoured over the complete data Fisher information $\mathcal{I}_c(\Psi)$ in the estimation of the standard errors is that, $\mathcal{I}_c(\Psi)$ tends to underestimate the standard errors. That is

$$\begin{aligned} \mathcal{I}(\Psi)^{-1} &= \mathcal{I}_c(\Psi)^{-1} + \triangle \boldsymbol{V} \,, \\ \triangle \boldsymbol{V} &= [\boldsymbol{I} - \mathcal{J}(\Psi)]^{-1} \mathcal{J}(\Psi) \mathcal{I}_c(\Psi)^{-1} \,. \end{aligned} \tag{5.4}$$

The quantity $\triangle \boldsymbol{V}$ measures the increase in the asymptotic variance due to missing information (McLachlan & Krishnan, 1997, Meng & Rubin, 1991) . It is determined by the rate of convergence matrix $\mathcal{J}(\Psi)$ and the complete data information $\mathcal{I}_c(\Psi)$.

The rate of convergence matrix $\mathcal{J}(\Psi)$ is a measure of the convergence speed of the algorithm. The derivation of the rate matrix $\mathcal{J}(\Psi)$ begins with a Taylor expansion of the mapping function $\Psi^{(it+1)} = \mathcal{F}(\Psi^{(it)})$ at the point $\Psi^{(it)} = \Psi^*$,

$$\mathcal{F}\left(\Psi^{(it)}\right) \approx \mathcal{F}(\Psi^*) + \left[\partial \mathcal{F}\left(\Psi^{(it)}\right) / \partial \Psi^{(it)}\right]_{\Psi^{(it)}=\Psi^*} \left(\Psi^{(it)} - \Psi^*\right) ,$$

which then gives

$$\begin{aligned} \Psi^{(it+1)} - \Psi^* &\approx \left[\partial \mathcal{F}\left(\Psi^{(it)}\right) / \partial \Psi^{(it)}\right]_{\Psi^{(it)}=\Psi^*} \left(\Psi^{(it)} - \Psi^*\right) . \\ &= \mathcal{J}(\Psi^*) \left(\Psi^{(it)} - \Psi^*\right) , \end{aligned} \tag{5.5}$$

where $\mathcal{J}(\Psi^*)$ is the rate matrix evaluated at Ψ^* (McLachlan & Krishnan, 1997). However, the computation of the rate matrix is usually difficult, which complicates the evaluation of the observed information $\mathcal{I}(\Psi)$ using equation (5.4). As a result, the approximation of the standard errors of parameter estimates becomes challenging.

Over the years, various approximation methods have been proposed in the literature to obtain the observed information $\mathcal{I}(\Psi)$. Among these methods, two of them are of direct relevance to this thesis.

(a) The first method avoids the computation of the rate of convergence $\mathcal{J}(\Psi)$ and approximates the observed information matrix using the score statistics. According to McLachlan & Krishnan (1997), $\mathcal{I}(\Psi)$ can be approximated using the conditional expectation of the gradient vector (i.e. the score statistic). The resulting matrix is called the empirical information matrix and is denoted as $\mathcal{I}_e(\Psi; \boldsymbol{Z})$. Under the i.i.d. assumption of the data $\boldsymbol{Z}_t$, matrix $\mathcal{I}_e(\Psi; \boldsymbol{Z})$ is constructed as

$$\begin{aligned} \mathcal{I}_e(\Psi; \boldsymbol{Z}) &= \sum_{t=1}^{T} I(\Psi; \boldsymbol{Z}_t) \\ &= \sum_{t=1}^{T} F(\boldsymbol{Z}_t; \Psi) F(\boldsymbol{Z}_t; \Psi)^\top - \frac{1}{T} \left[\sum_{t=1}^{T} F(\boldsymbol{Z}_t; \Psi)\right] \left[\sum_{t=1}^{T} F(\boldsymbol{Z}_t; \Psi)\right]^\top , \end{aligned}$$

where $I(\Psi; \boldsymbol{Z}_t)$ and $F(\boldsymbol{Z}_t; \Psi)$ are the observed information and score statistic for the t-th observation. On evaluating the equation at the MLE Ψ^*, the second term can be treated as approximately zero, giving the approximation

$$\mathcal{I}_e(\Psi^*; \boldsymbol{Z}) \approx \sum_{t=1}^{T} F(\boldsymbol{Z}_t; \Psi^*) F(\boldsymbol{Z}_t; \Psi^*)^\top \,. \tag{5.6}$$

In addition, it has been shown in McLachlan & Krishnan (1997) that $F(\boldsymbol{Z}_t; \Psi)$ can be obtained using the complete data log-likelihood as

$$F(\boldsymbol{Z}_t; \Psi) = \mathbf{E}\left[\frac{\partial \mathcal{L}_c(\Psi; \boldsymbol{Z}_t)}{\partial \Psi} \,\middle|\, \boldsymbol{Z}; \Psi \right] \,. \tag{5.7}$$

This means, the computation of the empirical information matrix could be carried out with just a little extra work in addition to the GEM procedure.

(b) The second method adopts a supplemented algorithm alongside the EM-type algorithm to approximate the rate of convergence matrix $\mathcal{J}(\Psi)$. Then the information matrix can be obtained using equation (5.4). Depending on the algorithms, $\mathcal{J}(\Psi)$ is linked to the observed information matrix $\mathcal{I}(\Psi)$ in specific ways (Liu & Rubin, 1994, McLachlan & Krishnan, 1997, Meng & Van Dyk, 1997). For example, equation (5.4) is for the standard EM algorithm; whereas the rate of convergence matrix of the AECM was derived in Meng & Van Dyk (1997) as

$$\mathcal{J}^{AECM}(\Psi) = \prod_{c=1}^{C} \left\{ \boldsymbol{I} - \mathcal{I}(\Psi) \mathcal{I}_c^{[c]}(\Psi)^{-1} \left(\boldsymbol{I} - \prod_{s=1}^{S^{[c]}} \boldsymbol{P}_s^{[c]} \right) \right\} \tag{5.8}$$

where

$$\boldsymbol{P}_s^{[c]} = \bigtriangledown_s^{[c]} \left\{ \left(\bigtriangledown_s^{[c]} \right)^\top \mathcal{I}_c^{[c]}(\Psi)^{-1} \left(\bigtriangledown_s^{[c]} \right) \right\}^{-1} \left(\bigtriangledown_s^{[c]} \right)^\top \mathcal{I}_c^{[c]}(\Psi)^{-1}$$
$$\bigtriangledown_s^{[c]} = \bigtriangledown g_s^{[c]}(\Psi^*) \,,$$

and $g_s^{[c]}(\boldsymbol{.})$ are constraint functions associated with the space-filling conditions. Based on these connections, supplemented algorithms were proposed to approximate the elements in the rate of convergence matrix through numerical differentiation (Meng & Rubin, 1991, Van Dyk *et al.*, 1995). Define r_{mn}^* as the (m, n)-th element of the matrix $\mathcal{J}(\Psi)$. After the EM-type algorithm converges, run a few more iterations beginning with a parameter vector $\Psi^{(it)}$ whose elements are close to, but slightly different from the converged Ψ^*. Compute

$$r_{m,n}^{(it)} = \frac{\tilde{\psi}_n^{(it+1)} - \psi_n^*}{\psi_m^{(it)} - \psi_m^*} \tag{5.9}$$

for $m,\, n\, = 1, \cdots, q$, where q is the dimension of the parameter set Ψ. The estimates $\tilde{\psi}_m^{(it+1)}$ in equation (5.9) can be obtained from the output of one iteration of the EM-type alorithm with input

$$\Psi^{(it)}(m) = \left(\psi_1^*, \cdots, \psi_{m-1}^*, \psi_m^{(it)}, \psi_{m+1}^*, \cdots, \psi_M^*\right) .$$

The iteration stops if a discrepancy measure $\delta(r_{m.n}^{(it+1)}, r_{m.n}^{(it)})$ is less than certain threshold. After convergence, set $r_{mn}^* = r_{mn}^{(it+1)}$. It is easy to see the connection between the supplemented algorithm and the derivation of $\mathcal{J}(\Psi)$ using the Taylor expansion (5.5). With the supplemented algorithms, the entire matrix $\mathcal{J}(\Psi)$ can be approximated. The observed information $\mathcal{I}(\Psi)$ can therefore be obtained.

The supplemented algorithm can be used to obtain the observed information matrix $\mathcal{I}(\Psi)$ of the AECM algorithm from equation (5.8). However, the implementation is non-trivial, especially when the algorithm involves multiple cycles, i.e. $C \geq 2$. Meng & Van Dyk (1997) mentioned using a corresponding $C = 1$ algorithm (provided it exists) in conjunction with the supplemented ECM algorithm to compute $\mathcal{I}(\Psi)$. Alternatively, for the specific case of $C = 2$, $\mathcal{I}(\Psi)$ might be obtained by solving a quadratic matrix equation

$$\boldsymbol{X}\boldsymbol{A}\boldsymbol{X} + \boldsymbol{B}\boldsymbol{X} + \boldsymbol{C}\boldsymbol{X} = \boldsymbol{D}\,,$$

where $\boldsymbol{X} = \mathcal{I}(\Psi)$ and $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ are constructed using elements associated with $\mathcal{I}_c^{[1]}(\Psi)$, $\mathcal{I}_c^{[2]}(\Psi)$, $\mathcal{J}(\Psi)$ and $g_s(\Psi)$. See Appendix C.1 for more explanation.

5.3.3 Practical results of the SS-FPCA

The two methods introduced in section 5.3.2 for approximating the observed information matrix $\mathcal{I}(\Psi)$ can be used to estimate the standard errors of MLEs of the parameters in the SS-FPCA model. In particular, the first method using $\mathcal{I}_e(\Psi)$ as defined in (5.6) is adopted in this thesis. Both method can suffer from numerical inaccuracies and instability, especially in high-dimensional settings (McLachlan & Krishnan, 1997). However, evaluating $\mathcal{I}_e(\Psi)$ based on the score statistics is supposed to be computationally more efficient than the supplemented algorithm, especially when the number of parameters is large. It is also easier to implement, provided there are neat analytical solutions to the partial derivatives of the complete data log-likelihood and the corresponding conditional expectations.

The complete data log-likelihood of the SS-FPCA model (4.27) is

$$
\begin{aligned}
&\mathcal{L}_c(\Psi; \boldsymbol{Z}_{1:T}, \boldsymbol{\beta}_{1:T}, \boldsymbol{\alpha}_{1:T}) \qquad (5.10)\\
=& -\frac{1}{2}\sum_{t=1}^{T}\left\{ n_t \log\left(\sigma^2\right) + \frac{1}{\sigma^2}\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_t - \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t\right)^\top \left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_t - \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t\right)\right\}\\
&-\frac{1}{2}\sum_{t=1}^{T}\left\{ \log\left(|\boldsymbol{H}|\right) + \left(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}\right)^\top \boldsymbol{H}^{-1}\left(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1}\right)\right\}\\
&-\frac{1}{2}\left\{ \log\left(|\boldsymbol{B}_0|\right) + \left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right)^\top \boldsymbol{B}_0^{-1}\left(\boldsymbol{\beta}_0 - \boldsymbol{\beta}\right)\right\}\\
&-\frac{1}{2}\sum_{t=1}^{T}\left\{ \log\left(|\boldsymbol{\Lambda}|\right) + \boldsymbol{\alpha}_t^\top \boldsymbol{\Lambda}^{-1}\boldsymbol{\alpha}_t\right\} + \text{constant}\ .
\end{aligned}
$$

The corresponding information matrix is block diagonal, consisting of one block for the second order partial derivatives with respect to σ^2 and $\boldsymbol{\theta}_p$, $p = 1, \cdots, P$, one block with respect to the second derivatives of λ_p, $p = 1, \cdots, P$, and one block associated with $\boldsymbol{H}$ (see Appendix C.2). Each block needs to be evaluated and inverted separately, in order to get the estimation of the standard errors. Specifically, the first and second derivatives of log-likelihood (5.10) with respect to σ^2, λ_p, $\boldsymbol{\theta}_p$ (the vector as an entity) and $\boldsymbol{H}$ (the matrix as an entity) can be derived as follows.

$$
\frac{\partial \mathcal{L}_c(\Psi)}{\partial \sigma^2} = -\frac{1}{2}\sum_{t=1}^{T}\left\{ \frac{n_t}{\sigma^2} - \frac{1}{\sigma^4}\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_t - \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t\right)^\top \left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_t - \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t\right)\right\}\ , \qquad (5.11)
$$

$$
\frac{\partial^2 \mathcal{L}_c(\Psi)}{\partial \sigma^2 \partial \sigma^2} = \frac{1}{2}\sum_{t=1}^{T}\left\{ \frac{n_t}{\sigma^4} - \frac{2}{\sigma^6}\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_t - \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t\right)^\top \left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_t - \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t\right)\right\}\ , \qquad (5.12)
$$

$$
\frac{\partial \mathcal{L}_c(\Psi)}{\partial \boldsymbol{\theta}_p} = \frac{1}{\sigma^2}\sum_{t=1}^{T}\left\{ \boldsymbol{\Phi}_t^\top\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_t\right)\alpha_{pt} - \alpha_{pt}^2\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t\boldsymbol{\theta}_p - \sum_{q\neq p}\alpha_{pt}\alpha_{qt}\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t\boldsymbol{\theta}_q\right\}\ , \qquad (5.13)
$$

$$
\frac{\partial^2 \mathcal{L}_c(\Psi)}{\partial \boldsymbol{\theta}_p \partial \boldsymbol{\theta}_p^\top} = -\frac{1}{\sigma^2}\sum_{t=1}^{T}\left\{ \alpha_{pt}^2\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t\right\}\ , \qquad (5.14)
$$

$$
\frac{\partial^2 \mathcal{L}_c(\Psi)}{\partial \boldsymbol{\theta}_p \partial \boldsymbol{\theta}_q^\top} = -\frac{1}{\sigma^2}\sum_{t=1}^{T}\left\{ \alpha_{pt}\alpha_{qt}\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t\right\}\ , \qquad (5.15)
$$

$$
\frac{\partial^2 \mathcal{L}_c(\Psi)}{\partial \boldsymbol{\theta}_p \partial \sigma^2} = -\frac{1}{\sigma^4}\sum_{t=1}^{T}\left\{ \boldsymbol{\Phi}_t^\top\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_t\right)\alpha_{pt} - \alpha_{pt}^2\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t\boldsymbol{\theta}_p - \sum_{q\neq p}\alpha_{pt}\alpha_{qt}\boldsymbol{\Phi}_t^\top\boldsymbol{\Phi}_t\boldsymbol{\theta}_q\right\}\ , \qquad (5.16)
$$

$$\frac{\partial \mathcal{L}_c(\Psi)}{\partial \lambda_p} = -\frac{1}{2}\sum_{t=1}^{T}\left\{\frac{1}{\lambda_p} - \frac{1}{\lambda_p^2}\alpha_{pt}^2\right\}, \tag{5.17}$$

$$\frac{\partial^2 \mathcal{L}_c(\Psi)}{\partial \lambda_p \partial \lambda_p} = \frac{1}{2}\sum_{t=1}^{T}\left\{\frac{1}{\lambda_p^2} - \frac{2}{\lambda_p^3}\alpha_{pt}^2\right\}, \tag{5.18}$$

$$\begin{aligned}\frac{\partial \mathcal{L}_c(\Psi)}{\partial \boldsymbol{H}} =& -\frac{1}{2}\sum_{t=1}^{T}\Big\{\mathbf{vec}(\boldsymbol{H}^{-\top})^\top + \left[(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^\top \otimes (\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^\top\right] \\ & \left(-\boldsymbol{H}^{-\top} \otimes \boldsymbol{H}^{-1}\right)\Big\},\end{aligned} \tag{5.19}$$

$$\begin{aligned}\frac{\partial^2 \mathcal{L}_c(\Psi)}{\partial \boldsymbol{H} \partial \boldsymbol{H}^\top} =& \frac{1}{2}\sum_{t=1}^{T}\Big\{\boldsymbol{I}_{K^2} - \boldsymbol{I}_{K^2} \otimes \left[(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^\top \otimes (\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^\top\right] \\ & \left[(\boldsymbol{I}_K \otimes \boldsymbol{T}_{K,K} \otimes \boldsymbol{I}_K)\left(\mathbf{vec}\left(\boldsymbol{H}^{-\top}\right) \otimes \boldsymbol{I}_{K^2}\right)\right]\Big\}\left(\boldsymbol{H}^{-\top} \otimes \boldsymbol{H}^{-1}\right),\end{aligned} \tag{5.20}$$

where $\boldsymbol{I}_K$, $\boldsymbol{I}_{K^2}$ are identity matrices of dimension K and K^2 (K is the basis dimension) respectively and $\boldsymbol{T}_{K,K}$ is the permutation matrix which satisfies $\mathbf{vec}(\boldsymbol{H})\boldsymbol{T}_{K,K} = \mathbf{vec}(\boldsymbol{H}^\top)$.

The evaluation of the conditional expectations of derivatives (5.11) to (5.18) uses the E-step outputs from cycle 2 of the AECM algorithm in section 4.4.2. By recognizing that $(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^\top \otimes (\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^\top = \mathbf{vec}\left[(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^\top\right]$, the only thing required for the conditional expectation of derivatives (5.19) and (5.20) is the expectation of matrix $(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^\top$ given all the data $\boldsymbol{Z}_{1:T}$, which is part of the routine of cycle 1. More details can be found in Appendix C.2

As an illustration of the approximation of the standard errors using the Fisher information matrix, an example based on the simulated data in section 5.2.1 is presented. This example does not intend to assess the asymptotic properties of the SS-FPCA model, but to provide some practical results only. Standard errors of the variance parameters, σ^2, λ_p, $p = 1, 2$, and h_k, $k = 1, \cdots, 7$, are investigated. Despite the fact that the MLEs of these parameters are biased in the majority of the simulation scenarios, it is interesting to investigate the behavior of their approximated standard errors. In this example, the Fisher information based on both the complete data log-likelihood and the score statistics, i.e. $\mathcal{I}_c(\Psi)$ and $\mathcal{I}_e(\Psi)$, were computed. The standard errors $\sigma(\Psi)$ were estimated using the negative inverse of the two. The lengths of the approximated 95% confidence intervals $[-1.96\,\sigma(\Psi), 1.96\,\sigma(\Psi)]$ (referred to as the approximated CI) were compared to the lengths of the 95% confidence intervals obtained from the 500 simulation replicates (referred to as the simulation CI). Tables 5.8 and 5.9 present the results from the replicates of simulation scenario S1 and S12. These two

scenarios correspond to a 'good' and a 'bad' cases respectively according to the sparsity and noise levels, so the results are somewhat representative.

TABLE 5.8: The lengths of the 95% confidence interval (CI) of the MLEs of σ^2, λ_1 and λ_2, based on the 500 replicates and on the approximations using information matrices $\mathcal{I}_c(\Psi)$ and $\mathcal{I}_e(\Psi)$ from the 100th, 300th and 500th replicates, from scenario 1 and 12.

	S1			**S12**		
	σ^2	λ_1	λ_2	σ^2	λ_1	λ_2
True value	0.0100	15.0633	2.8196	0.2500	15.0633	2.8196
CI Simulation	0.0192	8.0067	2.1039	0.0425	11.5399	2.5345
MLE (100th rep)	0.0669	12.7253	1.6715	0.3332	18.0375	2.7422
CI from $\mathcal{I}_c$	0.0052	7.1368	0.9472	0.0316	10.1313	1.5528
CI from $\mathcal{I}_e$	0.0110	8.0725	1.0786	0.0450	11.7779	2.2763
MLE (300th rep)	0.0636	11.2393	1.6222	0.3199	16.7336	2.1976
CI from $\mathcal{I}_c$	0.0049	6.3891	0.9173	0.0303	6.3199	0.7165
CI from $\mathcal{I}_e$	0.0102	7.7386	1.1547	0.0482	7.3977	1.2379
MLE (500th rep)	0.0717	11.2594	3.0367	0.3280	16.2105	4.2377
CI from $\mathcal{I}_c$	0.0056	6.3440	1.7190	0.0311	9.1179	2.4128
CI from $\mathcal{I}_e$	0.0111	7.4402	1.7612	0.0477	10.4642	2.7803

It can be seen in Table 5.8 and 5.9 that, the lengths of the 95% confidence intervals obtained from the complete data information matrix $\mathcal{I}_c(\Psi)$ are shorter than those obtained from the empirical observed information matrix $\mathcal{I}_e(\Psi)$. This is expected as the complete data information matrix tends to underestimate the standard errors of the MLEs. The differences between $\mathcal{I}_c(\Psi)$ and $\mathcal{I}_e(\Psi)$ are relatively larger in scenario S12 (see Table 5.9). A possible explanation is that, although 'missing information' in the AECM algorithm are in fact 'latent variables', the higher proportion of missing observations means the ratio of unknown versus known is larger. Therefore $\mathcal{I}_c(\Psi)$ tends to deviate more from $\mathcal{I}_e(\Psi)$.

There is no clear pattern in terms of the lengths of the 95% simulation CIs as compared to the lengths of the two approximated CIs. In the case of $\hat{\sigma}^2$, the simulation CIs are wider than the two approximated CIs. The simulation CIs of $\hat{\lambda}_1$ and $\hat{\lambda}_2$ tend to be wider than the approximated CIs, but with some exceptions. The situation with respect to $\hat{h}_k$, $k = 1, \cdots, 7$, is even more unpredictable. In addition, the lengths of the 95% confidence intervals for $\hat{\sigma}^2$, $\hat{\lambda}_1$ and $\hat{\lambda}_2$ using three different methods are comparable, but those for $\hat{h}_k$, $k = 1, \cdots, 7$, appear to vary greatly.

A possible explanation for the discrepancies in the 95% CIs of h_k from three different methods is as follows. Information matrices $\mathcal{I}_c(\hat{h}_k)$ and $\mathcal{I}_e(\hat{h}_k)$ were computed using the MLE of h_k and the Kalman smoothed series $\{\boldsymbol{\beta}_{t|T}\}_{t=1}^{T}$, based on the current replicate. As the Kalman

TABLE 5.9: The lengths of the 95% confidence interval (CI) of the MLEs of $h_1, \cdots, h_7$, based on the on the 500 replicates and on the approximations using information matrices $\mathcal{I}_c(\Psi)$ and $\mathcal{I}_e(\Psi)$ from the 100th, 300th and 500th replicates, from scenario S1 and S12.

				S1			
	h_1	h_2	h_3	h_4	h_5	h_6	h_7
True value	0.3364	0.2500	0.4225	0.2500	0.2704	0.6241	0.2809
CI Simulation	0.4356	0.2376	0.3551	0.0807	0.1552	0.1477	0.0595
MLE (100th rep)	0.2017	0.1659	0.2213	0.2162	0.3003	0.5631	0.2050
CI from $\mathcal{I}_c$	0.1141	0.0938	0.1260	0.1221	0.1698	0.3182	0.1159
CI from $\mathcal{I}_e$	0.3149	0.2030	0.6490	0.2004	0.2605	0.4117	0.2043
MLE (300th rep)	0.2034	0.1858	0.2053	0.1889	0.3024	0.5602	0.2189
CI from $\mathcal{I}_c$	0.1150	0.1050	0.1169	0.1067	0.1710	0.3166	0.1237
CI from $\mathcal{I}_e$	0.3393	0.2594	0.5483	0.2035	0.2682	0.4102	0.1970
MLE (500th rep)	0.1705	0.2172	0.2835	0.1779	0.2749	0.4940	0.1904
CI from $\mathcal{I}_c$	0.0967	0.1229	0.1612	0.1005	0.1554	0.2791	0.1076
CI from $\mathcal{I}_e$	0.3418	0.2384	0.5341	0.2090	0.2552	0.3630	0.2022
				S12			
	h_1	h_2	h_3	h_4	h_5	h_6	h_7
True value	0.3364	0.2500	0.4225	0.2500	0.2704	0.6241	0.2809
CI Simulation	0.2804	0.2518	0.5149	0.1847	0.3099	0.4057	0.1696
MLE (100th rep)	0.2638	0.1923	0.2540	0.2707	0.3624	0.6548	0.2497
CI from $\mathcal{I}_c$	0.1492	0.1088	0.1446	0.1529	0.2050	0.3701	0.1412
CI from $\mathcal{I}_e$	0.5797	0.5611	1.4697	0.5379	0.8082	0.8377	0.4743
MLE (300th rep)	0.2454	0.2461	0.1749	0.1911	0.3246	0.5702	0.1811
CI from $\mathcal{I}_c$	0.1389	0.1392	0.0999	0.1079	0.1835	0.3223	0.1024
CI from $\mathcal{I}_e$	0.6556	0.6426	0.9923	0.3674	0.6735	0.6533	0.3540
MLE (500th rep)	0.1690	0.2690	0.3856	0.2798	0.2599	0.5223	0.1952
CI from $\mathcal{I}_c$	0.0957	0.1522	0.2184	0.1582	0.1470	0.2954	0.1104
CI from $\mathcal{I}_e$	0.5695	0.6760	1.0130	0.6466	0.5515	0.6493	0.4679

filter/smoother relies heavily on the data, matrices $\mathcal{I}_c(\hat{h}_k)$ and $\mathcal{I}_e(\hat{h}_k)$ would also be closely related to the data in the current replicate, perhaps more so than the information matrices of $\hat{\sigma}^2$, $\hat{\lambda}_1$ and $\hat{\lambda}_2$. As a result, the approximated CIs vary across the replicates. On the contrary, the simulation CI is a summary of 500 data sets, which accommodates the variation of all the replicates. Therefore, the large discrepancies are not unrealistic due to the variations brought by the random noises in the simulated data sets.

Finally, turn to the question of whether the approximated CIs from the information matrices can be used as approximations of the standard errors of the MLEs of the SS-FPCA model. The answer based on the above results would be, 'yes, but with caution'. In particular, when the proportion of missing information (in terms of the latent variables) is large, the discrepancy between the complete data information matrix and the observed data information

matrix would also be large. Hence the reliability of the approximation using (5.6) becomes questionable. In practice, if it is computationally feasible to obtain the bootstrap standard errors or the simulated confidence intervals, then it should be considered as a better solution. Otherwise, the asymptotic results may be used.

5.4 Application to the sparse Lake Victoria data

Application 1: LSWT The SS-FPCA model was applied to the sparse Lake Victoria LSWT data. The same data were modelled using the mixed model FPCA in Chapter 3, with the LSWT images assumed to be independent from each other. Now with the SS-FPCA model, both the temporal and spatial patterns can be taken into account. As a reminder, the data set is a subset of the Lake Victoria LSWT data, consisting of 202 monthly images, each defined on a grid of 49×57. The grid is constructed by trimming off boundary pixels and some extensions of the water body at the edge of the lake (see section 3.3 for a detailed description). The data were centered using a monthly mean initially.

First of all, the basis dimension K and the K-L expansion order P were chosen using the two stage method described in section 5.1.2. The selection of the basis used the MM-FPCA approach and the appropriate basis according to the AIC and BIC values is of dimension $K = 7 \times 7$ (refer to Figure 3.12 in Chapter 3). In terms of the expansion order, both the selection through fitting the SS-FPCA model and the one using the variance proportion of $\delta\% \geq 95\%$ suggested $P = 4$. A filtering threshold of $\chi\% = 95\%$ was used. That is, any images with less than 5% data available would not be filtered and would only be smoothed based on the information from their neighbouring images.

The procedure described in section 4.4.3 was used to initialize the model parameters. In particular, the initial values of the state space component are $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\boldsymbol{H}^{(0)} = \sigma_h^2 \boldsymbol{I}$, where $\sigma_h^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Var}[Z_t(x_i, y_i)]$. The initial values of the FPCA component, $\boldsymbol{\Lambda}^{(0)}$ and $\boldsymbol{\Theta}^{(0)}$ are computed using the decomposition of the covariance matrix of the column stacked data. The initial value of the residual variance $\sigma^{2(0)}$ was set to the estimated residual variance from the initialization of the FPCA component, $\sigma^{2(0)} = \frac{1}{\sum_t n_t}\sum_{t=1}^{T} \hat{\boldsymbol{r}}_t^\top \hat{\boldsymbol{r}}_t$.

The AECM algorithm converged after 8 iterations, under the convergence criterion of the change of the estimated log-likelihood being $\leq 0.025\%$. The average computation time for one AECM iteration was about 38 minutes. The estimated residual variance of the model is $\hat{\sigma}^2 = 0.0872$. The four eigenvalues after ortho-normalization are $\hat{\lambda}_1 = 6704.36.7$,

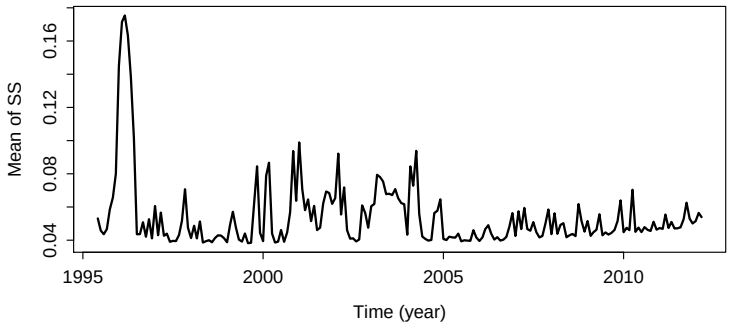

FIGURE 5.17: Time series plot of the spatial means of the element in the covariance matrix of the dynamic component (or SS component) $\boldsymbol{\Phi}\boldsymbol{B}_{t|T}\boldsymbol{\Phi}^\top$.

$\hat{\lambda}_2 = 1466.46$, $\hat{\lambda}_3 = 319.43$ and $\hat{\lambda}_4 = 166.51$. A time series plot of the spatial means of the variance of the dynamic component, which are computed as the average of the elements in $\boldsymbol{\Phi}\boldsymbol{B}_{t|T}\boldsymbol{\Phi}^\top$ at each time point t, is displayed in Figure 5.17. This helps to examine the relative scales of the variance of $\boldsymbol{\Phi}\boldsymbol{\beta}_t$. In this case, the variances at the beginning of the observing period appear to be larger than the rest of the time. This can be explained by the higher missing percentages in this period, which includes months with no observation. In addition, the dynamic component $\boldsymbol{\Phi}\boldsymbol{\beta}_t$ appears to have greater influence in explaining the total variation than the FPCA component $\boldsymbol{\Phi}\boldsymbol{\Theta}\boldsymbol{\alpha}_t$. This can be seen from the RSS using three different residuals defined in (5.2), with RSS of the model residuals $\boldsymbol{\epsilon}^m$ being 0.0811, RSS of residuals after accounting for dynamic component $\boldsymbol{\epsilon}^d$ being 0.1253 and that of residuals after accounting the FPCA component $\boldsymbol{\epsilon}^s$ being 0.2166.

Figure 5.18 shows the image of the eigenfunctions and the scores of the leading two principal components. The first PC accounts for 77.44% of the variation captured by the FPCA component; the second PC explains about 16.94% of the variation. The third and fourth PC each explain less than 4% of the variation, and therefore are regarded as less important in this case. Specifically, the first eigenfunction (top left) displays a spatial pattern with positive loadings in the middle, northeast of the lake and negative loadings in the south, northwest of the lake. The second eigenfunction (bottom left) shows a contrast between the west and east half of the lake. As in the MM-FPCA, the times series of PC scores can be viewed as the evolution of the spatial pattern. In this example, no distinctive temporal trend is found in the time series of PC score (two panels on the right).

Then the reconstructions were computed using the fitted SS-FPCA model. Figure 5.19

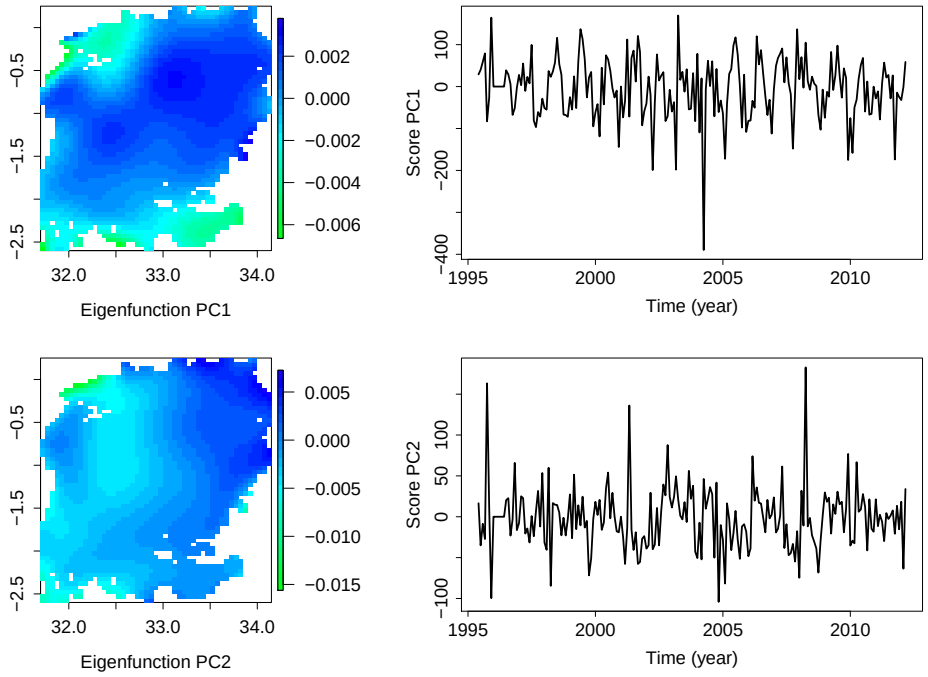

FIGURE 5.18: The plots of the eigenfunctions and the scores of the PC1 (top) and PC2 (bottom). The horizontal and vertical axes of the eigenimages represent longitude and latitude respectively.

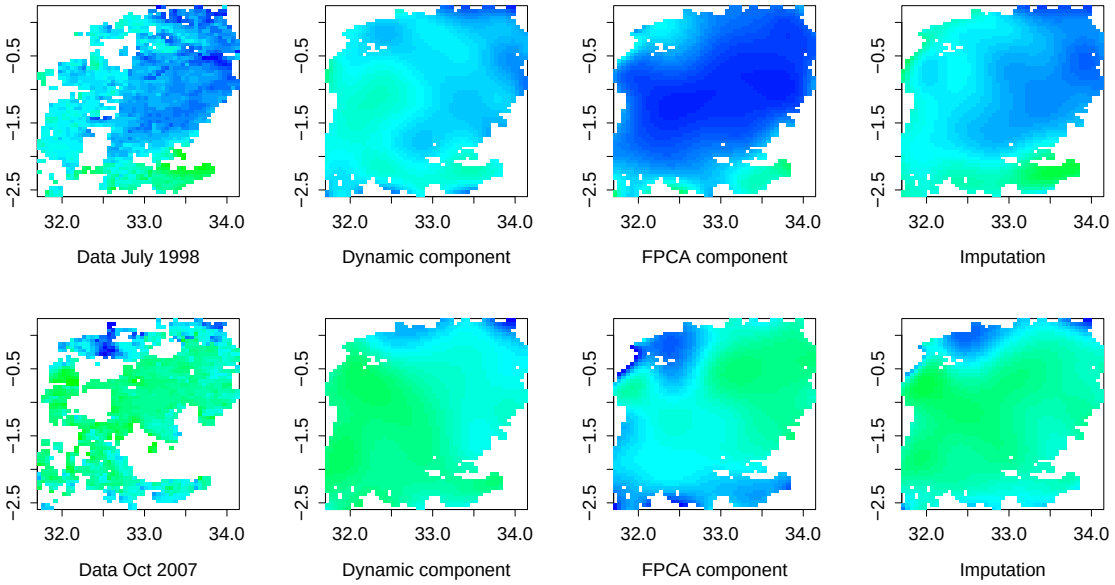

FIGURE 5.19: Observations, dynamic component, FPCA component and data reconstruction using the SS-FPCA model for the July 1998 data (top) and the October 2007 data (bottom). The horizontal and vertical axes are longitude and latitude respectively.

presents, from left to right, the images of the observed data, the estimated dynamic component, the estimated FPCA component and the data reconstruction, from July 1998 (upper panels) and October 2007 (lower panels). Images in each row were plotted using the same colour scheme, so that the comparison can be made easily. In general, data reconstructions can be regarded as the joint contribution of the dynamic and the FPCA component, each accounting for different variation patterns in the data. For instance, in the October 2007 model, the dynamic and the FPCA components both show the contract between the northeast and the southwest of the lake. However, the combination of the two results in a smooth image which reflects the spatial patterns in the original data.

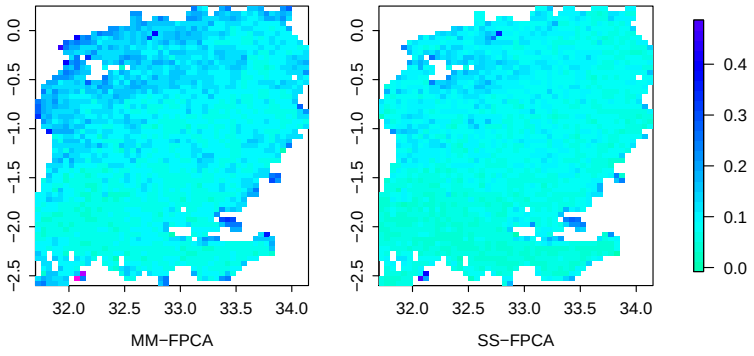

FIGURE 5.20: Maps of the RSS of each individual pixel from the MM-FPCA (left) and the SS-FPCA (right). The horizontal and vertical axes are longitude and latitude respectively.

Finally, a comparison of the SS-FPCA to the MM-FPCA was carried out to see if there is any actual improvement from incorporating the temporal dependence. The same 7×7 tensor spline basis was used in the MM-FPCA. The expansion order was set to $P = 4$ to be consistent with the SS-FPCA model. The MM-FPCA has $\hat{\sigma}^2 = 0.1211$ and RSS $= 0.1207$ from image reconstruction, larger than the RSS of the SS-FPCA model.

In terms of data imputation, the SS-FPCA is able to catch more detail in the observed data. This may be difficult to spot from the reconstructed images, but can be illustrated by the RSS computed using residuals in individual pixels (or pixel RSS). The same measure has been used in Chapter 3 for a comparison between the MM-FPCA reconstruction and the ARC-Lake reconstruction. Figure 5.20 presents two plots of the pixel RSS, where the left panel shows the MM-FPCA and right panel shows the SS-FPCA. The two images were produced using the same colour scheme for the ease of comparison. The cyan end of the palette corresponds to small RSS values and the blue end corresponds to high RSS values. It can be seen that RSS in the left panel are slightly larger than those in the right panel,

especially in the northwest corner where the missing percentages are higher, indicating an improvement brought by incorporating the temporal dependence.

In terms of the computation time, applying the MM-FPCA took much less time than fitting the SS-FPCA model, especially for lower expansion order cases. In this application, the computation of the MM-FPCA took only 62.84 seconds. However, the MM-FPCA ignores the potential temporal correlations among the images, which can be inappropriate in some situations. Whereas the SS-FPCA model accounts for it. The advantages of accounting for temporal dependence may not be seen straightforwardly in the above comparison, but the potential of obtaining a more reliable estimation of the functional PCs and a better data imputation is still attractive.

Application 2: Chlorophyll Analogous to the demonstration at the end of section 3.3, an application of the SS-FPCA model to the Lake Victoria Chlorophyll data is presented. The same subset taken from the 3×3 spatially aggregated Lake Victoria Chl data set was investigated here. This subset is of dimension $72 \times 72 \times 112$ and the missing percentage is 5.3%. The log transformation was applied to the data initially and they were then centered by a monthly mean. The same 7×6 basis as in section 3.3 was used. The expansion order was taken to be $P = 8$. The filtering threshold was set to $\chi\% = 95\%$.

Under the convergence criterion of $\varepsilon \leq 0.01\%$, the SS-FPCA model converged after 4 iterations, each taking about 20 minutes. Two dominant PCs were identified in this case, each explaining 37.30% and 34.49% of the variation in the FPCA component. The eigenfunctions of the two leading PCs show the contrast between northwest and southeast, northeast and southwest as the first two eigenfunctions from the MM-FPCA. However, the detailed patterns are different. The fitted model has RSS of 0.0376, smaller than the RSS from the MM-FPCA, which is 0.0539. There are still problems with respect to the discontinuities in certain images, but the reconstructions from the SS-FPCA model capture more details than the MM-FPCA, even with a smaller expansion order. This can be regarded as a benefit from incorporating the temporal dependence. An analysis on the three different types of model residuals shows that the contributions from the dynamic component and the FPCA component to the total variation are comparable. Neither is dominant in explaining the total variation. In general, the SS-FPCA appears to provide a better fit to the data than the MM-FPCA. However, it should be noted that the discontinuities in the data need to be handled with caution. Similar

to the MM-FPCA, the SS-FPCA tend to smooth over the rough images, which could result in a loss of information if the discontinuity is the interest of the analysis.

5.5 Remarks

The SS-FPCA model is designed to handle tasks such as dimension reduction, imputing missing observations and modelling spatial/temporal patterns in the data. It provides a solution to the three challenges presented at the beginning of the thesis with respect to the high-dimensional, sparse remote-sensing data. It is possible to gauge the scales of the spatial/temporal variations of the data through computing the variance of the dynamic and the FPCA components. The estimated coefficients of the dynamic component, the eigenfunctions and PC scores may also help to identify the dominant spatial/temporal patterns, provided they exist. All of these are of interest in terms of the analysis of remote-sensing environmental data. In general, the SS-FPCA is a useful model.

The criticisms of this method include the relatively high computational cost and slow convergence of the AECM algorithm in some situations. In particular, the computation time of the SS-FPCA model can be significantly longer than the MM-FPCA; whereas the convergence of the algorithm can be slow if the initial values are selected inappropriately. A detailed timer shows that the majority of the computation time is consumed by the Kalman filter, where high-dimensional matrix inversions are sometimes involved. Another problem is the identification of the dynamic and the FPCA components. The simulation study in section 5.2.1 showed that SS-FPCA model fitted using the 2-cycle AECM algorithm, while capturing the main patterns in the dynamic and the FPCA components, appeared to under/over-estimate the variances of the model components. This problem is associated with the spatial confounding, which is common to many spatial and spatio-temporal models and has been investigated in literature for various types of analysis (Hodges & Reich, 2010, Hughes & Haran, 2013, Paciorek, 2010). The situation for the SS-FPCA model is prone to be non-trivial. Detailed investigations are required to help improve the identifiability of the model.

The SS-FPCA model presented in Chapter 4 and 5 is in its basic form. Extensions can be made by modifying the specifications of various model components, e.g. the design of system dynamic and the dependence of random components. In model (4.27), the system dynamic is supposed to follow a local level model. Alternatively, vector AR models, PDE and IDE can be used if there are evidences supporting such a specification (Cressie & Wikle, 2011). The

covariance matrix $\boldsymbol{H}$ of the state transition equation may also be parameterized to reflect a more structured temporal dependence. Parameterizations summarised in section 4.3.1 provide some handy options. As for the dependence of the random components, it would be interesting to investigate the influence of the independence assumption of $\boldsymbol{\beta}_t$ and $\boldsymbol{\alpha}_t$ given data $\boldsymbol{Z}_{1:T}$. The challenge lies in the derivation and evaluation of the conditional distribution of $\boldsymbol{\alpha}_t$, $\boldsymbol{\beta}_t|\boldsymbol{Z}_{1:T}$, if the independence assumption is dropped. The fixed rank filtering method by Cressie *et al.* (2010) offers some intuitions to this problem. It is generally believed that accounting for this dependence would improve the model estimation and the subsequent statistical inferences.

Chapter 6

Conclusion

In this thesis, statistical methodologies for the analysis of remote-sensing image time series were investigated and developed, including methods for dimension reduction, missing data imputation and spatio-temporal modelling. The analyses in this research were motivated by the practical problems presented by the features of high-resolution, sparse remote-sensing image time series.

The research began with the exploratory analysis of the remote-sensing image time series of the Lake Victoria LSWT data, where drawbacks of investigations using conventional statistical methods were identified. To seek a more efficient way of analysing the data, statistical methods in the field of smoothing and functional data analysis were carefully studied. Considering the relatively high missing percentages of the remote-sensing images, the mixed model FPCA (MM-FPCA) was adopted to tackle this problem. However, the MM-FPCA did not account for the temporal dependence between the spatial images, which may be problematic in some situations. Therefore, methods for incorporating temporal dependence were explored. A new spatio-temporal model, SS-FPCA, consisting of a state space component (or dynamic component) and a FPCA component, was developed based on the dynamic spatio-temporal modelling framework. An estimation method based on the AECM algorithm was proposed and implemented using code developed in `R`[i]. A detailed investigation of the new model, including a simulation study on model performance, were carried out. The new model was shown to have the potential of identifying general spatial/temporal patterns in the image time series, improving data imputation while handling the task of dimension reduction. The MM-FPCA and SS-FPCA were applied to the sparse LSWT and Chlorophyll

[i]The `R` scripts for implementing the MM-SPCA, the Kalman filter with threshold and the SS-FPCA are all available on request.

data of Lake Victoria obtained from the AATSR and MERIS sensors on the European Space Agency's Envisat satellite.

6.1 General comments

6.1.1 On the MM-FPCA

The MM-FPCA was originally developed by James *et al.* (2000) for the analysis of longitudinal data. Zhou & Pan (2014) extended the methods to spatial data and showed that it is a powerful tool to extract principal components in irregularly sampled or sparse data sets. This thesis applied this method to the series of high resolution, sparse remote-sensing images of LSWT and Chl of Lake Victoria. The motivation was to conducting dimension reduction of the data set, while accounting for the missing observations. The implementation of the MM-FPCA using the EM algorithm was carried out with the `R` code developed based on package `fpca` (Peng & Paul, 2013). The estimation of the model using the EM algorithm has already been shown to be robust in earlier literature (James *et al.*, 2000, Rice & Wu, 2001), but the application of the model on sparse remote-sensing images has so far received little attention. Therefore, a simulation study was carried out to examine the influence of missing percentages and spatial missing patterns on the performance of the MM-FPCA. The results suggested that model estimates were robust and the RSS of the data reconstruction was relatively small, provided the missing percentage was moderate. However, statistical inference with respect to certain regions with substantive missingness throughout time needs to be interpreted with caution.

Implications The MM-FPCA was considered as the baseline model for the analysis of remote-sensing image time series. Its effectiveness in reducing data dimension and imputing missing observations has been demonstrated. First of all, FPCA provides two levels of dimension reduction, one through the functional data representation and the other through the truncation of functional PCs. The mixed effect model specification accounts for the missing observations by making use of the entire data set in the estimation of the overall mean function and the random effect of each individual subject. In the application to the Lake Victoria LSWT data in section 3.3, four PCs using a basis of degrees of freedom 49 were retained to reconstruct the original data set of size over 500,000, to a precision of RSS = 0.1207. The computation of the model took about one minute, which was far

more efficient than the thin-plate spline regression model with spatial covariance structure investigated in section 2.2. The results from the application to the Chl data may be less satisfactory due to the lack of smoothness of the data, but the MM-FPCA captured the general spatial patterns in the Chl images. Therefore, it is sufficient to say that the MM-FPCA offers an appealing solution to two challenges addressed at the beginning of the thesis, i.e. dimension reduction and missing data imputation.

The MM-FPCA also has the potential of identifying important spatial patterns in the data. These patterns are often reflected by the eigenfunctions of the leading PCs. This is a different way of describing the spatial patterns as compared to the covariogram models. To some extent, the FPCA offers more flexibility than the parameter models used in covariogram fitting, such as Gussian, exponential and Matérn models. The problem with FPCA, however, is that the interpretation of the PCs may not always be easy and the existence of leading PCs is not guaranteed. In the situation where the dominant spatial variation patterns can be extracted, the mathematical realizations of these patterns (i.e. PC scores) could be used for further analysis. In terms of remote-sensing image time series, the functional PC scores would reflect the strength of the patterns in different images. The changes in the scores can be interpreted as the evolution of the spatial pattern over time. It is possible to apply time series models to the scores to detect temporal trend, change points, etc.

The main drawback of the MM-FPCA is that it assumes independence for all individual subjects (images in this case). This may not be a problem for some remote-sensing image data after trend and seasonality are removed appropriately. However, neglecting temporal correlations in spatio-temporal data is generally not recommended, because it may lead to over or underestimated standard errors, resulting in inefficient statistical inference. It is essential to account for the temporal dependence where appropriate. Another issue that has not been tackled in this thesis is the shapes of the images, i.e. the irregular lake boundaries. The strategy in this thesis was to trim the grid to get a rectangle which contains as few redundant land pixels as possible and then apply a tensor spline basis. This was justified by the substantially larger retrieval errors towards the lake boundaries and the computational complexity of a shape adapted basis. The ideal solution would be to take care of the shapes of the images and the boundary uncertainties simultaneously. However, for the problem in this thesis, it might not be worthwhile modelling the shapes unless the retrieval errors in boundary pixels can be dealt with first.

6.1.2 On the SS-FPCA

The SS-FPCA model was proposed to account for the temporal dependence between individual images, which has not been considered in the MM-FPCA. The thesis approached this problem by developing a spatio-temporal model using FPCA based on the spatio-temporal random effect (STRE) model framework. In particular, the fixed mean function in the MM-FPCA was updated to a time-varying mean function (i.e. from $\boldsymbol{\Phi\beta}$ to $\boldsymbol{\Phi\beta}_t$), with a dynamic structure specified to the coefficient vector $\boldsymbol{\beta}_t$ to describe the spatio-temporal dynamic of the process. The proposed model falls into the category of dynamic spatio-temporal models (DSTM) as described in Cressie & Wikle (2011), which has the advantage of covering a wide range of spatio-temporal structures through flexible dynamic and hierarchical design.

The development of the SS-FPCA model was motivated by the STRE model presented in Cressie *et al.* (2010), from which valuable information with respect to the specification of the SS-FPCA model was acquired. The STRE model consists of a dynamic component accounting for the spatio-temporal dynamic of the data and a non-dynamic random component accounting for the remaining spatial variations. The proposed SS-FPCA model uses the same dynamic specification as in the STRE model, with a first-order random walk categorizing the evolution of the time varying coefficient, $\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{u}_t$. Whereas the unstructured random effect in the STRE was replaced by a truncated K-L expansion, $\sum_{p=1}^{P} \boldsymbol{\Phi\theta}_p \alpha_{pt}$, inheriting the same assumptions as in the MM-FPCA. This modification allows the non-dynamic random component to reflect more informative spatial patterns (in the form of the FPCA) than the somewhat vague 'remaining spatial variation'.

The estimation method of the SS-FPCA was motivated by that of the STRE model in Katzfuss & Cressie (2011), where the Kalman filter/smoother was used to estimate the system dynamic component and the EM algorithm was proposed to estimate the model parameters. However, due to the complexity of the SS-FPCA model, the standard EM algorithm was extended to the more powerful AECM algorithm. The AECM algorithm exploits several different data-augmentation schemes in the iterations to simplify the computation of the MLEs in a complex structured model. In this thesis, a 2-cycle AECM algorithm was developed such that analytical solutions are available for the MLEs of all parameters in the SS-FPCA model. The algorithm inherits the favourable properties of the classic Kalman filter/smoother and the robust MM-FPCA estimation method. The simulation study conducted on 1-dimensional data in section 5.2.1 suggested that the estimated results from the SS-FPCA model using the

2-cycle AECM algorithm was robust under different simulation scenarios, e.g. noise levels, sparsity and initial values.

The simulation study also showed that the SS-FPCA model was able to capture the dynamic structures and the FPCA components with a desirable precision. However, the estimation of the variances of different model components appeared to be biased. This phenomenon could be interpreted as the confounding between various model components, an issue common to many spatial and spatio-temporal models. A summary of explanations to this problem can be found in Hodges & Reich (2010). Further investigations are required to obtain a better understanding of the problem with respect to the SS-FPCA, so that the estimation precision can be improved. This might not be easy, because 'truth' is often unknown in reality and it is somehow impossible to judge whether the variances are under or over estimated.

Implications With the time-varying mean function describing the dynamic of the data, the SS-FPCA model was able to account for the spatio-temporal dependence in the remote-sensing image time series. It provided a solution to the potential limitation of the MM-FPCA, where individual functions were assumed to be independent. After removing the influence of the temporal structure through the dynamic component, the remainder of the variation explained by the FPCA component could be regarded as temporally independent. In other words, the fundamental assumption of the FPCA, or in fact any PCA would be fulfilled.

Different types of variation patterns can be identified from the SS-FPCA model. For instance, the spatio-temporal pattern may be interpreted from the system transition model, $\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{u}_t$, and its residual covariance matrix $\boldsymbol{H}$; pure spatial variation patterns may be displayed by the eigenfunctions of the functional PCs, with their evolution reflected by the PC scores. In view of these, the SS-FPCA model can be regarded as a suitable approach to the spatio-temporal modelling problem, which is of great interest in this research.

It is also believed that data imputation would be improved, as the SS-FPCA combines information from both space and time to enhance model fitting. The application of the SS-FPCA model to the Lake Victoria LSWT data suggested that, using the same basis dimension and expansion order, the RSS from the SS-FPCA model was smaller than its counterpart from the MM-FPCA (0.0811 v.s. 0.1207). Similar improvement was found in the application to the Lake Victoria Chl data. This means, a better data reconstruction can be achieved without sacrificing the degrees of freedom of the model. The pixel-wise RSS showed evidence of improvement in individual pixels, especially those with higher missing

percentages. This is no doubt a desirable feature concerning the objective of missing data imputation in this research.

One critical aspect of the SS-FPCA model is its relatively long computation time. The application to the Lake Victoria LSWT data took more than 5 hours, compared to only 1 minute using the MM-FPCA. It might be challenged whether it is worthwhile spending so much effort when the improvement in the results is not necessarily substantive. The answer based on the investigation in this thesis would be, 'Yes, in situations where there is evidence of temporal correlation'. Even if the temporal dependence is not as strong, there might still be gains in terms of the accuracy of data reconstruction. For example, if missing data imputation is crucial to the problem under study, then the effort spent in estimating the model might pay off. In addition, a detailed timer indicated that the majority of the computation time of the 2-cycle AECM algorithm was consumed by the Kalman filter, where high-dimensional matrix inversions might still be required. Therefore, if the code for the Kalman filter can be made more efficient, the computation time might be significantly reduced.

Another issue is related to the model inference. The parameters estimated using the EM-type algorithm do not come with standard errors or confidence intervals. So it is difficult to assess the uncertainty associated with the estimates. There are methods to obtain the confidence intervals, such as bootstrap and simulation, but the computation burden can be a problem for the SS-FPCA applied to a massive data set. This thesis proposed the use of asymptotic results, such as the inverse of the observed Fisher information matrix $\mathcal{I}(\Psi)$, which can be approximated using complete data score functions or the rate of convergence matrix. However, the precision of both approximations can be problematic. Based on the investigation, this thesis suggested that the asymptotic results to be used when bootstrap or simulation are considered as (computationally) infeasible.

6.2 Future work

The MM-FPCA and the SS-FPCA have been shown to be effective in the analysis of the sparse remote-sensing image time series in this thesis. However, they can still be improved in many ways.

An important future work with respect to the SS-FPCA model is to improve the identifiability of the model components, $\boldsymbol{\Phi}\boldsymbol{\beta}_t$, $\boldsymbol{\Phi}\boldsymbol{\Theta}\boldsymbol{\alpha}_t$ and $\boldsymbol{\epsilon}_t$. The simulation study in Chapter 5 showed that the model could under or overestimate the variances of these components. This problem has

been attributed to the spatial confounding, but detailed investigations are yet to be carried out. It would be interesting to explore the causes of the problem and to find some potential solutions, if possible.

In the meantime, various extensions can be made to the SS-FPCA model, by exploiting its flexible design, to describe different spatio-temporal structures.

(a) A straightforward extension of the SS-FPCA is to specify a more advanced model for the system dynamic, $\boldsymbol{\beta}_t = \boldsymbol{M}\boldsymbol{\beta}_{t-1} + \boldsymbol{u}_t$. This thesis used $\boldsymbol{M} = \boldsymbol{I}$ for a random walk process, but other designs of propagator matrix $\boldsymbol{M}$ can be used, such as the vector auto-regressive model and PDE/IDE based on physical/chemical laws. Reviews of different designs can be found in Cressie & Wikle (2011), Wikle & Hooten (2010). It is recommended that scientific based designs to be used, wherever possible, to inform the type of system dynamic. One thing to bear in mind is the identifiability of the parameters in matrix $\boldsymbol{M}$ and the computational cost. As the algorithm for estimating the SS-FPCA model is relatively complicated, the additional computation burden brought by a parameterized $\boldsymbol{M}$ needs to be handled carefully.

(b) Closely related to the above extension is the specification of the covariance matrix of the residuals in the system transition equation, $\boldsymbol{H} = \mathbf{Cov}[\boldsymbol{u}_t]$. No restraint was imposed on matrix $\boldsymbol{H}$ in this thesis, apart from the essential symmetric, non-negative definite requirement. However, it would be of interest to parameterize this matrix, so that specific spatio-temporal dynamic structure might be captured. Xu & Wikle (2007) presented some examples on parameterizing $\boldsymbol{H}$, such as the diagonal and the conditional auto-regressive settings. The structure imposed on $\boldsymbol{H}$ matrix may improve its estimation and inference.

(c) Another extension with respect to the specification of the SS-FPCA model is to use different bases for different model components. For example, $\boldsymbol{\Phi}_\beta$ for the state space component and $\boldsymbol{\Phi}_\xi$ for the FPCA component. This would be beneficial to the modelling of the variations in different spatial scales (Berliner *et al.*, 2000) or variations of very different natures, e.g. smooth trend and waves. It might also improve the identifiability of different model components and generate more meaningful results. The difficulty of this extension lies in the selection of the basis and the computation of the model, because using two bases might invalidate the simplification used in the computation of the current SS-FPCA.

(d) In this thesis, it was assumed that the components $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ in the SS-FPCA are independent and conditionally independent on data $\boldsymbol{Z}_{1:T}$. This assumption was meant for computational efficiency initially and was verified in section 4.3.2. In general, it is better to relax the conditional independence assumption due to the iterative nature of the estimation method. Unfortunately, this is non-trivial, because the derivation of the joint posterior distribution $\boldsymbol{\alpha}_t, \boldsymbol{\beta}_t | \boldsymbol{Z}_{1:T}$ is prone to be difficult. In order to get the conditional expectation of $\mathbf{E}[\boldsymbol{\alpha}_t \boldsymbol{\beta}_t | \boldsymbol{Z}_{1:T}]$, sampling techniques such as importance sampling, Metropolis-Hastings may be required. However, this could be complicated and the computation time might increase immensely. This extension would involve a great amount of effort, but the results could be influential.

(e) As discussed in section 6.1.1, it is sometimes helpful to use a basis that accounts for the shapes of the images. The applications of the MM-FPCA and SS-FPCA model in this thesis both used tensor spline basis on a rectangular grid. Alternatively, spline bases defined on triangulation as described in Ettinger *et al.* (2012), Zhou & Pan (2014) could be used to model the shapes of the images. This could be further extended to a manifold as in Lindgren *et al.* (2011) for modelling the global temperature surface. The soap film penalty proposed in Wood *et al.* (2008) could be another option. These methods should be able to minimize the influence of pixels that are irrelevant to the images in a regular grid. In terms of the remote-sensing lake measurements, it is also important to account for the higher retrieval uncertainties towards the lake boundaries. Otherwise, the efforts spent in modelling the shapes might end up having limited gains because the data close to the boundaries are far less reliable.

(f) Both the MM-FPCA and the SS-FPCA are designed for analysing smooth data. Their applications to data that involves discontinuities may be problematic. This was illustrated in the application of the two methods on the Lake Victoria Chl data, where the general patterns in the images were captured, but the edges of the discontinuity were blurred. This is not a desirable feature of the discontinuity is of main concern, e.g. the algae bloom, the boundaries of the pollution in the lakes. The MM-FPCA and the SS-FPCA may not be suitable in such situations. However, instead of abandoning the two methods completely, it may be interesting to investigate the potential modifications so that the discontinuities could be taken care of. Pre-processing of data that are not smooth by nature could be one approach. Alternatively, seeking a type of localized basis functions that automatically handles the discontinuities in the data could be another solution.

Finally, there is the issue of speeding up computation and monitoring convergence. Two directions to be considered in order to speed up the computation are, enhancing the `R` code and improving the design of the algorithm. As mentioned above, the majority of the computation time is taken up by the Kalman filter. Hence, it would be attractive if the code for the filter (for partially missing data) can be accelerated. As for monitoring convergence, one approach is to examine the structure of the rate of convergence matrix $\mathcal{J}(\Psi)$. McLachlan & Krishnan (1997) presents a summary of how the matrix and its eigenvalues can be interpreted to inform the convergence of the algorithm to a local maximum, saddle point, etc. Although obtaining $\mathcal{J}(\Psi)$ of the SS-FPCA model using numerical differentiation is computationally demanding, putting this idea into practice would certainly benefit the assessment of the model performance.

Appendix A

Appendix for Chapter 3

A.1 Hilbert-Schmidt operator, Mercer's theorem & Karhunen-Loève expansion

Hilbert-Schmidt operator. Let $\mathcal{D} \subset \mathbb{R}^n$ be a bounded domain. A function $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is a Hilbert-Schmidt kernel if

$$\int_{\mathcal{D}} \int_{\mathcal{D}} |k(x, y)|^2 dxdy < \infty \, .$$

That is, $k \in L^2(\mathcal{D} \times \mathcal{D})$, where $L^2(\centerdot)$ represents the L^2 norm. Define the integral operator, $\mathcal{K} : v \to \mathcal{K}v$, for $v \in L^2(\mathcal{D})$, as

$$[\mathcal{K}v](x) = \int_{\mathcal{D}} k(x, y)v(y)dy \, . \tag{A.1}$$

The mapping $\mathcal{K}$ is then called a Hilbert-Schmidt operator. It can be shown that $\mathcal{K} : L^2(\mathcal{D}) \to L^2(\mathcal{D})$. The eigenproblem of the functional PCA is based on this operator, with the covariance function, if exist (i.e. finite), being the Hilbert-Schmidt kernel.

Mercer's theorem. Let $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ be a continuous function on $\mathcal{D} = [a, b] \subset \mathbf{R}$. Suppose further that the corresponding Hilbert-Schmidt operator $\mathcal{K} : L^2(\mathcal{D}) \to L^2(\mathcal{D})$ of k is positive. If $\{\lambda_i\}$ and $\{e_i\}$ are the eigenvalues and the corresponding eigenvectors of $\mathcal{K}$, then for all $s, t \in \mathcal{D}$,

$$k(s, t) = \sum_{i=1}^{\infty} \lambda_i e_i(s)e_i(t) \, , \tag{A.2}$$

where convergence is absolute and uniform on $\mathcal{D} \times \mathcal{D}$.

Karhunen-Loève expansion. Define the covariance function of a zero mean continuous stochastic process $\{X_t\}_{t\in\mathcal{D}}$ as $V(s,t) = \mathbf{E}[X_sX_t]$ where $s,t \in \mathcal{D}$. Let $\mathcal{K}$ be defined as in equation (A.1) with $k(s,t) = V(s,t)$. Then using Mercer's theorem (A.2), it can be shown that $\mathcal{K}$ has a complete set of real eigenvalues $\{\lambda_i\}$ and eigenvfunctions $\{e_i\}$ in $L^2(\mathcal{D})$ such that $\mathcal{K}e_i = \lambda_i e_i$, and the stochastic process $\{X_t\}$ can be expanded as follows

$$X(t) = \sum_{i=1}^{\infty} f_i e_i(t), \qquad f_i = \int_{\mathcal{D}} X(t)e_i(t)dt\ . \tag{A.3}$$

The K-L expansion converges in mean square to the original process as $i \to \infty$ (Alexanderian, 2013). The expansion used in the mixed model for functional PCA is a truncated version of expansion (A.3) where the summation is operated on a finite set of eigenfunctions. The dimensions of the eigenfunction space is determined by the dimensions of the basis functions used in creating the functional data.

A.2 Additional information on the simulation study

The simulation study in section 3.2.3 considered 20 scenarios with different missing percentages and spatial missing patterns. Each scenario was repeated 200 times. The reason for this study to choose 200 replicates is the computational cost. The simulation study was run on the 'Dual 8 core HT Intel(R) Xeon(R) CPU E5-2640 v3' server. The computation time for each single replicates ranged from 19s to 60s. The total running time was 34.53 hours [i].

A re-sampling procedure is used to examine if the estimations from 200 replicates are robust (Law & Kelton, 1984). First, 50 replicates are randomly selected from 200 replicates. Then the same measures as in the simulation study (σ^2, MISE, etc) are computed using the 50 re-sampled repetitions. Repeat this for 10 times and compare the results to those estimated from 200 replicates. If 200 replicates are enough to give a robust estimation of a quantity θ using estimator $\hat{\theta}$, then the sample mean and sample variance should provide a good approximation to $\mathbf{E}[\hat{\theta}]$ and $\mathbf{Var}[\hat{\theta}]$. It is also presumed that the estimation from the re-sampled 50 replicates should have sample variance being approximately $\frac{1}{50}\mathbf{Var}[\hat{\theta}]$. Therefore, the following statistics from the 200 replicates are computed

$$\mathbf{E}[\hat{\theta}] \approx \bar{\hat{\theta}} = \frac{1}{200}\sum_{r=1}^{200}\hat{\theta}_r$$

[i] The computation can be made much faster now as the `R` code has been improved two times since then.

$$\mathbf{Var}[\hat{\theta}] \approx \frac{1}{200}\sum_{r=1}^{200}\left\{\hat{\theta}_r - \mathbf{E}[\hat{\theta}]\right\}^2 \approx \frac{1}{200}\sum_{r=1}^{200}\left[\hat{\theta}_r - \bar{\hat{\theta}}\right]^2$$

A 95% confidence interval of the estimator using the 50 re-sampled replicates, $\hat{\theta}_{50}$, can be constructed as

$$\left[\mathbf{E}[\hat{\theta}] \pm 1.96 \times \sqrt{\frac{1}{50}\mathbf{Var}[\hat{\theta}]}\right] \approx \left[\bar{\hat{\theta}} \pm 1.96 \times \sqrt{\frac{1}{50} \times \frac{1}{200}\sum_{i=1}^{200}\left[\hat{\theta}_r - \bar{\hat{\theta}}\right]^2}\right]$$

Check how many $\hat{\theta}_{50}$ out of M samples lie outside this interval. If the majority of $\hat{\theta}_{50}$ falls in the interval, then it can be concluded that the estimations of $\mathbf{E}[\hat{\theta}]$ and $\mathbf{Var}[\hat{\theta}]$ based on 200 replicates are reliable. In other words, 200 replicates are sufficient. The results show that all the estimates of $\hat{\sigma}^2_{50}$ from the re-sampled sets of 50 replicates fall within the 95% confidence intervals. The majority of the $\widehat{\text{MISE}}_{50}$ also fall within the 95% confidence intervals. Table A.1 shows the 95% confidence intervals of $\widehat{\text{MISE}}_{50}$, together with the sample means from four re-sampled sets of 50 replicates. Red colour indicates an excess of the 95% confidence interval. Therefore, it can be concluded that 200 replicates are sufficient to produce robust results to be used in inference.

As mentioned in section 3.2.3, the estimations of the coefficient vectors of the first eigenfunction, $\hat{\boldsymbol{\theta}}_1$, throughout the simulation scenarios are robust. This can be seen from the five panels in Figure A.1, which display the estimated $\hat{\boldsymbol{\theta}}_1$ from 200 replicates under five different missing specifications, all paired with the spatial variation scenario I. The point-wise 95% confidence intervals were also produced and plotted as the dashed curves in each panel. The general patterns of the values of elements in vector $\hat{\boldsymbol{\theta}}_1$ do not appear to vary substantively across the scenarios. The last panel, corresponding to 50% missing with spatial clusters, appear to be slightly different from the first four panel, but the difference is not distinctive. The same applies to the rest three spatial variation scenarios. Related plots are omitted here to avoid redundancy.

TABLE A.1: The 95% confidence intervals of $\widehat{\text{MISE}}_{50}$ from the re-sampled replicates of sample size 50, together with the sample means from four of the re-sampled data sets. The numbers in red indicate those exceeding the confidence intervals.

	spatial I $d = 1.5, \sigma^2_{ng} = 0.01$	**spatial II** $d = 1.5, \sigma^2_{ng} = 0.1$	**spatial III** $d = 1, \sigma^2_{ng} = 0.01$	**spatial IV** $d = 1, \sigma^2_{ng} = 0.1$
none	(0.2557, 0.2600)	(1.1494, 1.1550)	(0.3729, 0.3830)	(1.2610, 1.2721)
	0.2575	1.1537	0.3790	1.2635
	0.2576	1.1531	**0.3727**	1.2664
	0.2585	1.1514	0.3780	1.2633
	0.2586	1.1530	0.3782	1.2664
no pattern 30%	(0.2562, 0.2604)	(1.1525, 1.1581)	(0.3737, 0.3838)	(1.2683, 1.2793)
	0.2580	1.1567	0.3794	1.2722
	0.2580	1.1562	**0.3731**	1.2745
	0.2590	1.1545	0.3783	**1.2678**
	0.2591	1.1561	0.3794	1.2760
no pattern 50%	(0.2579, 0.2619)	(1.1597, 1.1655)	(0.3798, 0.3901)	(1.2879, 1.2986)
	0.2599	1.1638	0.3882	1.2911
	0.2601	1.1633	0.3818	1.2902
	0.2610	1.1607	0.3864	1.2899
	0.2602	1.1631	0.3859	1.2944
pattern 30%	(0.2580, 0.2633)	(1.1590, 1.1664)	(0.4016, 0.4117)	(1.3064, 1.3154)
	0.2617	1.1627	0.4068	1.3085
	0.2609	1.1659	0.4056	1.3087
	0.2605	1.1612	0.4090	1.3089
	0.2604	1.1622	0.4081	1.3113
pattern 50%	(0.3002, 0.3084)	(1.2331, 1.2449)	(0.4850, 0.4984)	(1.4226, 1.4366)
	0.3079	1.2386	0.4901	1.4326
	0.3019	**1.2452**	0.4885	1.4250
	0.3059	**1.2318**	0.4948	1.4306
	0.3063	1.2386	0.4908	1.4314

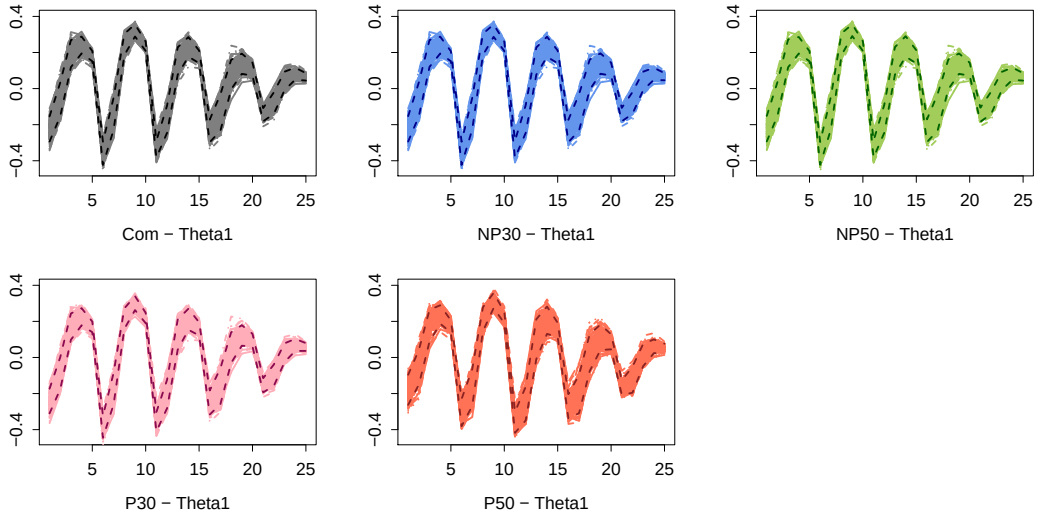

FIGURE A.1: The estimated coefficient vector of the first eigenfunction $\hat{\boldsymbol{\theta}}_1$ from 200 replicates under spatial variation scenario I. From top left to bottom middle are no missing, missing 30%, 50% without pattern, missing 30%, 50% with pattern respectively. The dashed curves indicate the point-wise 95% confidence intervals.

Appendix B

Appendix for Chapter 4

B.1 The target functions of the 2-cycle AECM algorithm

This section provides some details on the E-step functions of the 2-cycle AECM algorithm developed in section 4.4.2. The computational form of target function $\mathcal{Q}^{[1]}(\Psi^{[1]}; \Psi^{(it)})$ is

$$
\begin{aligned}
&\mathcal{Q}^{[1]}\left(\Psi^{[1]}\,;\,\Psi^{(it-1)}\right) \qquad\qquad (B.1)\\
&= \mathbf{E}\left[-2\mathcal{L}\left(\Psi^{[1]}; \boldsymbol{Z}^{[1]}, \widetilde{\Psi}^{[1]}\right)\middle|\, \boldsymbol{Z}_{1:T}, \Psi^{(it-1)}\right]\\
&= \mathbf{E}\left[-2\log f\left(\boldsymbol{Z}_{1:T}, \boldsymbol{\beta}_{0:T}\,;\, \boldsymbol{H}, \boldsymbol{\Theta}^{(it-1)}, \boldsymbol{\Lambda}^{(it-1)}, \sigma^{2(it-1)}\right)\middle|\, \boldsymbol{Z}_{1:T}, \Psi^{(it-1)}\right]\\
&= \sum_{t=1}^{T}\left[\log\left(\left|\boldsymbol{G}_t^{(it)}\right|\right) + \operatorname{tr}\left\{\left(\boldsymbol{G}_t^{(it)}\right)^{-1}\left(\boldsymbol{\Phi}_t\boldsymbol{B}_{t|T}\boldsymbol{\Phi}_t^{\top} + \left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_{t|T}\right)\left(\boldsymbol{Z}_t - \boldsymbol{\Phi}_t\boldsymbol{\beta}_{t|T}\right)^{\top}\right)\right\}\right]\\
&\quad + \log(|\boldsymbol{B}_0|) + \operatorname{tr}\left\{\boldsymbol{B}_0^{-1}\left(\boldsymbol{B}_{0|T} + \left(\boldsymbol{\beta}_{0|T} - \boldsymbol{\beta}\right)\left(\boldsymbol{\beta}_{0|T} - \boldsymbol{\beta}\right)^{\top}\right)\right\}\\
&\quad + T\log(|\boldsymbol{H}|) + \operatorname{tr}\left\{\boldsymbol{H}^{-1}\left(\boldsymbol{V}_{11} - 2\boldsymbol{V}_{10} + \boldsymbol{V}_{00}\right)\right\} + \text{constant}
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{G}_t^{(it)} &= \boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)}\boldsymbol{\Theta}^{(it-1)\top}\boldsymbol{\Phi}_t^{\top} + \sigma^{2(it-1)}\boldsymbol{I}\\
\boldsymbol{V}_{11} &= \sum_{t=1}^{T}\left(\boldsymbol{B}_{t|T} + \boldsymbol{\beta}_{t|T}\boldsymbol{\beta}_{t|T}^{\top}\right)\\
\boldsymbol{V}_{00} &= \sum_{t=1}^{T}\left(\boldsymbol{B}_{t-1|T} + \boldsymbol{\beta}_{t-1|T}\boldsymbol{\beta}_{t-1|T}^{\top}\right)\\
\boldsymbol{V}_{10} &= \sum_{t=1}^{T}\left(\boldsymbol{B}_{t,t-1|T} + \boldsymbol{\beta}_{t|T}\boldsymbol{\beta}_{t-1|T}^{\top}\right)\ .
\end{aligned}
$$

The smoothed states $\{\boldsymbol{\beta}_{t|T}\}_{t=1}^{T}$ are obtained by Kalman filter/smoother with $\boldsymbol{H}^{(it-1)}$ and $\boldsymbol{G}_t^{(it)}$. This is essentially the same as the target function of the EM algorithm of the reduced rank state space model in Cressie & Wikle (2011), with propagator matrix $\boldsymbol{M} = \boldsymbol{I}$.

The computational form of target function $\mathcal{Q}^{[2]}(\Psi^{[2]}; \Psi^{(it,it-1)})$ is

$$
\begin{aligned}
&\mathcal{Q}^{[2]}\left(\Psi^{[2]}\,;\,\Psi^{(it,it-1)}\right) \qquad\qquad (B.2)\\
=&\,\mathbf{E}\left[-2\mathcal{L}\left(\Psi^{[2]};\boldsymbol{Z}^{[2]},\widetilde{\Psi}^{[2]}\right)\middle|\,\boldsymbol{Z}_{1:T},\Psi^{(it,it-1)}\right]\\
=&\,\mathbf{E}\left[-2\log f\left(\boldsymbol{Z}_{1:T},\boldsymbol{\beta}_{0:T},\boldsymbol{\alpha}_{1:T}\,;\,\boldsymbol{\Theta},\boldsymbol{\Lambda},\sigma^2,\boldsymbol{H}^{(it)}\right)\middle|\,\boldsymbol{Z}_{1:T},\Psi^{(it,it-1)}\right]\\
=&\sum_{t=1}^{T}\Bigg[n_t\log\left(\sigma^2\right)+\frac{1}{\sigma^2}\mathrm{tr}\left\{\boldsymbol{\Phi}_t\boldsymbol{B}_{t|T}\boldsymbol{\Phi}_t^\top+\left(\boldsymbol{Z}_t-\boldsymbol{\Phi}_t\boldsymbol{\beta}_{t|T}\right)\left(\boldsymbol{Z}_t-\boldsymbol{\Phi}_t\boldsymbol{\beta}_{t|T}\right)^\top\right\}\\
&+\frac{1}{\sigma^2}\mathrm{tr}\left\{\boldsymbol{\Phi}_t\boldsymbol{\Theta}\widehat{\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^\top}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top\right\}-\frac{2}{\sigma^2}\mathrm{tr}\left\{\boldsymbol{\Phi}_t\boldsymbol{\Theta}\hat{\boldsymbol{\alpha}}_t\boldsymbol{Z}_t^\top-\boldsymbol{\Phi}_t\boldsymbol{\Theta}\widehat{\boldsymbol{\alpha}_t\boldsymbol{\beta}_t^\top}\boldsymbol{\Phi}_t^\top\right\}\Bigg]\\
&+T\log\left(\left|\boldsymbol{H}^{(it)}\right|\right)+\mathrm{tr}\left\{\left(\boldsymbol{H}^{(it)}\right)^{-1}\left(\boldsymbol{V}_{11}-2\boldsymbol{V}_{10}+\boldsymbol{V}_{00}\right)\right\}\\
&+\log(|\boldsymbol{B}_0|)+\mathrm{tr}\left\{\boldsymbol{B}_0^{-1}\left(\boldsymbol{B}_{0|T}+\left(\boldsymbol{\beta}_{0|T}-\boldsymbol{\beta}\right)\left(\boldsymbol{\beta}_{0|T}-\boldsymbol{\beta}\right)^\top\right)\right\}\\
&+T\log(|\boldsymbol{\Lambda}|)+\sum_{t=1}^{T}\mathrm{tr}\left\{\boldsymbol{\Lambda}^{-1}\widehat{\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^\top}\right\}+\text{ constant }.
\end{aligned}
$$

The smoothed states $\{\boldsymbol{\beta}_{t|T}\}_{t=1}^{T}$ are obtained with $\boldsymbol{H}^{(it)}$ and $\boldsymbol{G}_t^{(it)}$, which are obtained from the previous cycle. The estimation of $\hat{\boldsymbol{\alpha}}_t$ and $\widehat{\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^\top}$ are based on the conditional distribution of $\boldsymbol{\alpha}_t|\boldsymbol{Z}_{1:T},\Psi^{(it,it-1)}$, with the influence of $\boldsymbol{\beta}_t$ accounted for. $\widehat{\boldsymbol{\alpha}_t\boldsymbol{\beta}_t^\top}$ is estimated under the independent assumption of $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$.

Specifically, the conditional distribution of $\boldsymbol{\alpha}_t|\boldsymbol{Z}_{1:T}$ under the current parameter estimates $\Psi^{(it,it-1)}$ can be derived in a similar way as that of the MM-FPCA model. However, $\boldsymbol{\beta}_t$ in the SS-FPCA is no longer the fixed effect coefficient vector, but the random coefficient of the state space component. Hence, its influence needs to be accounted for appropriately.

(a) The covariance matrix of $\boldsymbol{Z}_t$ is now $\boldsymbol{\Phi}\mathbf{Cov}[\boldsymbol{\beta}_t]\boldsymbol{\Phi}_t^\top+\boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top+\sigma^2\boldsymbol{I}$.

(b) As $\boldsymbol{\beta}_{t|T}$ is obtained knowing only the variance, not the value, of $\boldsymbol{\alpha}_t$, giving information of $\boldsymbol{Z}_{1:T},\Psi^{(it,it-1)}$ is sufficient to determine the values of $\boldsymbol{\beta}_{t|T}$ and $\boldsymbol{B}_{t|T}$. Meanwhile, $\boldsymbol{\alpha}_t$ is assumed to be independent from $\boldsymbol{\beta}_t$ and is only related to $\boldsymbol{Z}_t$, not the observations at the rest of the time points. In consequence, the following dependence can be deduced,

$$
\left[\boldsymbol{\alpha}_t\middle|\boldsymbol{Z}_{1:T},\Psi^{(it,it-1)}\right]=\left[\boldsymbol{\alpha}_t\middle|\boldsymbol{Z}_{1:T},\boldsymbol{\beta}_{t|T},\boldsymbol{B}_{t|T},\Psi^{(it,it-1)}\right]=\left[\boldsymbol{\alpha}_t\middle|\boldsymbol{Z}_t,\boldsymbol{\beta}_{t|T},\boldsymbol{B}_{t|T},\Psi^{(it,it-1)}\right]\,.
$$

(c) It is relatively easy to figure out the conditional distribution of $\boldsymbol{\alpha}_t|\boldsymbol{Z}_t$, given the multivariate normal distribution of $(\boldsymbol{\alpha}_t, \boldsymbol{Z}_t)^\top$, which is

$$\mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \boldsymbol{\Phi}_t\mathbf{E}[\boldsymbol{\beta}_t] \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Lambda} & \boldsymbol{\Lambda}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top \\ \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\Lambda} & \boldsymbol{\Phi}_t\mathbf{Cov}[\boldsymbol{\beta}_t]\boldsymbol{\Phi}_t^\top + \boldsymbol{\Phi}_t\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^\top\boldsymbol{\Phi}_t^\top + \sigma^2\boldsymbol{I} \end{pmatrix}\right),$$

but the task remains to find $\mathbf{E}[\boldsymbol{\beta}_t]$ and $\mathbf{Cov}[\boldsymbol{\beta}_t]$. This can be done using the property of double expectation in basic probability, resulting in $\mathbf{E}[\boldsymbol{\beta}_t] = \mathbf{E}[\boldsymbol{\beta}_{t|T}]$ and $\mathbf{Cov}[\boldsymbol{\beta}_t] = \mathbf{E}[\boldsymbol{B}_{t|T}]$. That is, the smoothed state $\boldsymbol{\beta}_{t|T}$ and its covariance matrix $\boldsymbol{B}_{t|T}$ could be used as the estimates of $\mathbf{E}[\boldsymbol{\beta}_t]$ and $\mathbf{Cov}[\boldsymbol{\beta}_t]$.

Putting all the information in (a), (b) and (c) together gives the following distributional results,

$$\begin{aligned}
\mathbf{E}\left[\boldsymbol{\alpha}_t|\boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}\right] &= \left(\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)}\right)^\top \left(\boldsymbol{\Sigma}^{(it)}\right)^{-1}\left(\boldsymbol{Z}_t - \Phi\boldsymbol{\beta}_{t|T}\right), \\
\mathbf{Cov}\left[\boldsymbol{\alpha}_t|\boldsymbol{Z}_{1:T}, \Psi^{(it,it-1)}\right] &= \boldsymbol{\Lambda}^{(it-1)} - \left(\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)}\right)^\top \left(\boldsymbol{\Sigma}^{(it)}\right)^{-1}\boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)},
\end{aligned}$$

where $\boldsymbol{\Sigma}^{(it)} = \boldsymbol{\Phi}_t\boldsymbol{B}_{t|T}\boldsymbol{\Phi}_t^\top + \boldsymbol{\Phi}_t\boldsymbol{\Theta}^{(it-1)}\boldsymbol{\Lambda}^{(it-1)}\boldsymbol{\Theta}^{(it-1)\top}\boldsymbol{\Phi}_t^\top + \sigma^{2(it-1)}\boldsymbol{I}$. Using the above conditional expectation and variance, both $\hat{\boldsymbol{\alpha}}_t$ and $\widehat{\boldsymbol{\alpha}_t\boldsymbol{\alpha}_t^\top}$ can be obtained.

B.2 The Kalman filter/smoother in the 2-cycle AECM algorithm

This section provides some information on the matrix operations used for simplifying the computation of the missing observation Kalman filter. First recall that the Kalman gain in a standard Kalman filtering process is defined as

$$\boldsymbol{K}_t = \boldsymbol{B}_{t|t-1}\boldsymbol{\Phi}^\top\left(\boldsymbol{\Phi}\boldsymbol{B}_{t|t-1}\boldsymbol{\Phi}^\top + \boldsymbol{G}\right)^{-1}. \tag{B.3}$$

The part that need to be simplified in actual computation is the inverse $\left(\boldsymbol{\Phi}\boldsymbol{B}_{t|t-1}\boldsymbol{\Phi}^\top + \boldsymbol{G}\right)^{-1}$. This inverse could be of very high dimension even in the reduced rank Kalman filtering/smoothing algorithm. One important matrix identity used in the computation is the Woodbury identity

$$\left(\boldsymbol{A} + \boldsymbol{C}\boldsymbol{D}\boldsymbol{C}^\top\right)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{C}\left(\boldsymbol{D}^{-1} + \boldsymbol{C}^\top\boldsymbol{A}^{-1}\boldsymbol{C}\right)^{-1}\boldsymbol{C}^\top\boldsymbol{A}^{-1}, \tag{B.4}$$

which is very attractive if the $\boldsymbol{A}$ is a diagonal matrix, or if the inverse of $\boldsymbol{A}$ and $\boldsymbol{D}^{-1} + \boldsymbol{C}^\top \boldsymbol{A}^{-1}\boldsymbol{C}$ can be obtained in an easy way.

In cycle 1 of the AECM algorithm, $\boldsymbol{G} = \boldsymbol{\Phi\Theta\Lambda\Theta}^\top\boldsymbol{\Phi}^\top + \sigma^2\boldsymbol{I}$ in the Kalman gain (B.3), giving

$$\begin{aligned}\left(\boldsymbol{\Phi}\boldsymbol{B}_{t|t-1}\boldsymbol{\Phi}^\top + \boldsymbol{G}\right)^{-1} &= \left(\boldsymbol{\Phi}\boldsymbol{B}_{t|t-1}\boldsymbol{\Phi}^\top + \boldsymbol{\Phi\Theta\Lambda\Theta}^\top\boldsymbol{\Phi}^\top + \sigma^2\boldsymbol{I}\right)^{-1} \\ &= \left\{\boldsymbol{\Phi}\left(\boldsymbol{B}_{t|t-1} + \boldsymbol{\Theta\Lambda\Theta}^\top\right)\boldsymbol{\Phi}^\top + \sigma^2\boldsymbol{I}\right\}^{-1} \\ &= \frac{1}{\sigma^2}\left\{\boldsymbol{\Phi}\left[\frac{1}{\sigma^2}\left(\boldsymbol{B}_{t|t-1} + \boldsymbol{\Theta\Lambda\Theta}^\top\right)\right]\boldsymbol{\Phi}^\top + \boldsymbol{I}\right\}^{-1} .\end{aligned} \tag{B.5}$$

Applying the matrix identity (B.4) directly to the above and using the fact that $\boldsymbol{\Phi}^\top\boldsymbol{\Phi} = I$ when the observations at time t are complete gives

$$\begin{aligned}&\frac{1}{\sigma^2}\left\{\boldsymbol{\Phi}\left[\frac{1}{\sigma^2}\left(\boldsymbol{B}_{t|t-1} + \boldsymbol{\Theta\Lambda\Theta}^\top\right)\right]\boldsymbol{\Phi}^\top + \boldsymbol{I}\right\}^{-1} \\ =&\frac{1}{\sigma^2}\left\{\boldsymbol{I} - \boldsymbol{\Phi}\left[\sigma^2\left(\boldsymbol{B}_{t|t-1} + \boldsymbol{\Theta\Lambda\Theta}^\top\right)^{-1} + \boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right]^{-1}\boldsymbol{\Phi}^\top\right\} \\ =&\frac{1}{\sigma^2}\left\{\boldsymbol{I} - \boldsymbol{\Phi}\left[\sigma^2\left(\boldsymbol{B}_{t|t-1} + \boldsymbol{\Theta\Lambda\Theta}^\top\right)^{-1} + \boldsymbol{I}\right]^{-1}\boldsymbol{\Phi}^\top\right\} .\end{aligned} \tag{B.6}$$

This is very simple to compute and involves only low dimension (equals to the degrees of freedom of the basis function K) matrix inversion.

The situation with respect to the partially missing data case is more complicated. Only limited simplification can be achieved, which is determined by the proportion of missing observations at each time point. However, it is still better than computing the high-dimensional matrix inverse directly. Define $\boldsymbol{\Phi}_{yes}$ as the matrix consists of the rows in $\boldsymbol{\Phi}$ corresponding to the observed locations, $\boldsymbol{\Phi}_{no}$ as the matrix consists of the rows in $\boldsymbol{\Phi}$ corresponding to the unobserved locations. Define $\boldsymbol{\Phi}_{obs} = (\boldsymbol{\Phi}_{yes}^\top \ \boldsymbol{0})^\top$ which has the same dimension as matrix $\boldsymbol{\Phi}$. Also define

$$\boldsymbol{G}_{obs} = \begin{pmatrix} \boldsymbol{G}_{yes} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}_{no} \end{pmatrix} ,$$

where $\boldsymbol{G}_{yes}$ and $\boldsymbol{G}_{no}$ are matrices consist of rows and columns in matrix $\boldsymbol{G}$, corresponding to the observed and unobserved locations respectively. Then the Kalman gain for a time point t with missing observations can be written as (Shumway & Stoffer, 2006)

$$\boldsymbol{K}_t = \boldsymbol{B}_{t|t-1}\boldsymbol{\Phi}_{obs}^\top\left(\boldsymbol{\Phi}_{obs}\boldsymbol{B}_{t|t-1}\boldsymbol{\Phi}_{obs}^\top + \boldsymbol{G}_{obs}\right)^{-1} . \tag{B.7}$$

Since

$$\boldsymbol{\Phi}_{obs}\boldsymbol{B}_{t|t-1}\boldsymbol{\Phi}_{obs}^{\top} = \begin{pmatrix} \boldsymbol{\Phi}_{yes} \\ \mathbf{0} \end{pmatrix} \boldsymbol{B}_{t|t-1} \begin{pmatrix} \boldsymbol{\Phi}_{yes}^{\top} & \mathbf{0} \end{pmatrix}$$

and

$$\boldsymbol{G}_{obs} = \begin{pmatrix} \boldsymbol{\Phi}_{yes}\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^{\top}\boldsymbol{\Phi}_{yes}^{\top} + \sigma^2\boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}_{no}\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^{\top}\boldsymbol{\Phi}_{no}^{\top} + \sigma^2\boldsymbol{I} \end{pmatrix}$$

the matrix to be inverted in the Kalman gain (B.7) becomes

$$\begin{aligned} &\boldsymbol{\Phi}_{obs}\boldsymbol{B}_{t|t-1}\boldsymbol{\Phi}_{obs}^{\top} + \boldsymbol{G}_{obs} \\ &= \begin{pmatrix} \boldsymbol{\Phi}_{yes} \\ \mathbf{0} \end{pmatrix} \left(\boldsymbol{B}_{t|t-1} + \boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^{\top}\right) \begin{pmatrix} \boldsymbol{\Phi}_{yes}^{\top} & \mathbf{0} \end{pmatrix} + \begin{pmatrix} \sigma^2\boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}_{no}\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^{\top}\boldsymbol{\Phi}_{no}^{\top} + \sigma^2\boldsymbol{I} \end{pmatrix} \\ &= \boldsymbol{\Phi}_{obs}\left(\boldsymbol{B}_{t|t-1} + \boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^{\top}\right)\boldsymbol{\Phi}_{obs}^{\top} + \begin{pmatrix} \sigma^2\boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}_{no}\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^{\top}\boldsymbol{\Phi}_{no}^{\top} + \sigma^2\boldsymbol{I} \end{pmatrix}. \end{aligned} \tag{B.8}$$

Applying the Woodbury identity (B.4) again, the inverse of $\boldsymbol{G}_{obs}$ can be resolved using the inverse of $\boldsymbol{\Phi}_{no}\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^{\top}\boldsymbol{\Phi}_{no}^{\top} + \sigma^2\boldsymbol{I}$, which is of much lower dimension than $\boldsymbol{G}_{obs}$ if the missing percentage is low. In particular, the entire inverse is

$$\begin{aligned} &\left(\boldsymbol{\Phi}_{obs}\boldsymbol{B}_{t|t-1}\boldsymbol{\Phi}_{obs}^{\top} + \boldsymbol{G}_{obs}\right)^{-1} \\ &= \frac{1}{\sigma^2}\left\{\boldsymbol{O}^{-1} - \boldsymbol{O}^{-1}\boldsymbol{\Phi}_{obs}\left[\sigma^2\left(\boldsymbol{B}_{t|t-1} + \boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^{\top}\right)^{-1} + \boldsymbol{\Phi}_{obs}\boldsymbol{O}^{-1}\boldsymbol{\Phi}_{obs}^{\top}\right]\boldsymbol{\Phi}_{obs}^{\top}\boldsymbol{O}^{-1}\right\}, \end{aligned} \tag{B.9}$$

where

$$\boldsymbol{O} = \begin{pmatrix} \sigma^2\boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Phi}_{no}\boldsymbol{\Theta}\boldsymbol{\Lambda}\boldsymbol{\Theta}^{\top}\boldsymbol{\Phi}_{no}^{\top} + \sigma^2\boldsymbol{I} \end{pmatrix}$$

The above results are used in the computation of the missing data Kalman filter in the AECM algorithm. The `R` functions developed for the models in this thesis were based on these results.

Appendix C

Appendix for Chapter 5

C.1 Convergence properties of the AECM algorithm

The convergence theorems with respect to the AECM algorithm are derived in Meng & Van Dyk (1997). The result comes from the convergence theorems of the GEM algorithm, with the additional space-filling condition.

Theorem C.1. Any AECM sequence increases (or maintains) $\mathcal{L}(\Psi)$ at every cycle and thus increases (or maintains) $\mathcal{L}(\Psi)$ at every iteration.

Theorem C.2. In addition to the regularity conditions in Wu (1983), suppose that (a) all the CM-steps are unique, (b) the AECM iteration mapping, $\mathcal{F}: \Psi^{(it)} \to \Psi^{(it+1)}$, does not depend on it. Then all the limit points of an AECM sequence $\{\Psi^{(it)}\}$ are stationary points of $\mathcal{L}(\Psi)$.

Theorem C.3. Suppose that the AECM iteration mapping is a composition of C fixed cycle mappings, all the CM-steps satisfy the Lagrange multiplier equations and $\Psi^{(it,c-1+\frac{s}{S^{[c]}})} \to \Psi^*$ as $it \to \infty$. Then the rate matrix of convergence of the AECM iteration is

$$\mathcal{J}^{AECM}(\Psi) = \prod_{c=1}^{C} \left\{ \boldsymbol{I} - \mathcal{I}(\Psi)\mathcal{I}_c^{[c]}(\Psi)^{-1} \left(I - \prod_{s=1}^{S^{[c]}} \boldsymbol{P}_s^{[c]} \right) \right\}, \tag{C.1}$$

$$\boldsymbol{P}_s^{[c]} = \bigtriangledown_s^{[c]} \left\{ \left(\bigtriangledown_s^{[c]} \right)^\top \mathcal{I}_c^{[c]}(\Psi)^{-1} \left(\bigtriangledown_s^{[c]} \right) \right\}^{-1} \left(\bigtriangledown_s^{[c]} \right)^\top \mathcal{I}_c^{[c]}(\Psi)^{-1}$$

$$\bigtriangledown_s^{[c]} = \bigtriangledown g_s^{[c]}(\Psi^*)\,.$$

Based on equation (C.1), it can be deducted that the observed information matrix $\mathcal{I}(\Psi)$ for a 2-cycle AECM algorithm (i.e. $C = 2$) is the solution of the matrix quadratic equation

$$\boldsymbol{D} = \boldsymbol{I} - \boldsymbol{X}\boldsymbol{A} - \boldsymbol{X}\boldsymbol{B} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}\boldsymbol{B} \,, \tag{C.2}$$

where $\boldsymbol{X} = \mathcal{I}(\Psi)$ is the unknown matrix to be solved, $\boldsymbol{D} = \mathcal{J}^{AECM}(\Psi)$ is the rate matrix of convergence and

$$\begin{aligned}
\boldsymbol{A} &= \mathcal{I}(\Psi)\mathcal{I}_c^{[1]}(\Psi)^{-1}\left(\boldsymbol{I} - \prod_{s=1}^{S^{[1]}} \boldsymbol{P}_s^{[1]}\right) \\
\boldsymbol{B} &= \mathcal{I}(\Psi)\mathcal{I}_c^{[2]}(\Psi)^{-1}\left(\boldsymbol{I} - \prod_{s=1}^{S^{[2]}} \boldsymbol{P}_s^{[2]}\right) \,.
\end{aligned}$$

Under the condition that $\boldsymbol{B}$ is invertible, equation (C.2) can be written as

$$(\boldsymbol{D} - \boldsymbol{I})\boldsymbol{B}^{-1} = \boldsymbol{X}\boldsymbol{A}\boldsymbol{X} - \boldsymbol{X} - \boldsymbol{X}\boldsymbol{A}\boldsymbol{B}^{-1} \,.$$

This is essentially the same as

$$\boldsymbol{D}^* = \boldsymbol{X}\boldsymbol{A}\boldsymbol{X} + \boldsymbol{B}^*\boldsymbol{X} + \boldsymbol{X}\boldsymbol{C}^* \,, \tag{C.3}$$

with $\boldsymbol{D}^* = (\boldsymbol{D} - \boldsymbol{I})\boldsymbol{B}^{-1}$, $\boldsymbol{B}^* = -\boldsymbol{I}$ and $\boldsymbol{C}^* = -\boldsymbol{A}\boldsymbol{B}^{-1}$. The solution of equation (C.3) exists if matrices $\boldsymbol{A}, \boldsymbol{B}^*, \boldsymbol{C}^*$ and $\boldsymbol{D}^*$ satisfy certain conditions These conditions and detailed solutions were described in Shurbet *et al.* (1974).

C.2 Derivatives of the complete data log-likelihood of the SS-FPCA

First of all, the information matrix with respect to the complete data log-likelihood of the SS-FPCA is block diagonal. In particular, the corresponding second derivative matrix takes the following form,

$$\begin{pmatrix}
\frac{\partial^2 \mathcal{L}_c(\cdot)}{\partial\sigma^2\partial\sigma^2} & \frac{\partial^2 \mathcal{L}_c(\cdot)}{\partial\sigma^2\partial\boldsymbol{\Theta}^\top} & \boldsymbol{0} & \boldsymbol{0} \\
\frac{\partial^2 \mathcal{L}_c(\cdot)}{\partial\boldsymbol{\Theta}\partial\sigma^2} & \frac{\partial^2 \mathcal{L}_c(\cdot)}{\partial\boldsymbol{\Theta}\partial\boldsymbol{\Theta}^\top} & \boldsymbol{0} & \boldsymbol{0} \\
\boldsymbol{0} & \boldsymbol{0} & \frac{\partial^2 \mathcal{L}_c(\cdot)}{\partial\boldsymbol{\Lambda}\partial\boldsymbol{\Lambda}^\top} & \boldsymbol{0} \\
\boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \frac{\partial^2 \mathcal{L}_c(\cdot)}{\partial\boldsymbol{H}\partial\boldsymbol{H}^\top}
\end{pmatrix} . \tag{C.4}$$

The component $\frac{\partial^2 \mathcal{L}_c(\cdot)}{\partial \boldsymbol{\Theta} \partial \boldsymbol{\Theta}^\top}$ consists of $\frac{\partial^2 \mathcal{L}_c(\cdot)}{\partial \boldsymbol{\theta}_p \partial \boldsymbol{\theta}_p^\top}$, $p = 1, \cdots, P$, and $\frac{\partial^2 \mathcal{L}_c(\cdot)}{\partial \boldsymbol{\theta}_p \partial \boldsymbol{\theta}_q^\top}$, $p \neq q$. The component $\frac{\partial^2 \mathcal{L}_c(\cdot)}{\partial \sigma^2 \partial \boldsymbol{\Theta}^\top}$ consists of $\frac{\partial^2 \mathcal{L}_c(\cdot)}{\partial \sigma^2 \partial \boldsymbol{\theta}_p^\top}$, $p = 1, \cdots, P$. In the approximation using score functions, however, only the first derivatives will be used.

To obtain the first and second derivatives with respect to matrix $\boldsymbol{H}$, first introduce some matrix derivative results used in the derivation,

$$\frac{\partial \log(|\boldsymbol{H}|)}{\partial \boldsymbol{H}} = \mathbf{vec}\left(\boldsymbol{H}^{-\top}\right)^\top \tag{C.5}$$

$$\frac{\partial \mathbf{vec}\left(\boldsymbol{H}^{-\top}\right)^\top}{\partial \boldsymbol{H}} = \frac{\partial \boldsymbol{H}^{-1}}{\partial \boldsymbol{H}} = -\boldsymbol{H}^{-\top} \otimes \boldsymbol{H}^{-1} \tag{C.6}$$

$$\frac{\partial \boldsymbol{A}^\top}{\partial \boldsymbol{A}} = \boldsymbol{T}_{nm} \tag{C.7}$$

$$\frac{\partial \boldsymbol{A}\boldsymbol{H}}{\partial \boldsymbol{H}} = \boldsymbol{I} \otimes \boldsymbol{A} \tag{C.8}$$

$$\frac{\partial \boldsymbol{A}\boldsymbol{H}\boldsymbol{B}}{\partial \boldsymbol{H}} = \boldsymbol{B}^\top \otimes \boldsymbol{A} \tag{C.9}$$

$$\frac{\partial (\boldsymbol{A} \otimes \boldsymbol{B})}{\partial \boldsymbol{B}} = (\boldsymbol{I}_m \otimes \boldsymbol{T}_{qn} \otimes \boldsymbol{I}_p)\,(\mathbf{vec}(\boldsymbol{A}) \otimes \boldsymbol{I}_{pq}) \tag{C.10}$$

where matrix $\boldsymbol{A}$ is of dimension $m \times n$, matrix $\boldsymbol{B}$ is of dimension $p \times q$ and $\boldsymbol{T}_{qn}$ is a permutation matrix satisfying $\mathbf{vec}(\boldsymbol{X}^\top) = \boldsymbol{T}_{qn}\mathbf{vec}(\boldsymbol{X})$ with $\boldsymbol{X}$ an $n \times q$ matrix.

The part of the complete data log-likelihood that involves matrix $\boldsymbol{H}$ is

$$-\frac{1}{2}\sum_{t=1}^{T}\left\{\log(|\boldsymbol{H}|) + (\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^\top \boldsymbol{H}^{-1}(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})\right\}. \tag{C.11}$$

From equation (C.11), using property (C.5) and (C.9), it is straightforward to see that

$$\frac{\partial \mathcal{L}_c(\Psi)}{\partial \boldsymbol{H}} = -\frac{1}{2}\sum_{t=1}^{T}\left\{\mathbf{vec}(\boldsymbol{H}^{-\top})^\top + \left[(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^\top \otimes (\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^\top\right]\left(-\boldsymbol{H}^{-\top} \otimes \boldsymbol{H}^{-1}\right)\right\}. \tag{C.12}$$

The derivation of the second derivative begins with equation (C.12). The derivative of $\mathbf{vec}(\boldsymbol{H}^{-\top})^\top$ w.r.t $\boldsymbol{H}$ can be obtained by directly applying property (C.6). The derivative of $-\boldsymbol{H}^{-\top} \otimes \boldsymbol{H}^{-1}$ is obtained by first applying the chain rule to get

$$\frac{\partial\left(-\boldsymbol{H}^{-\top} \otimes \boldsymbol{H}^{-1}\right)}{\partial \boldsymbol{H}} = -\frac{\partial\left(\boldsymbol{H}^{-\top} \otimes \boldsymbol{H}^{-1}\right)}{\partial \boldsymbol{H}^{-1}}\frac{\partial \boldsymbol{H}^{-1}}{\partial \boldsymbol{H}}$$

Then with property (C.10), it can be shown that

$$\frac{\partial\left(\boldsymbol{H}^{-\top} \otimes \boldsymbol{H}^{-1}\right)}{\partial \boldsymbol{H}^{-1}} = (\boldsymbol{I}_K \otimes \boldsymbol{T}_{K,K} \otimes \boldsymbol{I}_K)\left(\mathbf{vec}\left(\boldsymbol{H}^{-\top}\right) \otimes \boldsymbol{I}_{K^2}\right)$$

The above results, together with property (C.6) and (C.8), would give

$$\frac{\partial^2 \mathcal{L}_c(\Psi)}{\partial \boldsymbol{H} \partial \boldsymbol{H}^\top} = \frac{1}{2} \sum_{t=1}^{T} \Big\{ \boldsymbol{I}_{K^2} - \boldsymbol{I}_{K^2} \otimes \left[(\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^\top \otimes (\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t-1})^\top \right] \left[(\boldsymbol{I}_K \otimes \boldsymbol{T}_{K,K} \otimes \boldsymbol{I}_K) \left(\mathbf{vec} \left(\boldsymbol{H}^{-\top} \right) \otimes I_{K^2} \right) \right] \Big\} \left(\boldsymbol{H}^{-\top} \otimes \boldsymbol{H}^{-1} \right) \tag{C.13}$$

Bibliography

Abrahamsen, P. (1997). A review of Gaussian random fields and correlation functions. (Available from http://publications.nr.no/directdownload/publications.nr.no/917_Rapport.pdf)

Alexanderian, A. (2013). A brief note on the Karhunen-Loéve expansion. (Available from http://users.ices.utexas.edu/~alen/articles/KL.pdf)

Allison, P. (2009). Chapter 4: Missing Data. *The SAGE Handbook of Quantitative Methods in Psychology*, SAGE Publications Ltd. 72–89.

Alvera-Azárate, A., Barth, A., Rixen, M. and Beckers, J. M. (2005). Reconstruction of incomplete oceanographic data sets using empirical orthogonal functions: application to the Adriatic Sea surface temperature. *Ocean Modelling* **9**, 325–346.

Ansley, C. F. and Kohn, R. (1982). A geometrical derivation of the fixed interval smoothing algorithm. *Biometrika* **69**, 2, 486–487.

Berliner, L. M., Wikle, C. K. and Cressie, N. (2000). Lond-lead prediction of Pacific SSTs via Bayesian Dynamic Modelling. *Journal of Climate* **13**, 3953–3968.

Brokeman Consult GmbH (2015). DIVERSITY II products user handbook - inland waters, issue 2.1. (Available from http://www.diversity2.info/products/documents/DEL8/DIV2_Products_User_Handbook_Inlandwaters_v2.1.pdf)

Cardot, H. (2000). Nonparametric estimation of smoothed principal component analysis of sampled noisy functions. *Journal of Nonparametric Statistics* **12**, 503–538.

Cressie, N. (1993) *Statistics for Spatial Data*. John Wiley & Sons, New York.

Cressie, N. and Huang, H. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association* **94**, 448, 1330–1340

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data. *Journal of Computational and Graphical Statistics* **19**, 3, 724–745

Cressie, N., Shi, T. and Kang, E. L. (2010). Fixed rank filtering for spatio-temporal data. *Journal of Computational and Graphical Statistics* **19**, 3, 724–745

Cressie, N. and Wikle, C. (2011) *Statistics for Spatio-temporal Data*. John Wiley & Sons, New Jersey

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1, 1–38.

Di, C., Crainiceanu, C. M., Caffo, B. S. and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *The Annals of Applied Statistics* **3**, 458–488.

Di, C., Crainiceanu, C. M. and Jank, W. S. (2014). Multilevel sparse functional principal component analysis. *Stats* **00**, 1–13.

Di Salvo, F., Ruggieri, M. and Plaia, A. (2015). Functional principal component analysis for multivariate multidimensional environmental data. *Environmental and Ecological Statistics* **22**, 4, 739–757.

Doerffer, R. and Schiller, H. (2008). MERIS regional coastal and lake case 2 water: atmospheric correction algorithm theoretical basis document. (Available from http://www.brockmann-consult.de/beam-wiki/download/attachments/1900548/meris_c2r_atbd_atmo_20080609_2.pdf)

Durbin, J. and Koopman, S. J. (2001) *Time Series Analysis by State Space Methods*. Oxford University Press, New York.

Ettinger, B., Guillas, S. and Lai, M. (2012). Bivariate splines for ozone concentration forecasting. *Environmentrics* **23**, 317–328.

Ferraty, F. and Romain, Y. (2011). *The Oxford Handbook of Functional Data Analysis*. Oxford University Press, New York.

Fessler, J. A. and Hero, A. O. (1994). Space-alternating generalized expectation-maximization algorithm. *IEEE Transactions on Signal Processing* **42**, 10, 2664–2677.

Gervini, D. (2009). Detecting and handling outlying trajectories in irregularly sampled functional datasets. *The Annals of Applied Statistics* **3**, 1758–1775.

Giraldo, R., Delicado, P. and Mateu, J. (2009). Continuous time-varying Kriging for spatial prediction of functional data: an environmental application. *Journal of Agricultural, Biological and Environmental Statistics* **15**, 1, 66-82.

Gneiting, T. (2006). Chapter 4: Geostatistical space-time models, stationarity, separability and full symmetry. *Statistical Methods for Spatio-Temporal Systems*. Chapman and Hall/CRC, Boca Raton, FL.

Goldsmith, J., Greven, S. and Crainiceanu, C. (2013). Corrected confidence bands for functional data using principal components. *Biometrics* **69**, 41–51.

Greven, S. and Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika* **97**, 4, 1–17.

Guillas, S. and Lai, M. (2010). Bivariate splines for spatial functional regression models. *Journal of Nonparametric Statistics* **22**, 4, 477-497.

Haggarty, R. A., Miller, C. A. and Scott, E. M. (2013). Spatially weighted functional clustering of river network data. *Journal of the Royal Statistical Society, Series C* **64**, 491–506.

Haslett, J. (1997). On the sample variogram and the sample autocovariance for non-stationary time series. *Journal of the Royal Statistical Society, Series D* **46**, 475–485.

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *The Annals of Statistics* **19**, 4, 2244–2253.

Ho, J. H. and Lin, T. I. (2010). Robust linear mixed model using the skew t distribution with application to schizophrenia data. *Biometrical Journal* **52**, 4, 449–469.

Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician* **64**, 4, 325–334.

Hout, J. P., Tait, H., Rast, M., Delwart, S., Bèzy, J. and Lervini, G. (2001). The optical imaging instruments and their applications: AATSR and MERIS. *EESA Bulletin* **106**, 56–66.

Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society, Series B* **75**, 1, 139–159.

Hurvich, C. M., Simonoff, J. S. and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion *Journal of the Royal Statistical Society, Series B* **60**, 2, 271–293.

Ivanescu, A. E. (2013). A note on bivariate smoothing for two-dimensional functional data. *International Journal of Statistics and Probability* **2**, 102–111.

James, G. M. (2011). Sparseness and functional data analysis. In *The oxford handbook of functional data analysis*. Oxford University Press, New York, 298–323.

James, G. M., Hastie, T. J. and Sugar, C. A. (2000). Principal component models for sparse functional data. *Biometrika*. **87**, 587–602.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering* **82**, 35–45.

Kammann, E. E. and Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society, Series C* **52**, 1–18.

Kang, E. L. and Cressie, N. (2010). Using temporal variability to improve spatial mapping with application to satellite data. *The Canadian Journal of Statistics* **38**, 2, 271–289.

Katzfuss, M and Cressie, N. (2011). Spatio-temporal smoothing and EM estimation for massive remote-sensing data sets. *Journal of Time Series Analysis* **32**, 430–446.

Katzfuss, M and Cressie, N. (2012). Bayesian hierarchical spatio-temporal smoothing for very large datasets. *Environmentrics* **23**, 94–107.

Kayombo, S. and Jorgensen, S. E. (2006). Lake Victoria experience and lessons learned. (Available from http://www.worldlakes.org/uploads/27_lake_victoria_27february2006.pdf)

Laird, N., Lange, N. and Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association* **82**, 97–105.

Law, M. and Kelton, W. D. (1984). Confidence intervals for steady-state simulations: I. A survey of fixed sample size procedures. *Operations Research* **32**, 1221–1239.

Lee, D., Rushworth, A. and Sahu, S. K. (2014). A Bayesian localized conditional autoregressive model for estimating the health effect of air pollution. *Biometrics* **70**, 419–429.

Lindgren, F., Rue, H. and Lindstrom, J.(2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society, Series B* **73**, 4, 423–498.

Lindgren, F. and Rue, H.(2015) Bayesian spatial modelling with R-INLA. *Journal of Statistical Software* **63**, 19,1–25.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*, John Wiley & Sons, New Jersey.

Liu, C. and Rubin, D. B. (1994) The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika* **81**, 4, 633–648.

Lu, G. and Copas, J. B. (2004) Missing at random, likelihood ignorability and model completeness. *The Annals of Statistics* **12**, 2, 754–765.

MacCallum, S. and Merchant, C. (2012). Surface water temperature observations of large lakes by optimal estimation. *Canadian Journal of Remote Sensing* **38**, 25–45.

MacCallum, S. and Merchant, C. (2013). ATSR reprocessing for climate lake surface water temperature: ARC-Lake: algorithm theoretical basis document. (Available from http://www.geos.ed.ac.uk/arclake/ARC-Lake-ATBD-v1.4.pdf)

Mardia, K., Goodall, C., Edwin, R. and Alonso, F. (1998). The kriged Kalman filter. *Sociedad de Estadística e Investigación Operativa* **7**, 217–285.

McLachlan, G. J. and Krishnan, T. (1997) *The EM Algorithm and Extensions*, John Wiley & Sons, New Jersey.

McLachlan, G. J., Peel, D. and Bean, R. W. (2003) Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis* **41**, 379–388.

McNicholas, P. D. and Murphy, T. B. (2008) Parsimonious Gaussian mixture models. *Statistical Computation* **18**, 285–296.

Meng, X. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *Journal of American Statistical Association* **86**, 416, 899–909.

Meng, X. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80**, 2, 267–278.

Meng, X. and Van Dyk, D. (1997). The EM algorithm - an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B* **59**, 3, 511–567.

Merchant, C. and Le Borgne, P. (2004). Retrieval of sea surface temperature from space, based on modeling of infrared radiative transfer: capabilities and limitations. *Journal of Atmospheric and Oceanic Technology* **21**, 1734–1746.

Nguyen, H., Katzfuss, M., Cressie, N. and Braveman, A. (2014). Spatio-temporal data fusion for very large remote sensing datasets. *Technometrics* **56**, 2, 174–185.

Nychka, D. W. (2000). Spatial process estimates as smoothers. *Smoothing and Regression. Approaches, Computation and Application*. Wiley, New York, 393–424.

Paciorek, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science* **25**, 1, 107–125.

Pannullo, F., Lee, D., Waclawski, E. and Leyland, A. (2016). How robust are the estimated effects of air pollution on health? Accounting for model uncertainty using Bayesian model averaging. *Spatial and Spatio-temporal Epidemiology* **18**, 53–62.

Peng, J. and Paul, D. (2009). A geometric approach to maximum likelihood estimation of functional principal components from sparse longitudinal data. *Journal of Computational and Graphical Statistics* **18**, 995–1015.

Peng, J. and Paul, D. (2013). fpca: Restricted MLE for Functional Principal Components Analysis (R package version 0.2-1). (Available from http://CRAN.R-project.org/package=fpca)

Petris, G., Petrone, S. and Campagnoli, P. (2009) *Dynamic Linear Models with R*. Springer, New York.

Petris, G. (2010) An R Package for Dynamic Linear Models. *Journal of Statistical Software* **36**, 1–16.

Pinheiro, J. C. and Bates, D. M. (2000) *Mixed-Effects Models in S and S-plus*. Springer, New York.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Core Team (2016) nlme: Linear and Nonlinear Mixed Effects Models (R package version 3.1-127). (Available from http://CRAN.R-project.org/package=nlme)

Pigorsch, W. W. and Bailer, A. J. (2005) *Analyzing Environmental Data*. John Wiley & Sons, Chichester, England.

Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*, 1st ed. Springer, New York, USA.

Ramsay, J. O., Wickham, H., Graves, S. and Hooker, G. (2013). fda: Functional Data Analysis (R package version 2.4.0). (Available from http://CRAN.R-project.org/package=fda)

Reich, B. J. and Hodges, J. S. (2008). Identification of the variance components in the general two-variance linear model. *Journal of Statistical Planning and Inference*. **138**, 1592–1604.

Reiss, P. T. and Ogden, R. T. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society, Series B*. **71**, 505–523.

Ribeiro Jr, P. J. and Diggle, P. J. (2016). geoR: Analysis of Geostatistical Data (R package version 1.7-5.2). (Available from https://CRAN.R-project.org/package=geoR)

Rice, J. A. and Wu, C. O. (2001). Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* **57**, 253–259.

Rodgers, C. D. (1990). Characterization and error analysis of profiles retrieved from remote sounding measurements. *Journal of Geophysical Research* **95**, 5587–5595.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

Schlather, M., Malinowski, A., Oesting, M., Boecker, D., Strokorb, K., Engelke, S., Martini, J., Ballani, F., Moreva, O., Menck, P. J., Gross, S., Ober, U., Berreth, C., Burmeister, K., Manitz, J., Morena, O., Ribeiro, P., Singleton, R., Pfaff, B. and R Core Team (2016) RandomFields: Simulation and Analysis of RandomFields (R package version 3.1.12). (Available from http://CRAN.R-project.org/package=RandomFields)

Seaman, S., Galati, J., Jackson, D. and Carlin, J.(2013) What is meant by 'missing at random'? *Statistical Science* **28**, 2, 257–268.

Shumway, R. H. and Stoffer, D. S. (1982) An approach to time series smoothing and forecasting using the EM algorithm. *Journal of Time Series Analysis* **3**, 253–264.

Shumway, R. H. and Stoffer, D. S. (2006) *Time Series Analysis and Its Application with R Examples*. Springer, New York.

Shurbet, G. L., Lewis, T. O. and Boullion, T. L. (1974) Quadratic matrix equations. *The Ohio Journal of Science* **74**, 5, 273–277.

Stroud, J. R., Muller, P. and Sanso, B. (2001) Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society, Series B* **64**, 4, 673–689.

Van Dyk, D. A., Meng, X. and Rubin, D. B. (1995). Maximum likelihood estimation via the ECM algorithm: computing the asymptotic variance. *Statistical Sinica* **5**, 1, 55–75.

Van Dyk, D. A. and Meng, X. (2001). The art of data augmentation. *Journal of Computational and Graphical Statistics* **10**, 1, 1–50.

Van Dyk, D. and Meng, X. (2010). Cross-fertilizing strategies for better EM mountain climbing and DA field exploration: a graphical guide book *Statistical Science* **25**, 4, 429–449.

Wakefield, J. (2007) Disease mapping and spatial regression with count data. *Biostatistics* **8**, 2, 158–183.

Wikle, C. and Cressie, N. (1999) A dimensional-reduced approach to space-time Kalman filtering. *Biometrika* **86**, 815–829.

Wikle, C. K., Milliff, R. F., Nychka, D. and Berliner, L. M.(2001) Spatiotemporal hierarchical Bayesian modelling: tropical ocean surface winds. *Journal of the American Statistical Association* **96** 454, 382–397.

Wikle, C. K. and Berliner, L. M.(2005) Combining information across spatial scales. *Technometrics* **47**, 1, 80–91.

Wikle, C. K. and Hooten. M. B. (2007) A general science-based framework for dynamical spatio-temporal models. *Test* **19**, 417–451.

Wood, S. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society, Series B* **62**, 1025–1036.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R.* Chapman and Hall, London.

Wood, S. N., Bravington, M. V. and Hedley, S. L. (2008). Soap film smoothing. *Journal of the Royal Statistical Society, Series B* **70**, 931–955.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society, Series B* **73**, 3–36.

Wu, C. F. (1983). On the convergence properties of the EM algorithm. *The Annal of Statistics* **11**, 1, 95–103.

Xu, K. and Wikle, C. (2007) Estimation of parameterized spatio-temporal dynamic models. *Journal of Statistical Planning and Inference* **137**, 567–588.

Yao, F., Müller, H. and Wang, J. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association* **100**, 557–590.

Zhang, B., Shen, X. and Mumford, S. (2012). Generalized degrees of freedom and adaptive model selection in linear mixed-effects models. *Computational Statistics and Data Analysis* **56**, 574–586.

Zhou, L. and Pan, H. (2014). Principal component analysis of two-dimensional functional data. *Journal of Computational and Graphical Statistics* **23**, 779–801.

Zipunikov, V., Caffo, B., Yousem, D., Davatzikos, C., Schwartz, B. and Crainiceanu, C. (2011). Multilevel functional principal component analysis for high-dimensional data. *Journal of Computational and Graphical Statistics* **20**, 4, 852–873.