PRINCIPAL COMPONENT ANALYSIS OF TWO-DIMENSIONAL FUNCTIONAL DATA WITH SERIAL CORRELATION

A Dissertation

by

SHIRUN SHEN

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Lan Zhou
Co-Chair of Committee,	Na Zou
Committee Members,	Suhasini Subba Rao
	Xianyang Zhang
Head of Department,	Brani Vidakovic

May 2022

Major Subject: Statistics

ABSTRACT

Functional principal component analysis is a popular technique used for analyzing the intrinsically infinite-dimensional functions. The functional principal components help explore the variation patterns of functions and achieve dimension reduction. Some functional data are sequentially observed on two-dimensional domains. How to analyze the serial correlated two-dimensional functional data is an important issue. This dissertation consists of two projects developing the dynamic two-dimensional functional principal component analysis to analyze the serial correlated functional data with Gaussian distribution or generally, with a distribution from exponential family.

The first project proposes a novel model to analyze serial correlated two-dimensional functional data observed sparsely and irregularly on a domain which may not be a rectangle. The approach employs a mixed effects model that specifies the principal component functions as bivariate splines on triangulations and the component scores as random effects which follow an auto-regressive model. We apply the roughness penalty for regularizing the function estimation and develop an effective EM algorithm along with Kalman filter and smoother for calculating the penalized likelihood estimates of the parameters. This approach was applied on simulated datasets and on Texas monthly average temperature data of 49 weather stations from January year 1915 to December year 2014.

The second project proposes the approach to analyze data which follow a distribution from exponential family and are observed over time on two-dimensional domain. Assuming that the natural parameter is a dynamic smooth function of the two-dimensional location, we propose a functional principal component model which models the natural parameter through the combination of smooth principal component functions on two-dimensional domain and principal component scores modeled by autoregressive processes. To address the problem of scalability of large data which is often seen in practice, a variational EM algorithm is proposed for fitting the model. Numerical results on simulated data and the motivating Arctic sea-ice-extent data demonstrate the good performance of the proposed approach.

DEDICATION

This Dissertation is dedicated to my parents

Songduan Shen

and

Xiu'e Xu

who have given me invaluable love and support

ACKNOWLEDGMENTS

First of all, I would like to thank my co-advisors Drs. Lan Zhou and Na Zou at Texas A&M University. Dr. Lan Zhou introduces me into the research topics of Functional Data Analysis. Her insightful guidance and patience in statistics are extremely helpful when I prepare my papers and make this dissertation possible. Dr. Na Zou leads me into another fantastic research topic: Network Analysis. Her enthusiasm in the network research encourages me to explore further interesting questions of social networks.

I also would like to thank Drs. Suhasini Subba Rao and Xianyang Zhang for their selfless service as my dissertation committee members. I appreciate their valuable comments and suggestions in my researches.

I am thankful to Dr. Kejun He, who is my advisor at Renmin University of China, and also the co-author of my papers. He always provides useful thoughts on my research projects and his profound knowledge and tireless hardworking have a long-standing impact on me. I thank Dr. Bohai Zhang of Nankai University, who is the collaborator of one of my papers, for his patience in working with me. I also owe my gratitude to Dr. Jianhua Huang for his encouragement when I was a junior graduate student. His ambition in statistical education and statistics researches always inspire me to keep moving forward.

I really appreciate Department of Statistics in Texas A&M University. The comfortable environment and financial support help me focus on my research without worrying much about the life. I am very grateful to the daily life with my professors and colleagues.

Finally, I would like to thank my families for their continuous support. I am especially thankful to Huiya Zhou. I could not make my path to achieve my degree and get through the hard time without her company and encouragement.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Dr. Lan Zhou (advisor) of Department of Statistics, Dr. Na Zou (co-advisor) of Department of Engineering Technology and Industrial Distribution, and Drs. Suhasini Subba Rao, Xianyang Zhang of Department of Statistics.

The work of Chapter 2 was conducted with Drs. Lan Zhou, and Kejun He of Institute of Statistics and Big Data at Renmin University of China. The work of Chapter 3 was conducted with Drs. Lan Zhou, Kejun He, as well as Bohai Zhang of School of Statistics and Data Science at Nankai University.

All the research work of the dissertation was completed by the student as the major contributor.

Funding Sources

In this dissertation, the research work of the student is funded by the Graduate Assistantship of Texas A&M University.

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iii

ACKNOWLEDGMENTS iv

CONTRIBUTORS AND FUNDING SOURCES v

TABLE OF CONTENTS vi

LIST OF FIGURES viii

LIST OF TABLES x

1. INTRODUCTION AND LITERATURE REVIEWS 1

    1.1 Principal Component Analysis of Serial Correlated 2-d Functional Data with Gaussian Distribution 1
    1.2 Principal Component Analysis of Serial Correlated 2-d Functional Data with A Distribution from Exponential Family 4
    1.3 Overall Structure 8

2. PRINCIPAL COMPONENT ANALYSIS OF SERIAL CORRELATED TWO DIMENSIONAL FUNCTIONAL DATA WITH GAUSSIAN DISTRIBUTION 9

    2.1 Mixed-effects Model for Serial Correlated 2-d Functional Data 9
        2.1.1 Bivariate Spline Basis Functions on a Triangulation 12
    2.2 Model Fitting 14
        2.2.1 Penalized Complete Data Log Likelihood 14
        2.2.2 EM Algorithm 18
        2.2.3 Model Selection 23
    2.3 Simulation Studies 25
    2.4 Texas Temperature Data Analysis 30

3. PRINCIPAL COMPONENT ANALYSIS OF SERIAL CORRELATED TWO DIMENSIONAL FUNCTIONAL DATA WITH A DISTRIBUTION FROM EXPONENTIAL FAMILY 37

    3.1 Mixed Effects Model for Serial Correlated 2-d Functional Data with A Distribution from Exponential Family 37

3.2 Penalized Complete Data Likelihood 41
3.3 The EM algorithm 44
3.3.1 The E-Step 44
3.3.1.1 Laplace approximation - Kalman filter and smoother approach 45
3.3.1.2 Variational inference approach 46
3.3.2 The M-Step 53
3.4 Model Selection 56
3.5 Simulation Studies 57
3.6 Arctic Sea-ice-extent Data Analysis 62

4. SUMMARY AND DISCUSSIONS 67

REFERENCES 68

APPENDIX A. APPENDIX OF CHAPTER 3 76

A.1 The Details of LapKFS Approach in the E Step 76

LIST OF FIGURES

FIGURE Page

1.1 The left panel is the distribution of locations of 49 weather stations in Texas. In the right panel are the connected curves of 3 stations in 2 years (2013–2014), colored corresponding to the points in the left panel. 1

1.2 The observations of the sea-ice-extent data on March, June, September, December in 2010, 2015, 2020, respectively. 6

2.1 An example of triangularization: a square with a hole in the middle used in the simulation study. 26

2.2 The temporal principal component functions in the first setting of simulation study. The first and second principal functions are depicted in the first and second rows, respectively. From left to right are the true PC functions, the estimation of the proposed tFPC, and of the alternative mFPC. 28

2.3 The individual functions in first setting of simulation study. From left to right are the contours at time points $t = 100$, 200, 300, and 400. The true function, the estimation of the proposed tFPC, and of the alternative mFPC are depicted in the first, second, and third rows, respectively. 29

2.4 Triangulation used in the application to the Texas temperature data analysis. 31

2.5 The estimated means of September, December in 1930, 1970, 2014, respectively. ... 32

2.6 Left: the estimated trends of different cities. Right: the estimated one period of different cities. 33

2.7 The first three principal component functions for the real data analysis. 34

2.8 Monthly prediction error (PE) at New Brunfels for two methods. 35

2.9 The prediction error of two methods for the monthly temperature in January–March, 2014. 36

3.1 The CPU time comparison between VEM and LapEM in both binary and Poisson distributions. The CPU time is counted in seconds. 59

3.2 The mean functions in the binary case of simulation study. From left to right are respectively the functions at time points $t = 50, 100, 150,$ and 200. The first row represents the true mean functions while the second row represents the estimated mean functions by VEM approach. 60

3.3 The principal component functions of binary case in simulation study. The first and second row are the first and second principal component functions respectively. The left column represents the true functions, while the right column represents the estimated functions. 60

3.4 The probability surfaces of binary case in simulation study. From left to right are respectively the probability functions at time points $t = 50, 100, 150,$ and 200. The first row represents the true probability functions while the second row represents the estimated functions by VEM approach. 61

3.5 The mean functions in the Poisson case of simulation study. From left to right are respectively the functions at time points $t = 50, 100, 150,$ and 200. The first row represents the true mean functions while the second row represents the estimated mean functions by VEM approach. 61

3.6 The principal component functions of Poisson case in simulation study. The first and second row are the first and second principal component functions respectively. The left column represents the true functions, while the right column represents the estimated functions. 62

3.7 The natural-parameter functions of Poisson case in simulation study. From left to right are respectively the natural-parameter functions at time points $t = 50, 100, 150,$ and 200. The first row represents the true natural-parameter functions while the second row represents the estimated functions by VEM approach. 63

3.8 The location visualization of the sea-ice-extent monthly data, where the red irregular region represents the data domain in Arctic Circle. The blue triangles are constructed for bivariate basis functions. 64

3.9 Top: observations of the sea-ice extent data on March, September in 2010, 2020; Bottom: the corresponding probability surface estimated by the proposed model. 65

3.10 The estimated univariate function $\widehat{\mu}_2(t)$ and bivariate function $\widehat{\mu}_1(x, y)$. 65

3.11 The estimated surface principal components. 66

3.12 The observations and forecasted probability surface on December, 2020. 66

LIST OF TABLES

TABLE Page

2.1 The means and standard errors of PAs and MISEs for the mean function $\mu_t(x, y)$ and the stochastic surface $Z_t(x, y)$. The results are based on 100 simulation runs. ... 27

2.2 The performance of parameters estimating in the simulation study for the noise level $\sigma^2 = 1$ and $(\sigma_1^2, \sigma_2^2) = (1, 0.1)$. Reported are the means of estimations and the standard errors (in parenthesis) based on 100 data replications. 30

2.3 The performance of parameters estimating in the simulation study for the noise level $\sigma^2 = 0.1$ and $(\sigma_1^2, \sigma_2^2) = (0.1, 0.01)$. Reported are the means of estimations and the standard errors (in parenthesis) based on 100 data replications. 30

3.1 The averages with standard deviations (in parenthesis) of different criteria over 100 repeated simulations with different sample size in the case of binary distribution. 58

3.2 The averages with standard deviations (in parenthesis) of different criteria over 100 repeated simulations with different sample size in the case of Poisson distribution. .. 58

1. INTRODUCTION AND LITERATURE REVIEWS

1.1 Principal Component Analysis of Serial Correlated 2-d Functional Data with Gaussian Distribution

Understanding the variation of weather patterns over time and geological locations is important in studying the climate change. To investigate the temperature change in Texas, the United States, we worked on one dataset from the U.S. Historical Climatology Network (Menne et al., 2009, USHCN), collected by National Oceanic Atmospheric Administration (NOAA). This dataset consists of monthly-average temperatures from year 1915 to year 2014, observed at 49 weather stations in Texas. Locations of the weather stations are shown in the left panel of Figure 1.1.

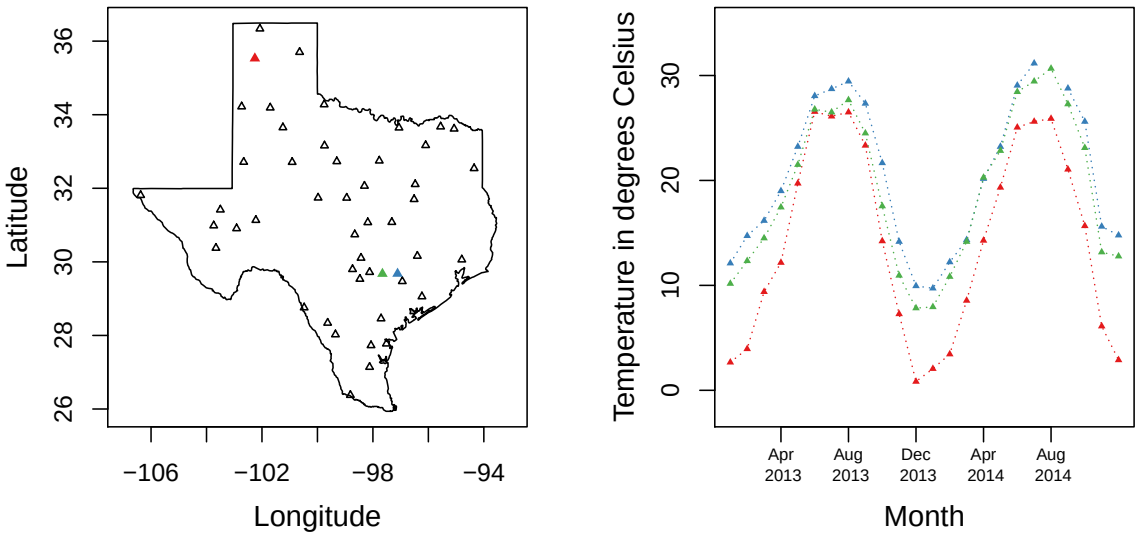

Figure 1.1: The left panel is the distribution of locations of 49 weather stations in Texas. In the right panel are the connected curves of 3 stations in 2 years (2013–2014), colored corresponding to the points in the left panel.

This dataset provides great challenges to data analysis. First, the temperature was affected by both the location of the weather station and the month of measurement. The monthly average

temperatures from 3 weather stations observed in year 2013 and 2014 are given in the right panel of Figure 1.1, where one weather station is in northern Texas, the other two are in the south and are close to each other. It can be seen that even though all temperatures from the three weather stations follow the same seasons, the two southern weather stations have similar temperatures while the northern weather station has lower temperatures in general and larger differences between winters and summers. On the other hand, there usually are serial correlation between temperatures observed over time, see for example, Jones et al. (1986) and Hansen et al. (2006), even when the temperatures were from the same weather station. In addition, the weather stations are sparsely and irregularly located in Texas, and there are missing observations in various months from different weather stations.

Previous studies include the topics of functional principal component analysis (FPCA), functional time series, and spatial functional data. There are different methods for the estimation of FPCA, such as local polynomials (Staniswalis and Lee, 1998; Yao et al., 2005a; Hall et al., 2006), and splines (Rice and Wu, 2001; James et al., 2000; Zhou et al., 2008). FPCA has shown its usefulness in many applications (James and Sugar, 2003; Yao et al., 2005b). The aforementioned papers focused on one-dimensional FPCA, which could not directly be applied to analyze higher-dimensional functional data. Zhou and Pan (2014a) proposed a mixed effects model-based FPCA for spatial (2-dimensional) data. Shi et al. (2022) also proposed a two-dimensional FPCA for feature extraction of image data without any smoothness techniques as the image data was densely samples in the domain. Chen and Jiang (2017) extended to a multi-dimensional FPCA that could be used for higher-dimensional functional data.

Recently, there are many extensions of FPCA to deal with functional data with different complex data structure in real applications. Di et al. (2009) proposed a multilevel (hierarchical) FPCA approach by incorporating the FPCA decomposition into the unknown functions of different levels in functional ANOVA model. Chiou et al. (2014) studied the multivariate FPCA, where the objects were a vector of random functions with the same support. Lin et al. (2016) proposed a interpretable FPCA method through penalizing the support of FPCs to derive the smooth FPCs that were non-

zero in the intervals where curves had major variations, while were strictly zero in others. Li et al. (2016) proposed a supervised FPCA approach borrowing the supervision information from the other datasets by employed the auxiliary variables into the FPC scores with a multivariate linear model. While Ding et al. (2022) also considered the supervised FPCA but incorporated the covariate variables into the mean and FPC functions instead of FPC scores. Lila et al. (2016) focused the smooth principal component analysis on functions whose domain is a two-dimensional manifold. Furthermore, Dai and Müller (2018) extended the FPCA approach to analyze the functional data on Riemannian manifolds. They achieved the Riemannian FPCA by mapping the manifold space onto the L_2 tangent space. Lin and Zhu (2019) proposed a multiscale functional PCA that could deal with heteroscedastic functional data and accurately estimate the high-order PCs. Shang (2014) provided a systematic review of FPCA in exploratory analysis, modeling and forecasting functional data,classification of functional data. Some advanced topics of FPCA can be found in the recent review papers Wang et al. (2016), Li et al. (2022). However, the above literature did not consider the serial correlations among functional objects.

Spatial functional data, which assumes the univariate functional data is spatially-correlated, and utilizes the methods of spatial statistics (e.g., Kriging) to fit the model, is another related topic. Previous works, for example, Zhou et al. (2010) extended the reduced-rank models to the spatially-correlated data, and incorporated the Matérn family to model the spatial correlation. Giraldo et al. (2012) considered the hierarchical clustering of functional data when they are spatially correlated. Li and Guan (2014) proposed the FPCA approach onto the spatiotemporal point processes. Some other additional studies can be found in the papers (Zhang et al., 2016; Zhang and Li, 2021; Kuenzer et al., 2021). A comprehensive survey of spatial functional data can be found in Delicado et al. (2010) and Ruiz-Medina (2012). Although spatial functional data borrowed the location information, they could not be applied for forecasting since they treated the time points as the realization of curves of time.

Functional time series is also a related topic that can shred some light on the challenges from the Texas temperature data. There were some approaches for functional time series, for example,

functional autoregressive (FAR) models (Bosq, 2000; Kokoszka and Reimherr, 2013), two-step methods through incorporating the time series onto FPC scores after employing a FPCA procedure (Shen and Huang, 2008; Shen, 2009; Hyndman and Ullah, 2007; Hyndman and Shang, 2009; Aue et al., 2015; Shang and Hyndman, 2017; Gao et al., 2019). Hörmann and Kokoszka (2012) provided a comprehensive and theoretical introduction of functional time series. However, the aforementioned works focused on the univariate functional time series. Even though these works took the serial correlation of the data into consideration, they did not consider the location effects of the data. Surface time series can be treated as the extension of one-dimensional functional time series. Spatiotemporal statistics (Cressie and Wikle, 2015) could be thought as surface data observed over time. They focused on the Gaussian processes, and achieved the good properties such as stationarity via the well-developed covariance structure. Martínez-Hernández and Genton (2020) provided a review to compare the topics of spatial functional data and surface time series from the perspective of spatiotemporal statistics. However, to the best of our knowledge, the methods from the perspective of functional data for surface time series are still desirable. Our first project contributes to the topic of surface time series from the viewpoint of functional data.

In Chapter 2, we propose a model for principal component analysis of serial correlated 2-dimensional functional data. More precisely, a unified model is used to characterize the serial correlation on the unobserved FPC scores of two-dimensional functional data to tame the curse of dimensionality. The latent FPC scores are modeled by multivariate autoregression. Triangularized bivariate splines are implemented to tackle the irregular shape of domain. All these ideas are integrated into a hidden Markov model (HMM) and an EM algorithm incorporated with Kalman filter and smoother is facilitated to estimate the unknown parameters in a single stage.

1.2 Principal Component Analysis of Serial Correlated 2-d Functional Data with A Distribution from Exponential Family

Arctic is an important component of the global climate system. The sea ice of Arctic Circle has drawn substantial attention in recent years (Vavrus and Harrison, 2003; Meier et al., 2007; Parkinson, 2014). The decreasing trend of ice cover in Arctic Circle has great impacts on the

ecosystem of Arctic and the global climate change. For instance, it changes the behaviors of species that use the ice for their breeding grounds or depend on its presence during their life circle (Stroeve et al., 2008; Meier et al., 2014a). It also changes the Arctic sea surface temperatures (Screen et al., 2013) and results in more extreme weather in mid-latitude regions (Sewall and Sloan, 2004; Cohen et al., 2014). Understanding both spatial and temporal variations of Arctic sea ice is an important issue (Peng and Meier, 2018).

National Oceanic and Atmospheric Administration (NOAA) and National Snow and Ice Data Center (NSIDC) collect a Climate Data Record (CDR) of sea ice concentration from passive microwave data (Meier et al., 2021). The dataset consists of the monthly sea-ice concentrations which are proportional values between 0 and 1 from January 2001 to December 2020. In each month, there are 20043 observations in Arctic Circle, resulting in total 4.8×10^6 samples within 20 years. The data densely cover Arctic Circle where each sample point represents a $25\text{km} \times 25\text{km}$ square. With the commonly-used 15% cut-off criterion (Peng et al., 2013; Zhang and Cressie, 2019), the sea-ice concentrations are transformed to the ice-water binary observations that declare whether a grid cell is covered by sea ice. To be specific, the sample point whose concentration is greater than 15% will be assigned as ice, otherwise it is water. The cumulative area of all grid cells having sea-ice concentrations beyond the cut-off is defined as sea-ice-extent (Parkinson et al., 1999). As an example, Figure 1.2 shows the sea ice cover on March, June, September, December of 2010, 2015, 2020, where the red, blue, and white represent the ice, water, and land regions, respectively. It indicates that there is a one-year periodicity and declining trend of sea ice extent.

Previous studies of Arctic sea ice data from geographers focused on the spatial effects or temporal effects separately. For example, Peng et al. (2013) mainly discussed the temporal variations of the data by visualizing the decreasing trend of time series of sea ice extent. Cavalieri and Parkinson (2012) and Peng and Meier (2018) researched on the regional effects of Arctic sea ice by dividing Arctic region into different regions and exploring their trends individually. Even though the aforementioned studies have provided useful information of Arctic sea ice cover, there are no uncertainty measures from their descriptive statistics. Recently, Zhang and Cressie (2019) employed

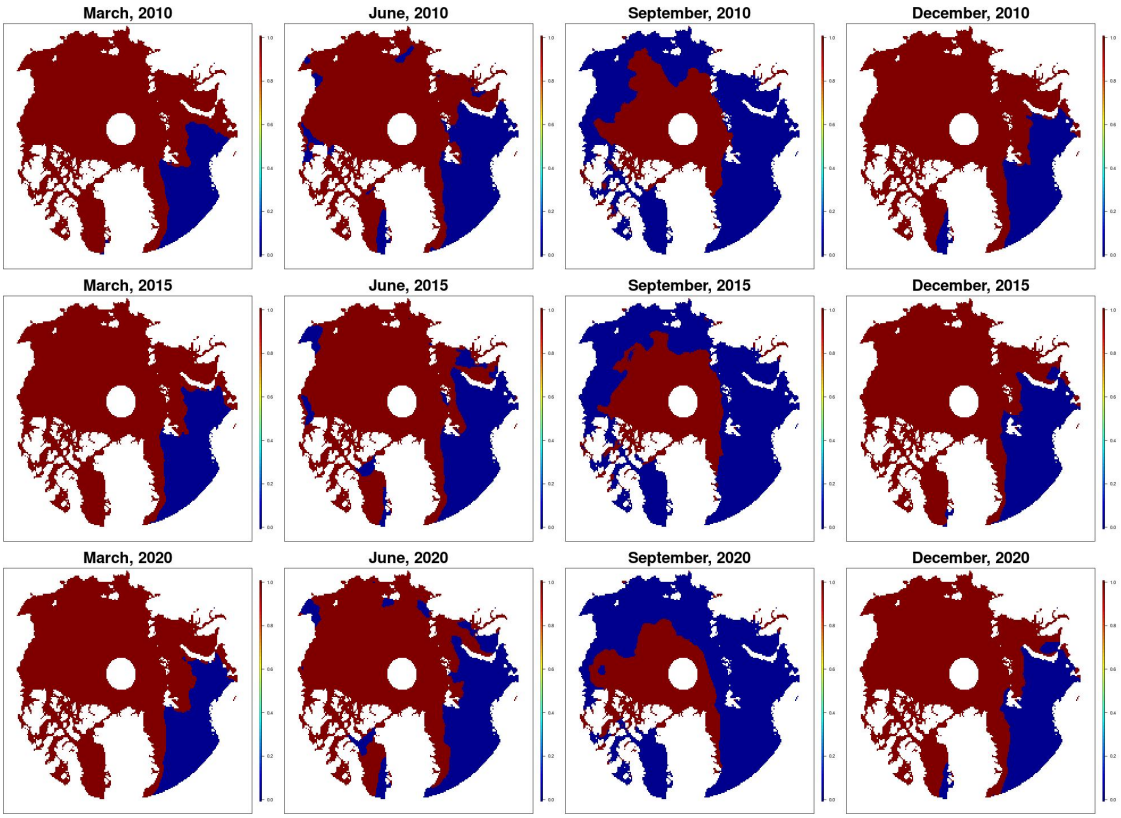

Figure 1.2: The observations of the sea-ice-extent data on March, June, September, December in 2010, 2015, 2020, respectively.

the dynamic spatial generalized linear mixed model to analyze the September sea-ice data over 20 years, while Zhang and Cressie (2020) extended the model to its Bayesian version and used the Markov chain Monte Carlo (MCMC) algorithms for inference. However, these two methods only tackled the single-month data in a short term (e.g., 5 time points) due to the high computational costs of their algorithms. The statistical models for understanding the spatiotemporal patterns of the Arctic sea-ice-extent data with binary outcomes and valid computational methods to tackle the computational challenges are still desired.

We aim to develop a statistical model for analyzing the Arctic sea-ice-extent dataset over 20 years from the perspective of functional data analysis. Some models have been proposed to analyze the binary functional data. For instance, Hall et al. (2008) proposed a latent Gaussian process approach by assuming that the generalized observations were inherited from an underlying unob-

served Gaussian process via the link function (logit link for binary and log-link for count data). While the Gaussian process could be decomposed via functional principal component analysis (FPCA). The model fitting was achieved by a local linear smoother. Serban et al. (2013) extended this approach to model the rare event data (small number of 1's in binary distribution). Goldsmith et al. (2015) considered the generalized function-on-scalar regression model where the response was curve, and the link function could be decomposed as the combinations of covariates with varying coefficient functions and the subject-specific functions and the random deviation functions, where the subject-specific and random deviation functions can be approximated via FPCA. Huang et al. (2014) proposed a method for joint modeling and clustering of non-Gaussian functional data such as binary and count data. Specifically, the authors used a generalized linear mixed model to link the binary observations to a latent longitudinal process that admitted an FPCA decomposition. They utilized a multinomial distribution for the clustering problem. They modeled the paired data similar to Zhou et al. (2008) in the latent process. Li et al. (2018) proposed an Exponential-family Functional PCA model with low-rank structure to model the one/two-way bivariate functional data. They utilized the singular value decomposition techniques and introduced the regularization for the singular vectors. They implemented the iteratively reweighted least squares (IRLS) method to update the parameters. The aforementioned papers, however, did not consider both the spatial effects and serial correlation of the data. They cannot directly be applied to analyze the sea-ice-extent data.

Motivated by analyzing the spatiotemporal-dependent and large-scale Arctic sea-ice-extent data, we propose a unified model in Chapter 3 to study the dynamic two-dimensional functional data with a distribution from exponential family. Borrowing the idea of functional principal component analysis (Zhou and Pan, 2014a), we develop a model for principal component analysis of serial correlated two-dimensional functional data. Incorporating with the exponential family of distributions, the proposed model can be applied to analyze the binary/count data observed over a continuous domain. Based on the bivariate functional PCA on the smooth natural-parameter function of the exponential-family distributions, the model integrates the vector autoregression

onto the functional principal component scores, to analyze both the spatial and temporal variabilities. Furthermore, to deal with the scalability issue of the Arctic sea-ice-extent data, a variational expectation-maximization algorithm has been implemented into the parameters estimation. Numerical examples including simulation studies and real data analysis show the good performance of our approach. Finally, the model can be applied to forecast the future observations, which provides useful information in the Arctic sea-ice-extent data analysis.

1.3 Overall Structure

The rest parts of this dissertation are discussed as below. Chapter 2 introduced the dynamic Gaussian principal component model for two dimensional functional data with serial correlation. Incorporating the autoregression into the functional principal component scores, the model can help analyze the serial correlations among the observations. The EM algorithm along with Kalman filter and smoother is proposed for model fitting. The numerical examples including a simulation study and Texas temperature data analysis indicate the good performance of the proposed model compared with the benchmark method.

Chapter 3 discussed the dynamic exponential-family functional principal component model for two-dimensional binary/count data. The binary/count functional data can be modeled via a distribution of exponential family. We assume the natural parameters of exponential-family distributions can be decomposed as the structure of functional principal components analysis discussed in Chapter 2. A variational EM algorithm is developed for model fitting to reduce the computational costs.

Chapter 4 provides the summary and discussions of the future extensions of this dissertation.

2. PRINCIPAL COMPONENT ANALYSIS OF SERIAL CORRELATED TWO DIMENSIONAL FUNCTIONAL DATA WITH GAUSSIAN DISTRIBUTION

This chapter is organized as follows. In Section 2.1, we propose the model to analyze the temporal-dependent two-dimensional functional data on an irregular domain. Section 2.2 investigates an EM algorithm to estimate the parameters where the E step is calculated through Kalman filter and smoother procedures. The empirical performance of the proposed method is illustrated via a simulation study and Texas temperature data analysis in Sections 2.3 and 2.4, respectively.

2.1 Mixed-effects Model for Serial Correlated 2-d Functional Data

Let Ω be a compact subset of $\mathbb{R}^2$, and (x, y) be the 2-dimensional index variable on Ω. Suppose $Z(x, y)$ is a stochastic process on Ω with finite second moment, $\int_\Omega \mathbb{E}\{Z^2(x, y)\}\mathrm{d}x\mathrm{d}y < \infty$. Denote the mean function of $Z(x, y)$ as $\mu(x, y) = \mathbb{E}\{Z(x, y)\}$ and the covariance function of $Z(x, y)$ as

$$\mathcal{K}(x_1, y_1; x_2, y_2) = \mathbb{E}\big[\{Z(x_1, y_1) - \mu(x_1, y_1)\}\{Z(x_2, y_2) - \mu(x_2, y_2)\}\big].$$

Under mild conditions, Mercer's lemma (Mercer, 1909) shows that there exists an orthonormal sequence $\{\phi_j\}_j$ in $L_2(\Omega)$ as eigenfunctions, and a decreasing non-negative sequence $\{\zeta_j\}_j$ as eigenvalues, such that the covariance function can be expanded as

$$\mathcal{K}(x_1, y_1; x_2, y_2) = \sum_{j=1}^{\infty} \zeta_j \phi_j(x_1, y_1)\phi_j(x_2, y_2).$$

The orthonormality of ϕ_j's means that $\int_\Omega \phi_j \phi_{j'}\,\mathrm{d}x\mathrm{d}y = \delta_{jj'}$, where $\delta_{jj'}$ is the Kronecker delta. Applying Karhunen-Loève theorem (Karhunen, 1946; Loève, 1946), the random surface $Z(x, y)$ admits the following expansion

$$Z(x, y) = \mu(x, y) + \sum_{j=1}^{\infty} \alpha_j \phi_j(x, y), \tag{2.1}$$

where α_j's are uncorrelated random variables with mean zero and variances $\{\zeta_j\}_j$. Following Ramsay and Silverman (2005), the random variable α_j and the eigenfunction $\phi_j(x, y)$ are called the j-th FPC score and principal component function, respectively.

Assume that $Z(x, y)$ can be well approximated by its projection on the space spanned by the first J eigenfunctions and treat the rest of terms as the noise, we arrive at the following model

$$Z(x, y) = \mu(x, y) + \sum_{j=1}^{J} \alpha_j \phi_j(x, y) + \epsilon(x, y), \tag{2.2}$$

where $\epsilon(x, y)$ is a random variable with mean 0 and variance σ^2.

When there are n independent copies of $Z(x, y)$, denoted by $Z_1(x, y)$, ..., $Z_n(x, y)$, Zhou and Pan (2014a) proposed a mixed effects model-based approach and model the PC scores as the random effects. When the random surfaces $Z_t(x, y)$, $t = 1, \ldots, n$, have time and location-dependent mean function $\mu_t(x, y) = \mathbb{E}\{Z_t(x, y)\}$ and $Z_t(x, y) - \mu_t(x, y)$ are serial correlated, we need consider both dependence between the locations and time points.

We first assume that location effect and the time effect are separable such that $\mu_t(x, y) = \mu_1(x, y)\mu_2(t)$. Note that if one multiplies $\mu_1(x, y)$ by a non-zero constant c and divides $\mu_2(t)$ by c, the value of $\mu_t(x, y)$ does not change. For identifiability purpose, we require that the L_2-norm of $\mu_1(x, y)$ be identity, i.e.,

$$\|\mu_1(x, y)\|_2 = 1. \tag{2.3}$$

Next borrowing the idea of FPCA as in (2.2), we propose the model

$$Z_t(x, y) = \mu_1(x, y)\mu_2(t) + \sum_{j=1}^{J} \alpha_{j,t} \phi_j(x, y) + \epsilon_t(x, y), \quad t = 1, \ldots, n, \tag{2.4}$$

where $\alpha_{j,t}$ is the j-th FPC score at time t, $\phi_j(x, y)$ is the j-th PC function which are orthonormal, i.e., $\int_\Omega \phi_j \phi_{j'} = \delta_{j,j'}$, with $\delta_{j,j'}$ being the Kronecker delta, and $\epsilon_t(x, y)$ is a white noise process with mean zero and variance σ^2.

Furthermore, $\boldsymbol{\alpha}_j = \{\alpha_{j,t}\}_{t=1}^n$, $j = 1, \ldots, J$, are independent stationary time series. For each j,

the time series $\{\alpha_{j,t}\}_{t=1}^{n}$, follows the p-th order autoregressive model (AR(p)). To be specific,

$$\alpha_{j,t} = \sum_{i=1}^{p} k_i \alpha_{j,t-i} + \eta_{j,t}, \quad \eta_{j,t} \overset{\text{ind}}{\sim} \mathrm{N}(0, \sigma_j^2), \quad j = 1, \ldots, J, \quad t = 1, \ldots, n, \tag{2.5}$$

where k_i's and $\eta_{j,t}$'s are the autoregressive coefficients and white noises of the AR(p) models, respectively. For the identifiability of the principal components, we assume that $\sigma_1^2 > \cdots > \sigma_J^2$.

Note that when $k_i = 0, i = 1, \ldots, p$, the FPC scores $\alpha_{j,t}$'s are mutually independent and normally distributed. Then the proposed model (2.4) degenerates to the mFPC model in Zhou and Pan (2014a). For notation simplicity consideration, we called the proposed model the temporal-dependent functional PC (tFPC, for short) model.

Assuming that the functions $\mu_1(x, y)$, $\mu_2(t)$ and $\phi_j(x, y)$ are smooth, we can approximate them by basis expansions. For the approximation of bivariate functions $\mu_1(x, y)$ and $\phi_j(x, y)$, we utilize the orthonormal bivariate spline basis functions constructed on triangulations (Lai and Schumaker, 2007) due to its advantage on irregular domains. In the following Section 2.1.1, we introduce the details of bivariate basis on triangulations. As for the basis expansion of univariate function $\mu_2(t)$, we can choose from the commonly used regression spline (de Boor, 1978), Bernstein polynomial (Lai and Schumaker, 2007), and Fourier basis (Ramsay and Silverman, 2005), etc.

To be specific, let $\mathbf{b}(x, y)$ denote n_b-dimensional vectors of orthonormal bivariate basis functions with

$$\int_{\Omega} \mathbf{b}(x, y)\mathbf{b}^{\mathsf{T}}(x, y)\mathrm{d}x\mathrm{d}y = \mathbf{I}_{n_b}, \tag{2.6}$$

and $\mathbf{c}(t)$ denote n_c-dimensional vectors of univariate basis functions. We write the basis expansions of the smooth functions as

$$\mu_1(x, y) = \mathbf{b}(x, y)^{\mathsf{T}}\boldsymbol{\theta}_b, \;\; \mu_2(t) = \mathbf{c}(t)^{\mathsf{T}}\boldsymbol{\theta}_c,$$

and

$$\phi_j(x, y) = \mathbf{b}(x, y)^{\mathsf{T}}\boldsymbol{\theta}_j, \quad j = 1, \ldots, J,$$

where the basis coefficients $\boldsymbol{\theta}_b \in \mathbb{R}^{n_b}$ with $\|\boldsymbol{\theta}_b\| = 1$, $\boldsymbol{\theta_c} \in \mathbb{R}^{n_c}$, and $\boldsymbol{\theta}_j \in \mathbb{R}^{n_b}$, $j = 1, \ldots, J$, are orthonormal. Denote $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J) \in \mathbb{R}^{n_b \times J}$ and $\boldsymbol{\alpha}_t = (\alpha_{1,t}, \ldots, \alpha_{J,t})^\mathsf{T}$, model (2.4) can be rewritten as

$$Z_t(x, y) = \mathbf{b}(x, y)^\mathsf{T} \boldsymbol{\theta}_b \boldsymbol{\theta}_c^\mathsf{T} \mathbf{c}(t) + \mathbf{b}(x, y)^\mathsf{T} \boldsymbol{\Theta} \boldsymbol{\alpha}_t + \epsilon_t(x, y), \tag{2.7}$$

and (2.5) becomes

$$\boldsymbol{\alpha}_t = \sum_{i=1}^{p} k_i \boldsymbol{\alpha}_{t-i} + \boldsymbol{\eta}_t,$$

where $\boldsymbol{\eta}_t \sim \mathrm{N}(\mathbf{0}, \mathbf{H}_J)$ with covariance matrix $\mathbf{H}_J = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_J^2)$.

Suppose the sparsely sampled 2-dimensional surfaces are observed at time $t = 1, \ldots, T$. At time point t, there are n_t randomly sampled points $(x_{t1}, y_{t1}), \ldots, (x_{tn_t}, y_{tn_t})$ on the surface.

Denote $\mathbf{z}_t = (Z_t(x_{t1}, y_{t1}), \ldots, Z_t(x_{tn_t}, y_{tn_t}))^\mathsf{T}$, $\mathbf{B}_t = (\mathbf{b}(x_{t1}, y_{t1}), \ldots, \mathbf{b}(x_{tn_t}, y_{tn_t}))^\mathsf{T}$, $\boldsymbol{\epsilon}_t = (\epsilon_t(x_{t1}, y_{t1}), \ldots, \epsilon_t(x_{tn_t}, y_{tn_t}))^\mathsf{T}$, and $\mathbf{c}_t = \mathbf{c}(t)$ for notational simplicity. Model (2.7) for both observed data and latent variables can then be written as

$$\begin{aligned} \mathbf{z}_t &= \mathbf{B}_t \boldsymbol{\theta}_b \boldsymbol{\theta}_c^\mathsf{T} \mathbf{c}_t + \mathbf{B}_t \boldsymbol{\Theta} \boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{n_t}), \\ \boldsymbol{\alpha}_t &= \sum_{i=1}^{p} k_i \boldsymbol{\alpha}_{t-i} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \mathrm{N}(\mathbf{0}, \mathbf{H}_J), \end{aligned} \tag{2.8}$$

and the identifiability constraints are the same as mentioned above.

Denote $\mathbf{k} = (k_1, \ldots, k_p)^\mathsf{T}$. For model (2.8), the unknown parameters to be estimated are $\Xi = \{\boldsymbol{\theta}_b, \boldsymbol{\theta}_c, \boldsymbol{\Theta}, \mathbf{k}, \sigma^2, \{\sigma_j^2\}_{j=1}^J\}$.

2.1.1 Bivariate Spline Basis Functions on a Triangulation

In this section, we discuss the construction of the 2-dimensional orthonormal basis function $\mathbf{b}(x, y)$. One trivial choice is the tensor-product B spline basis functions, i.e., $\mathbf{b}(x, y) = \mathbf{b}_1(x) \otimes \mathbf{b}_2(y)$. However, the tensor-product B spline basis functions will cause two problems: (1) the computational cost is usually expensive due to a large number of tensor-product basis functions; and (2) this basis can only used in regular regions like rectangle. To overcome these challenges, we alternatively introduce the Bernstein bivariate polynomial splines on triangulations. The book

Lai and Schumaker (2007) presented the mathematical properties of the bivariate spline. Zhou and Pan (2014a) applied such bivariate splines into their mFPC model. Due to the great properties of Bernstein bivariate spline, there emerged many applications in spatial statistical models (Yu et al., 2020; Wang et al., 2020). Figure 2.1 depicts the triangulation example that will be used in our simulation study. As we can see, unlike the commonly used tensor product of univariate basis functions, this bivariate basis can easily handle the irregular shapes on $\mathbb{R}^2$.

Denote δ as a triangle, which has the counter-wise vertices $(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$. Then for any point $\mathbf{v} \in \mathbb{R}^2$, there is a unique representation in the form $\mathbf{v} = b_1\mathbf{v}_1 + b_2\mathbf{v}_2 + b_3\mathbf{v}_3$. The three coefficients (b_1, b_2, b_3) are called the barycentric coordinates of $\mathbf{v}$ with respect to the triangle δ. Given a non-negative integer d and for any i, j, k such that $i + j + k = d$, the Bernstein polynomials of degree d relative to triangle δ are defined as

$$B^d_{ijk}(\mathbf{v}) = \frac{d!}{i!j!k!} b_1^i b_2^j b_3^k.$$

Let $P_d(\delta)$ be the space of polynomials defined on the triangle δ with degree d. Then the Bernstein polynomials $B^d_{ijk}, i + j + k = d$, form a basis for $P_d(\delta)$. That is, for any function $s \in P_d(\delta)$, we have

$$s(\mathbf{v}) = \sum_{i+j+k=d} \gamma_{ijk} B^d_{ijk}(\mathbf{v}).$$

For an irregular domain, we can construct a triangulation $\Delta = \{\delta_1, \ldots, \delta_M\}$ whose union covers the irregular region Ω (see, for example, Lai and Schumaker, 2007). We construct the Bernstein polynomial basis functions with respect to each δ_i, and the collection of all such polynomials form a basis for $P_d(\Delta)$, the space of continuous piecewise polynomials of degree d on Δ. With additional smoothness conditions that the derivatives up to r degree are continuous, the bivariate basis functions $\mathbf{b}(x, y)$ are constructed. The details of smoothness conditions are referred to Zhou and Pan (2014a).

2.2 Model Fitting

2.2.1 Penalized Complete Data Log Likelihood

Following model (2.8), it is natural to estimate the unknown parameters Ξ by maximizing the log likelihood function with some penalization to regularize the estimates of the smooth functions. However, since the latent FPC scores $\{\boldsymbol{\alpha}_t\}_{t=1}^T$ follow an AR(p) model, it is not feasible to integrate out $\{\boldsymbol{\alpha}_t\}_{t=1}^T$ to get an analytical form of the log likelihood function of Ξ. By treating $\{\boldsymbol{\alpha}_t\}_{t=1}^T$ as missing data, we can get an analytical form of the complete data log likelihood and then apply the EM algorithm (Dempster et al., 1977) for parameter estimation.

The negative twice log likelihood is

$$-2l_c(\Xi; \{\mathbf{z}_t\}_{t=1}^n, \{\boldsymbol{\alpha}_t\}_{t=1}^n) = -2\log p(\mathbf{z}_1, \ldots, \mathbf{z}_n, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n).$$

The joint probability $p(\mathbf{z}_1, \ldots, \mathbf{z}_n, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n)$ can be decomposed into the multiple of the probability density $p(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n)$ and the likelihood of observations given the latent variables, i.e., $p(\mathbf{z}_1, \ldots, \mathbf{z}_n | \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n)$. Thus, the negative twice log likelihood can be written as

$$\begin{aligned} &-2\log p(\mathbf{z}_1, \ldots, \mathbf{z}_n, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n) \\ &\qquad = -2\log p(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n) - 2\log p(\mathbf{z}_1, \ldots, \mathbf{z}_n | \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n), \end{aligned} \tag{2.9}$$

The AR(p) time structure indicates that the first part in (2.9) can be decomposed as the joint density of initial states $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p$ and the later states $\boldsymbol{\alpha}_t | \boldsymbol{\alpha}_{t-1}, \ldots, \boldsymbol{\alpha}_{t-p}$ for $t = p+1, \ldots, n$

$$\begin{aligned} -2\log p(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n) &= -2\log p(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p) - 2\log p(\boldsymbol{\alpha}_{p+1}, \ldots, \boldsymbol{\alpha}_n | \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p) \\ &= -2\log p(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_p) - 2\sum_{t=p+1}^{n} \log p(\boldsymbol{\alpha}_t | \boldsymbol{\alpha}_{t-1}, \ldots, \boldsymbol{\alpha}_{t-p}). \end{aligned} \tag{2.10}$$

The second term of (2.10) has the explicit expression as

$$
\begin{aligned}
&-2\sum_{t=p+1}^{n}\log p(\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1},\ldots,\boldsymbol{\alpha}_{t-p}) \\
&\qquad=\sum_{j=1}^{J}\sum_{t=p+1}^{n}\left\{\log\sigma_j^2+\frac{1}{\sigma_j^2}(\alpha_{j,t}-k_1\alpha_{j,t-1}-\cdots-k_p\alpha_{j,t-p})^2\right\},
\end{aligned}
$$

which comes from the conditional probability $\boldsymbol{\alpha}_t|\boldsymbol{\alpha}_{t-1},\ldots,\boldsymbol{\alpha}_{t-p}\sim \mathrm{N}(\sum_{i=1}^{p}k_i\boldsymbol{\alpha}_{t-i},\mathbf{H}_J)$. As for the initial states, according to the independence of α_j for $j=1,\ldots,J$, and referred to Box et al. (2015), the first term of (2.10) can be derived as

$$
\begin{aligned}
-2\log p(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_p)&=-2\sum_{j=1}^{J}\log p(\alpha_{1j},\ldots,\alpha_{pj}) \\
&=\sum_{j=1}^{J}\left\{p\log\sigma_j^2-\log|\mathbf{M}_j|+\frac{1}{\sigma_j^2}S_{pj}(\mathbf{k})\right\},
\end{aligned}
$$

where $S_{pj}(\mathbf{k})=\widetilde{\boldsymbol{\alpha}}_j^{\mathsf{T}}\mathbf{M}_j\widetilde{\boldsymbol{\alpha}}_j=\sum_{i=1}^{p}\sum_{k=1}^{p}m_{ik}^{(j)}\alpha_{j,i}\alpha_{j,k}$ are the residual sum of squares, with $\widetilde{\boldsymbol{\alpha}}_j=(\alpha_{j,1},\ldots,\alpha_{j,p})^{\mathsf{T}}$ and $\mathbf{k}=(k_1,\ldots,k_p)^{\mathsf{T}}$. Further, the inverse covariance matrices are

$$
\mathbf{M}_j=\begin{pmatrix}\gamma_0 & \gamma_1 & \cdots & \gamma_{p-1}\\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2}\\ \vdots & \vdots & \ddots & \vdots\\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0\end{pmatrix}^{-1}\in\mathbb{R}^{p\times p} \tag{2.11}
$$

where the elements in the matrices $\gamma_i=\mathbb{E}(\alpha_{j,t+i}\alpha_{j,t})/\sigma_j^2=\mathbb{E}(\alpha_{j,t}\alpha_{j,t+i})/\sigma_j^2$, $i=0,\ldots,p-1$, and $j=1,\ldots,J$. Therefore, the first part of (2.9) is

$$
-2\log p(\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_n)=\sum_{j=1}^{J}\left\{n\log\sigma_j^2-\log|\mathbf{M}_j|+\frac{1}{\sigma_j^2}S_j(\mathbf{k})\right\},
$$

where the total residual sum of squares is

$$S_j(\mathbf{k}) = S_{pj}(\mathbf{k}) + \sum_{t=p+1}^{n} (\alpha_{j,t} - k_1\alpha_{j,t-1} - \cdots - k_p\alpha_{j,t-p})^2 = \widetilde{\mathbf{k}}^{\mathsf{T}}\mathbf{D}_j\widetilde{\mathbf{k}},$$

with $\widetilde{\mathbf{k}}^{\mathsf{T}} = (1, \mathbf{k}^{\mathsf{T}}) = (1, k_1, \ldots, k_p)$ and

$$\mathbf{D}_j = \begin{pmatrix} D_{11,j} & -D_{12,j} & -D_{13,j} & \ldots & -D_{1,p+1,j} \\ -D_{12,j} & D_{22,j} & D_{23,j} & \ldots & D_{2,p+1,j} \\ \vdots & \vdots & \vdots & & \vdots \\ -D_{1,p+1,j} & D_{2,p+1,j} & D_{3,p+1,j} & \ldots & D_{p+1,p+1,j} \end{pmatrix}, \tag{2.12}$$

and the elements $D_{ik,j} = D_{ki,j} = \alpha_{j,i}\alpha_{j,k} + \alpha_{j,i+1}\alpha_{j,k+1} + \cdots + \alpha_{j,n+1-k}\alpha_{j,n+1-i}$. Further details can be referred to Box et al. (2015, Appendix A7.4). Following the relationship between observed variables $\mathbf{z}_t$ and latent variables $\boldsymbol{\alpha}_t$, the second part of the complete data log likelihood (2.9), i.e., the likelihood of the observed data given latent variables can be written as

$$\begin{aligned} &-2\log\{p(\mathbf{z}_1, \ldots, \mathbf{z}_n|\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n)\} \\ &\quad = -2\sum_{t=1}^{n} \log p(\mathbf{z}_t|\boldsymbol{\alpha}_t) \\ &\quad = \sum_{t=1}^{n} n_t \log\sigma^2 + \frac{1}{\sigma^2}\sum_{t=1}^{n} (\mathbf{z}_t - \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t)^{\mathsf{T}}(\mathbf{z}_t - \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t). \end{aligned}$$

Letting $l_c(\Xi; \{\mathbf{z}_t\}_{t=1}^n, \{\boldsymbol{\alpha}_t\}_{t=1}^n)$ denote the complete data log likelihood, it hence can be written as

$$\begin{aligned} &-2l_c(\Xi; \{\mathbf{z}_t\}_{t=1}^n, \{\boldsymbol{\alpha}_t\}_{t=1}^n) \\ &\quad = \sum_{j=1}^{J} \left\{ n\log\sigma_j^2 - \log|\mathbf{M}_j| + \frac{1}{\sigma_j^2} S_j(\mathbf{k}) \right\} + \sum_{t=1}^{n} n_t \log\sigma^2 \\ &\qquad + \frac{1}{\sigma^2}\sum_{t=1}^{n} (\mathbf{z}_t - \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t)^{\mathsf{T}}(\mathbf{z}_t - \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t), \end{aligned} \tag{2.13}$$

We next introduce the roughness penalty for the regularization of the estimates of the smooth

functions. For the univariate function $\mu_2(t)$, as in Zhou et al. (2008), we use the integrated squared second derivatives and the roughness penalty takes the form

$$\int_T \left\{ \frac{\partial^2 \mu_2(t)}{\partial t^2} \right\}^2 \mathrm{d}t = \boldsymbol{\theta}_c^\mathsf{T} \left[\int_T \left\{ \frac{\partial^2 \mathbf{c}(t)}{\partial t^2} \frac{\partial^2 \mathbf{c}(t)^\mathsf{T}}{\partial t^2} \right\} \mathrm{d}t \right] \boldsymbol{\theta}_c := \boldsymbol{\theta}_c^\mathsf{T} \mathbf{P} \boldsymbol{\theta}_c.$$

For a generic bivariate function $f(x, y)$, we use the thin plate penalization (Ruppert et al., 2003) which is defined as

$$\int_\Omega \left[\left\{ \frac{\partial^2 f(x,y)}{\partial x^2} \right\}^2 + 2 \left\{ \frac{\partial^2 f(x,y)}{\partial x \partial y} \right\}^2 + \left\{ \frac{\partial^2 f(x,y)}{\partial y^2} \right\}^2 \right] \mathrm{d}x \mathrm{d}y.$$

Denote

$$\boldsymbol{\Gamma} = \int_\Omega \left\{ \frac{\partial^2 \mathbf{b}(x,y)}{\partial x^2} \frac{\partial^2 \mathbf{b}(x,y)^\mathsf{T}}{\partial x^2} + 2 \frac{\partial^2 \mathbf{b}(x,y)}{\partial x \partial y} \frac{\partial^2 \mathbf{b}(x,y)^\mathsf{T}}{\partial x \partial y} + \frac{\partial^2 \mathbf{b}(x,y)}{\partial y^2} \frac{\partial^2 \mathbf{b}(x,y)^\mathsf{T}}{\partial y^2} \right\} \mathrm{d}x \mathrm{d}y.$$

With basis expansions, the thin plate penalty for $\mu_1(x, y)$ and $\phi_j(x, y)$ can be written as, respectively, $\boldsymbol{\theta}_b \boldsymbol{\Gamma} \boldsymbol{\theta}_b$ and $\boldsymbol{\theta}_j \boldsymbol{\Gamma} \boldsymbol{\theta}_j$, $j = 1, \ldots, J$.

Thus, the penalized complete data log likelihood has the expression

$$\begin{aligned} & -2 l_c(\Xi; \{\mathbf{z}_t\}_{t=1}^n, \{\boldsymbol{\alpha}_t\}_{t=1}^n) + \mathrm{Penalty}(\lambda; \Xi) \\ & \quad = \sum_{j=1}^J \left\{ n \log \sigma_j^2 - \log |\mathbf{M}_j| + \frac{1}{\sigma_j^2} S_j(\mathbf{k}) \right\} + \sum_{t=1}^n n_t \log \sigma^2 \\ & \qquad + \frac{1}{\sigma^2} \sum_{t=1}^n (\mathbf{z}_t - \mathbf{B}_t \boldsymbol{\theta}_b \boldsymbol{\theta}_c^\mathsf{T} \mathbf{c}_t - \mathbf{B}_t \boldsymbol{\Theta} \boldsymbol{\alpha}_t)^\mathsf{T} (\mathbf{z}_t - \mathbf{B}_t \boldsymbol{\theta}_b \boldsymbol{\theta}_c^\mathsf{T} \mathbf{c}_t - \mathbf{B}_t \boldsymbol{\Theta} \boldsymbol{\alpha}_t), \\ & \qquad\quad + \lambda_{\mu_s} \boldsymbol{\theta}_b^\mathsf{T} \boldsymbol{\Gamma} \boldsymbol{\theta}_b + \lambda_{\mu_t} \boldsymbol{\theta}_c^\mathsf{T} \mathbf{P} \boldsymbol{\theta}_c + \lambda_{pc} \sum_{j=1}^J \boldsymbol{\theta}_j^\mathsf{T} \boldsymbol{\Gamma} \boldsymbol{\theta}_j, \end{aligned} \tag{2.14}$$

where $\lambda = (\lambda_{\mu_s}, \lambda_{\mu_t}, \lambda_{pc})$ are the regularization parameters.

2.2.2 EM Algorithm

To estimate the parameters, instead of minimizing (3.8), the EM algorithm iteratively minimizes

$$Q(\Xi|\Xi^{(0)}) = \mathbb{E}\big[-2l_c(\Xi; \{\mathbf{z}_t\}_{t=1}^n, \{\boldsymbol{\alpha}_t\}_{t=1}^n)|\{\mathbf{z}_t\}_{t=1}^n, \Xi^{(0)}\big] + \text{Penalty}(\lambda; \Xi),$$

where $\Xi^{(0)}$ are the current guesses of the parameter values (i.e.values of the parameters from the previous iteration).

The E step. In the E step, we calculate $Q(\Xi|\Xi^{(0)})$. Denote $\widehat{\boldsymbol{\alpha}}_t = \mathbb{E}(\boldsymbol{\alpha}_t|\{\mathbf{z}_t\}_{t=1}^n, \Xi^{(0)})$, $\widehat{\boldsymbol{\Sigma}}_t = \text{Var}(\boldsymbol{\alpha}_t|\{\mathbf{z}_t\}_{t=1}^n, \Xi^{(0)})$, and $\widehat{\mathbf{D}}_j = \mathbb{E}(\mathbf{D}_j|\{\mathbf{z}_t\}_{t=1}^n, \Xi^{(0)})$, $j = 1, \ldots, J$. We obtain

$$\widehat{S}_j(\mathbf{k}) := \mathbb{E}[S_j(\mathbf{k})|\{\mathbf{z}_t\}_{t=1}^n, \Xi^{(0)}] = (1, \mathbf{k}^\mathsf{T})\mathbb{E}(\mathbf{D}_j|\{\mathbf{z}_t\}_{t=1}^n, \Xi^{(0)})(1, \mathbf{k}^\mathsf{T})^\mathsf{T} = (1, \mathbf{k}^\mathsf{T})\widehat{\mathbf{D}}_j(1, \mathbf{k}^\mathsf{T})^\mathsf{T}. \tag{2.15}$$

Using (2.13)–(2.15), it shows that

$$\begin{aligned} Q(\Xi|\Xi^{(0)}) = \sum_{j=1}^J \big\{ n \log \sigma_j^2 - \log|\mathbf{M}_j| + \frac{1}{\sigma_j^2}\widehat{S}_j(\mathbf{k})\big\} + \sum_{t=1}^n n_t \log \sigma^2 \\ + \frac{1}{\sigma^2}\sum_{t=1}^n \big\{ (\mathbf{z}_t - \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^\mathsf{T}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}\widehat{\boldsymbol{\alpha}}_t)^\mathsf{T}(\mathbf{z}_t - \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^\mathsf{T}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}\widehat{\boldsymbol{\alpha}}_t) \\ + \text{tr}(\mathbf{B}_t\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_t\boldsymbol{\Theta}^\mathsf{T}\mathbf{B}_t)\big\} \\ + \lambda_{\mu_s}\boldsymbol{\theta}_b^\mathsf{T}\boldsymbol{\Gamma}\boldsymbol{\theta}_b + \lambda_{\mu_t}\boldsymbol{\theta}_c^\mathsf{T}\mathbf{P}\boldsymbol{\theta}_c + \lambda_{pc}\boldsymbol{\theta}_j^\mathsf{T}\boldsymbol{\Gamma}\boldsymbol{\theta}_j, \end{aligned} \tag{2.16}$$

Hence, to calculate the value of (2.16), we only need calculate $\widehat{\boldsymbol{\alpha}}_t$, $\widehat{\boldsymbol{\Sigma}}_t$, and $\widehat{\mathbf{D}}_j$, $j = 1, \ldots, J$.

Note that model (2.8) can be viewed as a state-space model (Durbin and Koopman, 2012). To be specific, denote $\boldsymbol{\beta}_t = (\boldsymbol{\alpha}_{t+p}^\mathsf{T}, \ldots, \boldsymbol{\alpha}_t^\mathsf{T})^\mathsf{T}$, $t = 1, \ldots, n$, where $\boldsymbol{\alpha}_t = 0$ when $t \geq n$. Denote

$\mathbf{S} = \big(\mathbf{0}, \cdots \mathbf{0}, \mathbf{I}\big) \in \mathbb{R}^{J\times(p+1)J}$ and

$$\mathbf{T} = \begin{pmatrix} \mathbf{k}^\mathsf{T} \otimes \mathbf{I}_{J\times J} & \mathbf{0}_{J\times J} \\ \mathbf{I}_{(pJ)\times(pJ)} & \mathbf{0}_{(pJ)\times J} \end{pmatrix} = \begin{pmatrix} k_1\mathbf{I} & k_2\mathbf{I} & \cdots & k_p\mathbf{I} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{(p+1)J\times(p+1)J},$$

model (2.8) can be rewritten as

$$\begin{aligned} \mathbf{z}_t &= \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^\mathsf{T}\mathbf{c}_t + \mathbf{B}_t\boldsymbol{\Theta}\mathbf{S}\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t, \\ \boldsymbol{\beta}_t &= \mathbf{T}\boldsymbol{\beta}_{t-1} + \boldsymbol{\xi}_t, \end{aligned} \tag{2.17}$$

where $\boldsymbol{\xi}_t = (\boldsymbol{\eta}_{t+p}^\mathsf{T}, \mathbf{0}, \cdots, \mathbf{0})^\mathsf{T} \sim \mathrm{N}(\mathbf{0}, \widetilde{\mathbf{H}})$ with $\widetilde{\mathbf{H}} = \mathrm{diag}(\mathbf{H}_J, \mathbf{0}, \cdots, \mathbf{0}) \in \mathbb{R}^{J(p+1)\times J(p+1)}$.

Now with the state-space model (2.17), Kalman filter and smoother (Durbin and Koopman, 2012) can be used to obtain $\widehat{\mathbf{b}}_t = \mathbb{E}(\boldsymbol{\beta}_t|\mathbf{z}_1, \ldots, \mathbf{z}_n)$ and $\mathbf{V}_t = \mathrm{Var}(\boldsymbol{\beta}_t|\mathbf{z}_1, \ldots, \mathbf{z}_n)$, and we get

$$\begin{aligned} \widehat{\boldsymbol{\alpha}}_t &= \mathbb{E}(\boldsymbol{\alpha}_t|\mathbf{z}_1, \ldots, \mathbf{z}_n, \Xi^{(0)}) = \mathbf{S}\widehat{\mathbf{b}}_t, \\ \widehat{\boldsymbol{\Sigma}}_t &= \mathrm{Var}(\boldsymbol{\alpha}_t|\mathbf{z}_1, \ldots, \mathbf{z}_n, \Xi^{(0)}) = \mathbf{S}\mathbf{V}_t\mathbf{S}^\mathsf{T}. \end{aligned}$$

Following (2.12), $\widehat{\mathbf{D}}_j$, $j = 1, \ldots, J$, can be obtained through computing $\widehat{D}_{i,k,j} := \mathbb{E}(D_{i,k,j}|\mathbf{Z}, \Xi^{(0)})$ as

$$\begin{aligned} \widehat{D}_{i,k,j} &= \mathbb{E}(\alpha_{j,i}\alpha_{j,k} + \cdots + \alpha_{j,n+1-k}\alpha_{j,n+1-i}|\mathbf{Z}, \Xi^{(0)}) \\ &= \mathbb{E}(\alpha_{j,i}|\mathbf{Z}, \Xi^{(0)})\mathbb{E}(\alpha_{j,k}|\mathbf{Z}, \Xi^{(0)}) + \cdots + \mathbb{E}(\alpha_{j,n+1-k}|\mathbf{Z}, \Xi^{(0)})\mathbb{E}(\alpha_{j,n+1-i}|\mathbf{Z}, \Xi^{(0)}) \\ &\quad + \mathrm{Cov}(\alpha_{i,j}, \alpha_{k,j}|\mathbf{Z}, \Xi^{(0)}) + \cdots + \mathrm{Cov}(\alpha_{j,n+1-k}, \alpha_{j,n+1-i}|\mathbf{Z}, \Xi^{(0)}) \\ &= \widehat{\alpha}_{j,i}\widehat{\alpha}_{j,k} + \ldots \widehat{\alpha}_{j,n+1-k}\widehat{\alpha}_{j,n+1-i} + \sum_{t=1}^{n+1-i-k} \mathbf{V}_{(1+p-i)J+j,(1+p-k)J+j,t}, \end{aligned}$$

where $\widehat{\alpha}_{j,t}$ is the j-th element of $\widehat{\boldsymbol{\alpha}}_t$ and $\mathbf{V}_{(1+p-i)J+j,(1+p-k)J+j,t}$ is the $((1+p-i)J+j, (1+p-$

$k)J + j)$-th element of $\mathbf{V}_t$, $t = 1, \ldots, n$.

The procedure of Kalman filter and smoother (Durbin and Koopman, 2012) involves first applying Kalman filter algorithm then applying Kalman smoother recursion, we give the details in the following paragraphs.

Let $\mathbf{b}_{t|t-1}$ and $\mathbf{Q}_{t|t-1}$ respectively be the one-step prediction of $\boldsymbol{\beta}_t$ and its uncertainty, i.e., $\mathbf{b}_{t|t-1} = \mathbb{E}\{\boldsymbol{\beta}_t|\mathbf{z}_{1:(t-1)}, \Xi^{(0)}\}$ and $\mathbf{Q}_{t|t-1} = \mathrm{Var}(\boldsymbol{\beta}_t|\mathbf{z}_{t-1}, \ldots, \mathbf{z}_1, \Xi^{(0)})$, $t = 1, \ldots, n$. Denote the estimation of the state at time t and its uncertainty as $\mathbf{b}_{t|t} = \mathbb{E}(\boldsymbol{\beta}_t|\mathbf{z}_t, \ldots, \mathbf{z}_1, \Xi^{(0)})$ and $\mathbf{Q}_{t|t} = \mathrm{Var}(\boldsymbol{\beta}_t|\mathbf{z}_t, \ldots, \mathbf{z}_1, \Xi^{(0)})$, respectively. The Kalman filter operates in a prediction-correction loop. The prediction step updates

$$\begin{cases} \mathbf{b}_{t|t-1} = \mathbf{T}^{(0)}\mathbf{b}_{t-1|t-1} \\ \mathbf{Q}_{t|t-1} = \mathbf{T}^{(0)}\mathbf{Q}_{t-1|t-1}(\mathbf{T}^{(0)})^{\mathsf{T}} + \widetilde{\mathbf{H}}^{(0)}, \end{cases}$$

where $\mathbf{T}^{(0)}$ and $\widetilde{\mathbf{H}}^{(0)}$ correspond to $\mathbf{T}$ and $\widetilde{\mathbf{H}}$ plugging in the current values of $\Xi^{(0)}$. In the correction step, we denote $\mathbf{F}_t = \mathbf{Q}_{t|t-1}(\mathbf{B}_t\boldsymbol{\Theta}^{(0)}\mathbf{S})^{\mathsf{T}}\{(\mathbf{B}_t\boldsymbol{\Theta}^{(0)}\mathbf{S})\mathbf{Q}_{t|t-1}(\mathbf{B}_t\boldsymbol{\Theta}^{(0)}\mathbf{S})^{\mathsf{T}} + \sigma^{2(0)}\mathbf{I}_{n_t}\}^{-1}$ as the Kalman gain matrix, and update

$$\begin{cases} \mathbf{b}_{t|t} = \mathbf{b}_{t|t-1} + \mathbf{F}_t\{\mathbf{z}_t - \mathbf{B}_t\boldsymbol{\theta}_b^{(0)}(\boldsymbol{\theta}_c^{(0)})^{\mathsf{T}}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}^{(0)}\mathbf{S}\mathbf{b}_{t|t-1}\} \\ \mathbf{Q}_{t|t} = \mathbf{Q}_{t|t-1} - \mathbf{F}_t(\mathbf{B}_t\boldsymbol{\Theta}^{(0)}\mathbf{S})\mathbf{Q}_{t|t-1}, \end{cases}$$

where $\boldsymbol{\Theta}^{(0)}$, $\sigma^{2(0)}$, $\boldsymbol{\theta}_b^{(0)}$, and $\boldsymbol{\theta}_c^{(0)}$ correspond to the current values of $\boldsymbol{\Theta}$, σ^2, $\boldsymbol{\theta}_b$, and $\boldsymbol{\theta}_c$, respectively. For the initialization, we adopt the commonly used non-informative values that $\mathbf{b}_{0|0} = \mathbf{0}$ and $\mathbf{Q}_{0|0} = \mathbf{0}$.

Next, we apply Kalman smoother recursion to obtain $\widehat{\mathbf{b}}_t$ and $\mathbf{V}_t$, $t = n-1, \ldots, 1$, using the updating formula,

$$\begin{cases} \widehat{\mathbf{b}}_t = \mathbf{b}_{t|t} + \mathbf{L}_t(\widehat{\mathbf{b}}_{t+1} - \mathbf{b}_{t+1|t}) \\ \mathbf{V}_t = \mathbf{Q}_{t|t} + \mathbf{L}_t(\mathbf{V}_{t+1} - \mathbf{Q}_{t+1|t})\mathbf{L}_t^{\mathsf{T}}, \end{cases}$$

where $\mathbf{L}_t = \mathbf{Q}_{t|t}(\mathbf{T}^{(0)})^\mathsf{T}\mathbf{Q}_{t+1|t}^{-1}$ with the initial values $\widehat{\mathbf{b}}_n = \mathbf{b}_{n|n}$ and $\mathbf{V}_n = \mathbf{Q}_{n|n}$ due to their definitions.

The M step. In M step, we find the minimizer of $Q(\Xi|\Xi^{(0)})$ in (2.16). The explicit form of minimizer is usually difficult to derive. Note that in (2.16), the parameters, $\boldsymbol{\theta}_b, \boldsymbol{\theta}_c, \boldsymbol{\Theta}, \mathbf{k}, \sigma^2$, and $\{\sigma_j^2\}_{j=1}^J$, are separated. Alternatively, we use block-wise optimization and discuss the updating rules for each parameter when the others are fixed at the current values.

The optimization problem with respect to $\boldsymbol{\theta}_b \in \mathbb{R}^{n_b}$ in (2.16) is equivalent to minimizing $(\boldsymbol{\theta}_b - \mathbf{m})^\mathsf{T}\mathbf{A}(\boldsymbol{\theta}_b - \mathbf{m})$, with the constraint that $\boldsymbol{\theta}_b^\mathsf{T}\boldsymbol{\theta}_b = 1$, where

$$\mathbf{m} = \left\{ \sum_{t=1}^{n} (\boldsymbol{\theta}_c^{(0)\mathsf{T}}\mathbf{c}_t)^2 \mathbf{B}_t^\mathsf{T}\mathbf{B}_t + \sigma^{2(0)}\lambda_{\mu_s}\boldsymbol{\Gamma} \right\}^{-1} \sum_{t=1}^{n} (\boldsymbol{\theta}_c^{(0)\mathsf{T}}\mathbf{c}_t)\mathbf{B}_t^\mathsf{T}(\mathbf{z}_t - \mathbf{B}_t\boldsymbol{\Theta}^{(0)}\widehat{\boldsymbol{\alpha}}_t)$$

and

$$\mathbf{A} = \sum_{t=1}^{n} (\boldsymbol{\theta}_c^{(0)\mathsf{T}}\mathbf{c}_t)^2 \mathbf{B}_t^\mathsf{T}\mathbf{B}_t + \sigma^{2(0)}\lambda_{\mu_s}\boldsymbol{\Gamma}.$$

By treating the sphere of $\boldsymbol{\theta}_b$ as an embedded sub-manifold of the n_b-dimensional Euclidean space, $\boldsymbol{\theta}_b$ can be updated using the gradient decent algorithm on a sub-manifold (Absil et al., 2009). The gradient descent iteration for a sub-manifold includes four steps:

i calculate the negative gradient of the objective function in the Euclidean space without any constraint;

ii project the obtained negative gradient function onto the tangent space of manifold;

iii evaluate the updating value along the direction of the projected negative gradient in step ii with a given step size;

iv retract the calculated value in step iii back to the manifold structure.

The step size in the above step iii can be determined using the Armijo backtracking method (see, for example, Chapter 4.2 of Absil et al., 2009). Algorithm 1 specializes our implementation for using the gradient decent algorithm on the sphere manifold to solve the optimization problem (2.16) with respect to $\boldsymbol{\theta}_b$ with more details therein.

Algorithm 1 Gradient decent algorithm to update $\boldsymbol{\theta}_b$.

Require: $\mathbf{A}, \mathbf{m}$, and scalars $\beta, \gamma \in (0,1)$. Initialization $\widehat{\boldsymbol{\theta}}_b^{(0)}$.
for $k = 1, 2, \ldots$ **do**
i) Compute $\boldsymbol{\eta}_k = -2\big\{\mathbf{A}(\widehat{\boldsymbol{\theta}}_b^{(k-1)} - \mathbf{m}) - \widehat{\boldsymbol{\theta}}_b^{(k-1)}\big(\widehat{\boldsymbol{\theta}}_b^{(k-1)}\big)^{\mathsf{T}}\mathbf{A}(\widehat{\boldsymbol{\theta}}_b^{(k-1)} - \mathbf{m})\big\}$.
ii) Find the smallest integer $n \geq 0$ such that $f\{R_{\widehat{\boldsymbol{\theta}}_b^{(k-1)}}(\beta^n\boldsymbol{\eta}_k)\} \leq f(\widehat{\boldsymbol{\theta}}_b^{(k-1)}) - \gamma\beta^n\boldsymbol{\eta}_k^{\mathsf{T}}\boldsymbol{\eta}_k$, where $R_{\widehat{\boldsymbol{\theta}}_b^{(k-1)}}(\beta^n\boldsymbol{\eta}_k) = (\widehat{\boldsymbol{\theta}}_b^{(k-1)} + \beta^n\boldsymbol{\eta}_k)/\|\widehat{\boldsymbol{\theta}}_b^{(k-1)} + \beta^n\boldsymbol{\eta}_k\|$.
iii) $\widehat{\boldsymbol{\theta}}_b^{(k)} = R_{\widehat{\boldsymbol{\theta}}_b^{(k-1)}}(\beta^n\boldsymbol{\eta}_k)$.
iv) Repeat until $\|\widehat{\boldsymbol{\theta}}_b^{(k)} - \widehat{\boldsymbol{\theta}}_b^{(k-1)}\|$ is small enough.
end for
return $\widehat{\boldsymbol{\theta}}_b = \widehat{\boldsymbol{\theta}}_b^{(k)}$.

We propose to update $\boldsymbol{\theta}_c$, σ^2, and σ_j^2, $j = 1, \ldots, J$, by setting the corresponding block-wise derivatives to be zero. To be specific, by taking derivative of (2.16) with respect to $\boldsymbol{\theta}_c$ and set it to zero, we update $\boldsymbol{\theta}_c$ by

$$\widehat{\boldsymbol{\theta}}_c = \Big\{\sum_{t=1}^{n}(\mathbf{B}_t\widehat{\boldsymbol{\theta}}_b\mathbf{c}_t^{\mathsf{T}})^{\mathsf{T}}(\mathbf{B}_t\widehat{\boldsymbol{\theta}}_b\mathbf{c}_t^{\mathsf{T}}) + \sigma^{2(0)}\lambda_{\mu_t}\mathbf{P}\Big\}^{-1}\sum_{t=1}^{n}(\mathbf{B}_t\widehat{\boldsymbol{\theta}}_b\mathbf{c}_t^{\mathsf{T}})^{\mathsf{T}}(\mathbf{z}_t - \mathbf{B}_t\boldsymbol{\Theta}^{(0)}\widehat{\boldsymbol{\alpha}}_t).$$

Analogously, the updating formula for σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{\sum_{t=1}^{n} n_t}\sum_{t=1}^{n}\Big\{(\mathbf{z}_t - \mathbf{B}_t\widehat{\boldsymbol{\theta}}_b\widehat{\boldsymbol{\theta}}_c^{\mathsf{T}}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}^{(0)}\widehat{\boldsymbol{\alpha}}_t)^{\mathsf{T}}(\mathbf{z}_t - \mathbf{B}_t\widehat{\boldsymbol{\theta}}_b\widehat{\boldsymbol{\theta}}_c^{\mathsf{T}}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}^{(0)}\widehat{\boldsymbol{\alpha}}_t) + \operatorname{tr}(\mathbf{B}_t\boldsymbol{\Theta}^{(0)}\widehat{\boldsymbol{\Sigma}}_t\boldsymbol{\Theta}^{(0)\mathsf{T}}\mathbf{B}_t^{\mathsf{T}})\Big\},$$

and the updating formula for σ_j^2 is $\widehat{\sigma}_j^2 = \widehat{S}_j(\mathbf{k}^{(0)})/n$, $j = 1, \ldots, J$, where $\widehat{S}_j(\mathbf{k}^{(0)})$ is defined in (2.15).

For $\boldsymbol{\Theta}$, we first update the columns of $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)$ sequentially. Minimizing (2.16) with respect to $\boldsymbol{\theta}_j$ is equivalent to minimizing

$$\sum_{t=1}^{n}\left\|\mathbf{z}_t - \mathbf{B}_t\widehat{\boldsymbol{\theta}}_b\widehat{\boldsymbol{\theta}}_c^{\mathsf{T}}\mathbf{c}_t - \sum_{j'\neq j}\mathbf{B}_t\boldsymbol{\theta}_{j'}\widehat{\alpha}_{tj'} - \mathbf{B}_t\boldsymbol{\theta}_j\widehat{\alpha}_{tj}\right\|^2 + \sum_{t=1}^{n}\operatorname{tr}(\mathbf{B}_t\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_t\boldsymbol{\Theta}^{\mathsf{T}}\mathbf{B}_t^{\mathsf{T}}) + \sigma^2\lambda_{pc}\boldsymbol{\theta}_j^{\mathsf{T}}\boldsymbol{\Gamma}\boldsymbol{\theta}_j,$$

which has an analytical form

$$\widehat{\boldsymbol{\theta}}_j = \Big\{ \sum_{t=1}^{n} (\widehat{\alpha}_{tj}^2 + \widehat{\boldsymbol{\Sigma}}_{t,jj}) \mathbf{B}_t^{\mathsf{T}} \mathbf{B}_t + \widehat{\sigma}^2 \lambda_{pc} \boldsymbol{\Gamma} \Big\}^{-1} \times \sum_{t=1}^{n} \mathbf{B}_t^{\mathsf{T}} \Big\{ (\mathbf{z}_t - \mathbf{B}_t \widehat{\boldsymbol{\theta}}_b \widehat{\boldsymbol{\theta}}_c^{\mathsf{T}} \mathbf{c}_t) \widehat{\alpha}_{tj} - \sum_{j' \neq j} (\widehat{\alpha}_{tj'} \widehat{\alpha}_{tj} + \widehat{\boldsymbol{\Sigma}}_{t,j'j}) \mathbf{B}_t \widehat{\boldsymbol{\theta}}_{j'} \Big\}.$$

To guarantee the orthonormality of $\widehat{\boldsymbol{\Theta}}$, we utilize the spectral decomposition of $\widehat{\boldsymbol{\Theta}} \widehat{\mathbf{H}}_J \widehat{\boldsymbol{\Theta}}^{\mathsf{T}}$, where $\widehat{\mathbf{H}}_J = \mathrm{diag}(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_J^2)$. To be specific, let $\widehat{\boldsymbol{\Theta}} \widehat{\mathbf{H}}_J \widehat{\boldsymbol{\Theta}}^{\mathsf{T}} = \widetilde{\mathbf{Q}} \widetilde{\mathbf{D}} \widetilde{\mathbf{Q}}^{\mathsf{T}}$, where $\widetilde{\mathbf{Q}}$ is orthonormal and $\widetilde{\mathbf{D}}$ is a diagonal matrix with decreasing diagonal elements. We then replace $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\mathbf{H}}_J$ by $\widetilde{\mathbf{Q}}$ and $\widetilde{\mathbf{D}}$, respectively. Furthermore, replace $\widehat{\boldsymbol{\alpha}}_t$ with $\widetilde{\mathbf{Q}}^{\mathsf{T}} \widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\alpha}}_t$, such transformation preserves the variance of $\widehat{\boldsymbol{\Theta}} \widehat{\boldsymbol{\alpha}}_t$.

Finally, we aim to minimize (2.16) with respect to $\mathbf{k}$, which is equivalent to minimizing

$$\sum_{j=1}^{J} \Big\{ -\log |\mathbf{M}_j| + \frac{1}{\widehat{\sigma}_j^2} \widehat{S}_j(\mathbf{k}) \Big\}, \tag{2.18}$$

where $\mathbf{M}_j$ and $\widehat{S}_j(\mathbf{k})$ are given in (2.11) and (2.15), respectively. Note that the value of $\log |\mathbf{M}_j|$ is invariant with the change of sample size n, while the second term in (2.18) is n-dependent. To simplify the computation, we use the second term $\sum_{j=1}^{J} \{(1/\widehat{\sigma}_j^2) \widehat{S}_j(\mathbf{k})\}$ to approximate (2.18). Since $\widehat{S}_j(\mathbf{k}) = (1, \mathbf{k}^{\mathsf{T}}) \widehat{\mathbf{D}}_j (1, \mathbf{k}^{\mathsf{T}})^{\mathsf{T}}$ has a quadratic form with respect to $\mathbf{k}$, using the weighted least squares, we obtain

$$\widehat{\mathbf{k}} = \Big(\sum_{j=1}^{J} \frac{1}{\widehat{\sigma}_j^2} \widehat{\mathbf{D}}_{pj} \Big)^{-1} \sum_{j=1}^{J} \frac{1}{\widehat{\sigma}_j^2} \widehat{\mathbf{d}}_j,$$

where $\widehat{\mathbf{d}}_j = (\widehat{D}_{12,j}, \ldots, \widehat{D}_{1(p+1),j})^{\mathsf{T}}$ and $\widehat{\mathbf{D}}_{pj}$ is the bottom right $p \times p$ major submatrix of $\widehat{\mathbf{D}}_j$.

2.2.3 Model Selection

The general guideline for constructing triangulation is that we should avoid having triangles with a very small interior angle and that there is no triangle that contains no data point. We refer to Chapter 4 of Lai and Schumaker (2007) for a detailed discussion of triangulations. When the

penalized spline method is used, the number of basis functions is not crucial in many applications as long as it is moderately large, since the roughness penalty helps regularize the estimation and prevent overfitting (Ruppert et al., 2003). Furthermore, the smoothness of the 2-dimensional basis functions $\mathbf{b}(x, y)$ is determined by the order d of polynomials and the order r of the smoothness parameter on the connected edges of triangulations. Practically, these two orders can also be given by the users based on the prior knowledge of the data. The number of principal component functions is determined by the empirical proportion of variances of temporal FPC scores. The order p of autoregressive model for the latent variables can be selected using a data-driven criteria like Akaike information criteria (AIC, Akaike, 1974) or Bayesian information criteria (BIC, Schwarz, 1978), such that p minimizes

$$\sum_{j=1}^{J} \left\{ n \log \widehat{\sigma}_j^2 + \frac{1}{\widehat{\sigma}_j^2} \widehat{S}_j(\widehat{\mathbf{k}}) \right\} + 2p,$$

or

$$\sum_{j=1}^{J} \left\{ n \log \widehat{\sigma}_j^2 + \frac{1}{\widehat{\sigma}_j^2} \widehat{S}_j(\widehat{\mathbf{k}}) \right\} + \log(n)p.$$

The regularization parameters λ_{μ_s}, λ_{μ_t}, and λ_{pc} can be determined by minimizing the value of K-fold leave-location-out cross validation (CV), with a typical choice of $K = 5$ or $K = 10$. Nevertheless, since there are three regularization parameters, the classical full grid-search will be impractical due to high computational cost. Alternatively, we propose to use the simplex method (Nelder and Mead, 1965) to find the local optima. The overall selection procedure for the regularization parameters $(\lambda_{\mu_s}, \lambda_{\mu_t}, \lambda_{pc})$ contains two steps. In the first step, we assign a few number of grid points sparse enough to cover a wide range of regularization parameters, and apply k-fold CV to determine the best candidate according to the crossed predictive errors. Afterwards, we treat the selected point as the initial value and use the simplex method to obtain the final selected regularization parameters with local optimality.

2.3 Simulation Studies

In this section, we present the results of a simulation study to assess the performance of the proposed tFPC method and compare with that of mFPC in Zhou and Pan (2014a). The irregular domain Ω was set to be a 2×2 square with a hole in the middle, as shown in Figure 2.1. We generated data on this region according to model (2.4) and (3.3) with $J = 2$ principal components and order of the AR model $p = 2$. The mean function and the PC functions are given as follows,

$$\mu_t(x, y) = \mu_1(x, y)\mu_2(t),$$

where

$$\begin{aligned}
\mu_1(x, y) &= 5\big\{ \exp\big(\sqrt{0.1x^2 + 0.2y}\big) + \exp\big(- \sqrt{0.1x^2 + 0.2y}\big)\big\},\\
\mu_2(t) &= 1 \ \text{ or } \ \mu_2(t) = \cos(2\pi t/12) + t/n,\\
\phi_1(x, y) &= 0.8578 \sin(x^2 + 0.5y^2),\\
\phi_2(x, y) &= 0.8721 \sin(0.3x^2 + 0.6y^2) - 0.2988 \sin(x^2 + 0.5y^2).
\end{aligned}$$

Note that the PC functions are orthonormal such that $\int_\Omega \phi_1^2(x, y)\mathrm{d}x\mathrm{d}y = 1$, $\int_\Omega \phi_2^2(x, y)\mathrm{d}x\mathrm{d}y = 1$ and $\int_\Omega \phi_1(x, y)\phi_2(x, y)\mathrm{d}x\mathrm{d}y = 0$.

In the simulation study, we considered four setups: i) $\mu_2(t) = \cos(2\pi t/12) + t/n$ with AR(2) coefficients $k_1 = 0.8$ and $k_2 = 0.1$; ii) $\mu_2(t) = \cos(2\pi t/12) + t/n$ with AR(2) coefficients $k_1 = k_2 = 0$; iii) $\mu_2(t) = 1$ with AR(2) coefficients $k_1 = 0.8$ and $k_2 = 0.1$; iv) $\mu_2(t) = 1$ with AR(2) coefficients $k_1 = k_2 = 0$. Note that in setup (ii) and (iv) the AR(2) model degenerates to a white noise model such that FPC scores are independent. In each setup, we used two levels of variances: $\sigma^2 = 1, (\sigma_1^2, \sigma_2^2) = (1, 0.1)$; or $\sigma^2 = 0.1, (\sigma_1^2, \sigma_2^2) = (0.1, 0.01)$. To simulate a data set, we set the number of time points $n = 500$. At each time t, $t = 1, \cdots, 500$, the number of observed locations was drawn from $\{50, 51, \ldots, 60\}$ uniformly and each location was randomly drawn from a uniform distribution on the irregular domain. We ran the simulation 100 times for each combination of the four setups and the two levels of variances. Both the proposed tFPC model and the mFPC model to were applied on each simulated data.

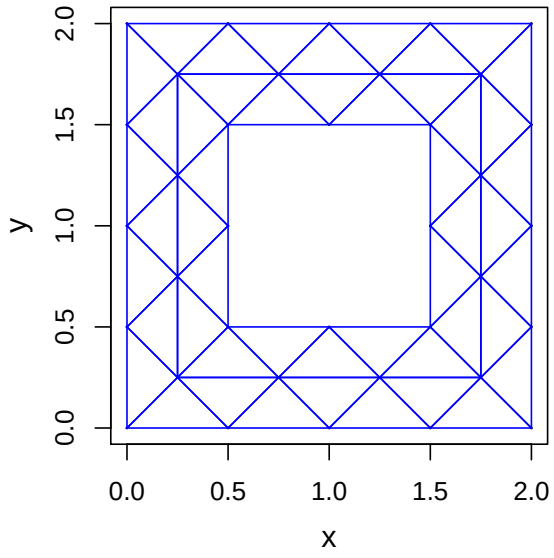

Figure 2.1: An example of triangularization: a square with a hole in the middle used in the simulation study.

Since the mFPC model assumes a mean function constant over t, to apply the mFPC model in setup i) and ii), we took a two-step approach: 1) estimated the mean function $\mu_t(x, y)$ by basis expansion, and 2) applied the mFPC model on the residuals. To construct the basis functions in step 1), We first ignored the location effect and constructed a crude estimate of overall time effect by regressing the data on time domain basis functions 1, t, plus Fourier basis functions $\sin(\frac{2k\pi t}{12})$, $\cos(\frac{2k\pi t}{12})$, $k = 1, \cdots, 10$; the fitted function is defined as $\tilde{\nu}(t)$. The basis functions for estimating $\mu_1(x, y)\mu_2(t)$ were generated by multiplying $\tilde{\nu}(t)$ with bivariate basis functions on the domain Ω described in next paragraph.

For both approaches, on the domain Ω, the triangulations were presented in Figure 2.1. The bivariate basis functions were constructed from Bernstein polynomials with $d = 3$ (cubic order splines) and $r = 1$ (continuous first derivative across the connected edges), same as that in Zhou and Pan (2014b). See Section 2.1.1 of Supplementary Materials for details on basis construction. On the time domain, in setup i) and ii), to model $\mu_2(t)$ with tFPC approach, we used basis functions 1, t, plus Fourier basis functions $\sin(\frac{2k\pi t}{12})$, $\cos(\frac{2k\pi t}{12})$, $k = 1, \cdots, 10$.

Setup	$\sigma^2, (\sigma_1^2, \sigma_2^2)$	Model	PA	MIAE $\mu_t(x, y)$	MIAE $Z_t(x, y)$
i)	1.0, (1.0, 0.1)	tFPC	4.8426 (0.0809)	0.1206 (0.0054)	0.1433 (0.0003)
		mFPC	6.4969 (0.3397)	0.2189 (0.0089)	0.1805 (0.0023)
	0.1, (0.1, 0.01)	tFPC	4.7622 (0.0780)	0.0377 (0.0017)	0.0452 (0.0001)
		mFPC	8.3236 (0.3948)	0.0737 (0.0032)	0.0589 (0.0008)
ii)	1.0, (1.0, 0.1)	tFPC	9.4752 (0.1582)	0.0478 (0.0011)	0.1409 (0.0003)
		mFPC	19.142 (0.8928)	0.0697 (0.0008)	0.1590 (0.0009)
	0.1, (0.1, 0.01)	tFPC	9.4025 (0.1642)	0.0145 (0.0003)	0.0442 (0.0001)
		mFPC	23.103 (0.3607)	0.0256 (0.0003)	0.0523 (0.0002)
iii)	1.0, (1.0, 0.1)	tFPC	4.6808 (0.0778)	0.1014 (0.0055)	0.1359 (0.0003)
		mFPC	5.1745 (0.2513)	0.1341 (0.0083)	0.1525 (0.0019)
	0.1, (0.1, 0.01)	tFPC	4.7334 (0.0775)	0.0543 (0.0029)	0.0430 (0.0001)
		mFPC	4.9064 (0.0920)	0.0603 (0.0038)	0.0475 (0.0001)
iv)	1.0, (1.0, 0.1)	tFPC	9.3640 (0.1545)	0.0291 (0.0011)	0.1362 (0.0003)
		mFPC	11.770 (0.6829)	0.0429 (0.0022)	0.1398 (0.0008)
	0.1, (0.1, 0.01)	tFPC	9.4162 (0.1608)	0.0108 (0.0003)	0.0434 (0.0001)
		mFPC	11.707 (0.7475)	0.0110 (0.0003)	0.0442 (0.0002)

Table 2.1: The means and standard errors of PAs and MISEs for the mean function $\mu_t(x, y)$ and the stochastic surface $Z_t(x, y)$. The results are based on 100 simulation runs.

In setup iii) and iv) the mean function is constant over t, which is the assumption in the mFPC model, for fair comparison between the two approaches, we modified the model fitting procedure in tFPC approach by assuming that $\mu_2(t) = 1$ is known.

For simplicity, the number of PCs and the order of AR in our simulation study were set to be the same as the true ones. We selected the three penalty parameters $(\lambda_{\mu_s}, \lambda_{\mu_t}, \lambda_{pc})$ by minimizing the value of 5-fold leave-location-out CV with simplex method.

To quantitatively measure the performance of the estimation of the mean function and the stochastic surfaces, we use the mean integrated absolute errors (MIAE) defined as

$$\frac{1}{n}\sum_{t=1}^{n}\int_{\Omega}|f(x,y,t)-\widehat{f}(x,y,t)|\mathrm{d}x\mathrm{d}y,$$

where the integration is evaluated as a scaled sum over 1976 grid points distributed evenly on the spatial domain (the grid points are 51×51 points evenly distributed on the rectangle with those in the hole been removed).

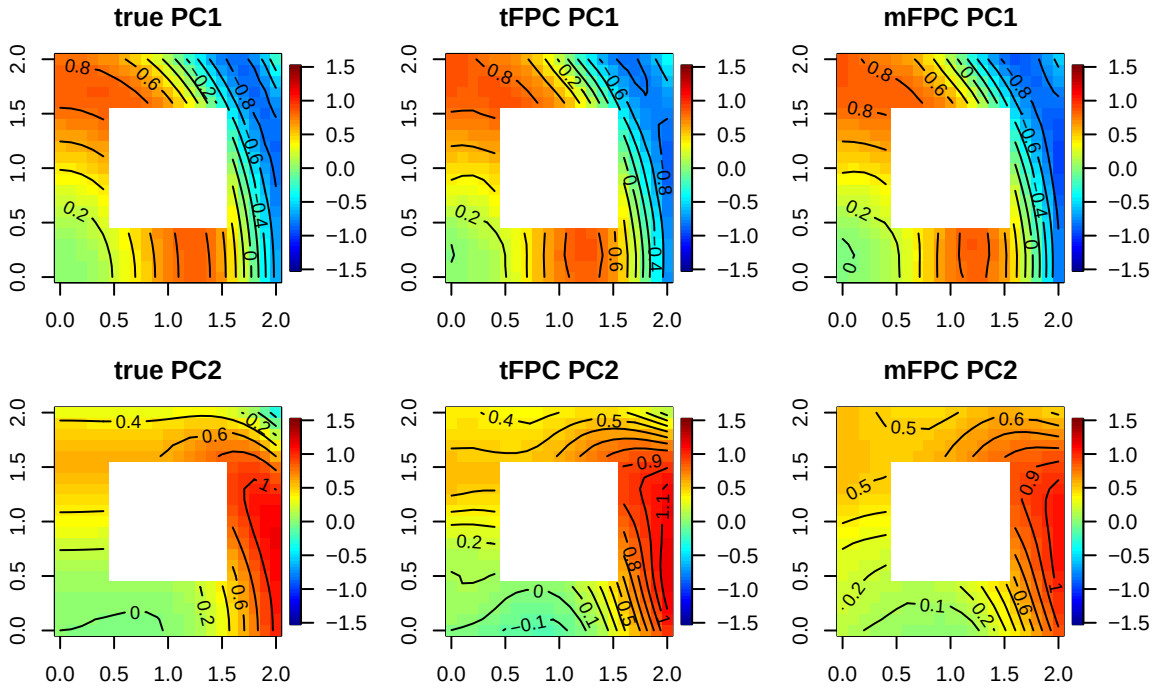

Figure 2.2: The temporal principal component functions in the first setting of simulation study. The first and second principal functions are depicted in the first and second rows, respectively. From left to right are the true PC functions, the estimation of the proposed tFPC, and of the alternative mFPC.

To evaluate the performance of the estimations of the principal component functions, We use the principal angle (PA) as follows: We first evaluated the principal component functions on 1976 grid points evenly distributed over the domain and obtained two matrices $\widehat{\mathbf{V}}$ and $\mathbf{V}$, corresponding to the estimated and the true principal component functions. We then compute the principal angle as $\text{angle} = \cos^{-1}(\rho) \times 180/\pi$ where ρ is the minimum singular value of the matrix $\mathbf{Q}_{\widehat{\mathbf{V}}}^{\mathsf{T}}\mathbf{Q}_{\mathbf{V}}$ with $\mathbf{Q}_{\widehat{\mathbf{V}}}$ and $\mathbf{Q}_{\mathbf{V}}$ being the orthonormal matrices of the QR decomposition of $\widehat{\mathbf{V}}$ and $\mathbf{V}$, respectively (Golub and Van Loan, 2013).

The means and standard errors of MIAEs of the mean function and stochastic surfaces and PAs of the PC functions from the 100 simulation runs are given in Table 2.1. It is clear from this table that the tFPC model outperforms the mFPC model in setup i), ii) and iii) where the tFPC model gives smaller average MIAEs and PAs. In setup iv), the two models have smaller average PAs but

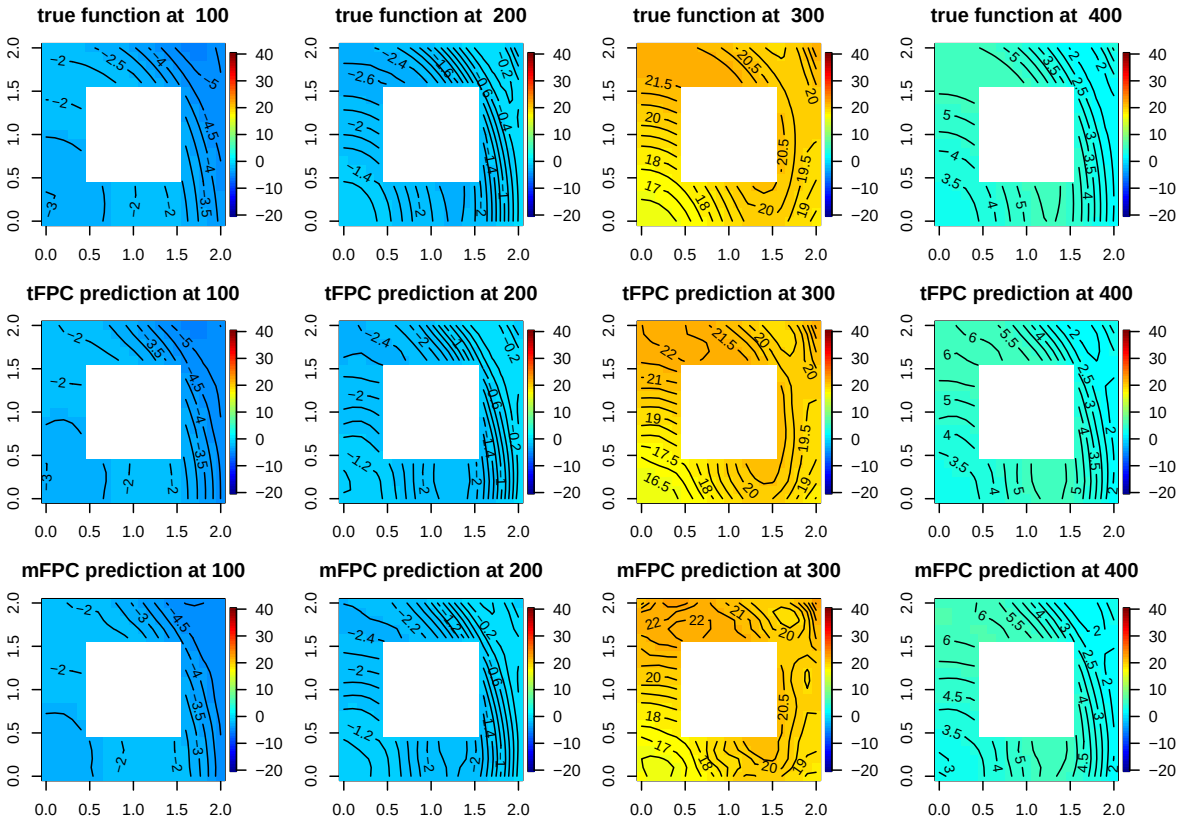

Figure 2.3: The individual functions in first setting of simulation study. From left to right are the contours at time points $t = 100$, 200, 300, and 400. The true function, the estimation of the proposed tFPC, and of the alternative mFPC are depicted in the first, second, and third rows, respectively.

similar results in mean function and stochatic surface estimation.

The contour plots and heat maps of the true and estimated temporal principal component functions of one random replicate in setting i) with $\sigma^2 = 0.1$ are depicted in Figure 2.2. Compared with mFPC, the estimated temporal PC functions using proposed tFPC is slightly closer to the true ones. We then depict the contour plot and heat map of the true individual functions and their best linear unbiased prediction (BLUP) by two methods in Figure 2.3. It is confirmed that the results of tFPC is consistently better in the sense that its predicted contours are more precise.

Finally, the estimated real-valued parameters, including the autoregressive coefficients and the variances of time series and observational noise, are summarized in Tables 2.2 and 2.3 for our

Setting	$\widehat{k}_1$ ($k_1 = 0.8$ or 0)	$\widehat{k}_2$ ($k_2 = 0.1$ or 0)	$\widehat{\sigma}^2$ ($\sigma^2 = 1$)	$\widehat{\sigma}_1^2$ ($\sigma_1^2 = 1$)	$\widehat{\sigma}_2^2$ ($\sigma_2^2 = 0.1$)
i)	0.8342 (0.0059)	0.0594 (0.0053)	0.9994 (0.0009)	1.0085 (0.0080)	0.0928 (0.0018)
ii)	0.0027 (0.0041)	0.0009 (0.0040)	0.9978 (0.0009)	0.9878 (0.0069)	0.0952 (0.0010)
iii)	0.8088 (0.0046)	0.0864 (0.0045)	1.0001 (0.0009)	0.9994 (0.0077)	0.0972 (0.0011)
iv)	-0.0002 (0.0042)	-0.0017 (0.0040)	0.9991 (0.0009)	0.9903 (0.0070)	0.0956 (0.0010)

Table 2.2: The performance of parameters estimating in the simulation study for the noise level $\sigma^2 = 1$ and $(\sigma_1^2, \sigma_2^2) = (1, 0.1)$. Reported are the means of estimations and the standard errors (in parenthesis) based on 100 data replications.

Setting	$\widehat{k}_1$ ($k_1 = 0.8$ or 0)	$\widehat{k}_2$ ($k_2 = 0.1$ or 0)	$\widehat{\sigma}^2$ ($\sigma^2 = 0.1$)	$\widehat{\sigma}_1^2$ ($\sigma_1^2 = 0.1$)	$\widehat{\sigma}_2^2$ ($\sigma_2^2 = 0.01$)
i)	0.8440 (0.0056)	0.0528 (0.0051)	0.0998 (0.0001)	0.1007 (0.0008)	0.0086 (0.0002)
ii)	0.0011 (0.0042)	-0.0002 (0.0040)	0.0997 (0.0009)	0.0989 (0.0007)	0.0094 (0.0001)
iii)	0.8063 (0.0044)	0.0865 (0.0043)	0.0998 (0.0001)	0.0997 (0.0008)	0.0098 (0.0001)
iv)	-0.0015 (0.0042)	-0.0027 (0.0039)	0.0998 (0.0001)	0.0989 (0.0007)	0.0096 (0.0001)

Table 2.3: The performance of parameters estimating in the simulation study for the noise level $\sigma^2 = 0.1$ and $(\sigma_1^2, \sigma_2^2) = (0.1, 0.01)$. Reported are the means of estimations and the standard errors (in parenthesis) based on 100 data replications.

proposed tFPC model. Overall, it shows that the parameters are estimated accurately in all settings using tFPC model. Note that settings ii) and iv) are the cases that the PC scores are i.i.d. random variables. Our proposed tFPC model is able to estimate the autoregressive coefficients close to zeros implying the robustness of this method.

2.4 Texas Temperature Data Analysis

In this section, we apply the proposed model to study the climate change of Texas by analyzing Texas temperature data downloaded from United States Historical Climatology Network, Version 2.5 (USHCN v2.5, `https://cdiac.ess-dive.lbl.gov/epubs/ndp/ushcn/ushcn.html`). Our data set consists of monthly average temperatures from January 1915 to December 2014 recorded by 49 weather stations located in Texas. The locations of these weather stations are shown in Figure 2.4.

With an area of 696,200 km^2, Texas has diverse physical geography and climate types. In general, in the eastern half of Texas, where lie the Gulf Coastal Plains and the North Central Plains,

the climate is humid subtropical; in the western half, where lie the deserts and tall mountains, climate is semi-arid. Due to various reasons, 6.82% of the data were missing. In particular, only 3 stations have complete records while about 13 stations miss more than 120 months of data. There is no clear pattern in the missing of the data.

To apply the proposed tFPC model on the data, we first smoothed the data by removing the location effect and time effect. We took two-step method. We first removed the location effect by fitting a nonparametric regression $z_t(x,y) = \mu(x,y) + \epsilon_t(x,y) = \mathbf{b}(x,y)^\mathsf{T}\boldsymbol{\theta}_\mu + \epsilon_t(x,y)$, where we utilized the same Bernstein polynomial basis on triangulations as the bivariate basis $\mathbf{b}(x,y)$. We used the penalized least squares approach with the roughness penalty matrix $\boldsymbol{\Gamma}$ discussed in Section 2.2. While the penalty parameter was selected by cross-validation. After removing the location effect and denoting $\widetilde{z}_t(x,y) = z_t(x,y) - \mathbf{b}(x,y)^\mathsf{T}\widehat{\boldsymbol{\theta}}_b$, we applied another nonparametric regression such that $\widetilde{z}_t(x,y) = \nu(t) + \eta_t(x,y) = \mathbf{c}(t)^\mathsf{T}\boldsymbol{\theta}_\nu + \eta_t(x,y)$. Similarly, we used the penalized least squares to obtain the time effect $\mathbf{c}(t)^\mathsf{T}\widehat{\boldsymbol{\theta}}_\nu$ and removed it from $\widetilde{z}_t(x,y)$. Finally, we obtained the data without both location effect and time effect $z_t(x,y)^{\text{demean}} = z_t(x,y) - \mathbf{b}(x,y)^\mathsf{T}\widehat{\boldsymbol{\theta}}_\mu - \mathbf{c}(t)^\mathsf{T}\widehat{\boldsymbol{\theta}}_\nu$.

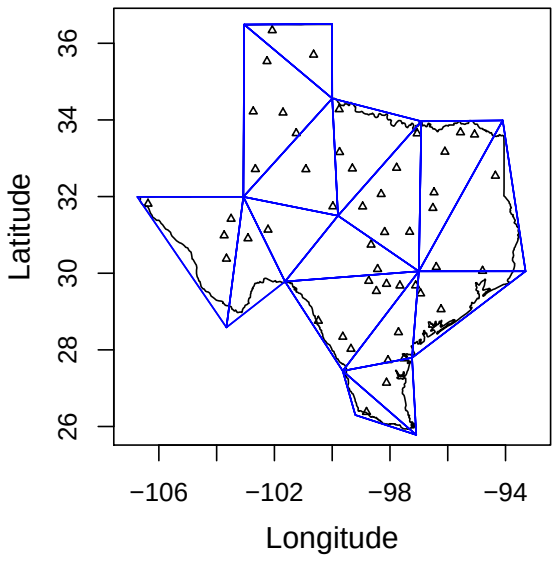

Figure 2.4: Triangulation used in the application to the Texas temperature data analysis.

The triangulations were constructed as in Figure 2.4 to cover the irregular domain of Texas and that there are some weather stations in each triangle. On these triangles, the Bernstein polynomial bivariate splines were constructed with $d = 3$ (cubic order splines) and $r = 1$ (continuous first derivative across the connected edges) as used in Zhou and Pan (2014b). As for the temporal basis, considered the seasonal effect and the climate change over the 100 years, we used the Fourier basis functions $\sin(\frac{2k\pi t}{12})$, $\cos(\frac{2k\pi t}{12})$, $k = 1, \cdots, 25$, plus cubic polynomials. Following Section 3.4, by checking the empirical proportions of variances of temporal FPC scores, the number of principal components was chosen as $J = 3$. Using a criteria like AIC, we selected the order of autoregressive model to be $p = 4$. The three penalty parameters $\lambda_{\mu_s}, \lambda_{\mu_t}$, and λ_{pc} were selected by 5-fold CV with simplex method.

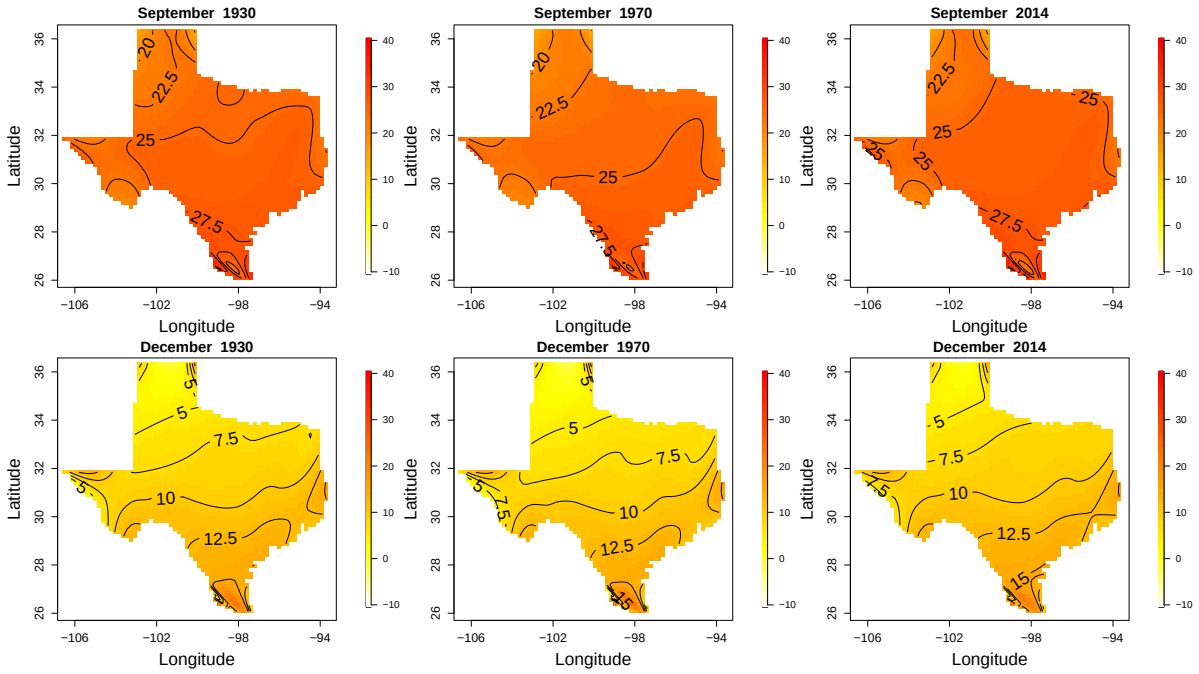

Figure 2.5: The estimated means of September, December in 1930, 1970, 2014, respectively.

The top row of Figure 2.5 depicts the estimated mean functions $\mu(x_i, y_i)+\nu(t_j)+\mu_1(x_i, y_i)\mu_2(t_j)$ (fixed t_j, and varied (x_i, y_i) within the domain) in September on 1930, 1970, 2014, respectively.

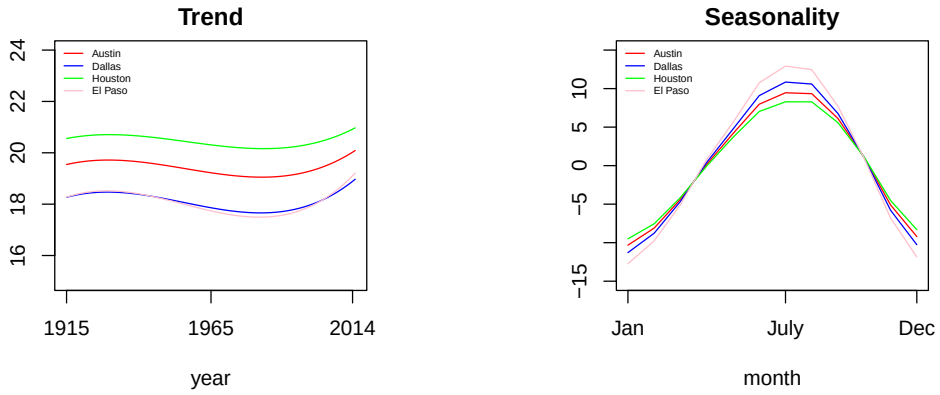

Figure 2.6: Left: the estimated trends of different cities. Right: the estimated one period of different cities.

While the bottom row of Figure 2.5 presents the estimated mean functions in December on the same years. The contours of these figures show that the overall temperature has a decreasing trend from 1930 to 1970, while there is a increasing trend from 1970 to 2014.

The left panel of Figure 2.6 shows the overall trends $\mu(x_i, y_i) + \nu(t) + \mu_1(x_i, y_i)\mu_2(t)$ (fixed (x_i, y_i), varied t within 1200 months.) of Austin, Dallas, Houston, and El Paso. The average temperature trends generally decrease when the latitudes of cities increase, while El Paso has a similar trend with Dallas. Besides, the right panel of Figure 2.6 depicts one-period of different cities. As we can see from this figure, the cities with lower latitude generally have smaller ranges in one period. While El Paso has a larger temperature range within one period compared with Dallas (which has higher latitude). It may be caused by the location in mountain region of El Paso compared with the locations in plain region of other cities.

To show the main patterns of spatial variation, we also depict the contour plot and heat map of the estimated principal component functions in Figure 2.7. The first PC surface varies slightly almost over the whole state of Texas, which indicates the average deviation of the temperature in Texas. The second PC surface varies in parallel with the latitude of Texas, which indicates that the variation of the temperature is due to the change of latitude. As we know, the altitude of western

Texas gradually decrease forward to eastern, while the eastern Texas is a plain. The contour lines of the third PC surface varies the same with the geospatial location of Texas meaning that the third PC function may be considered as the altitude effects.

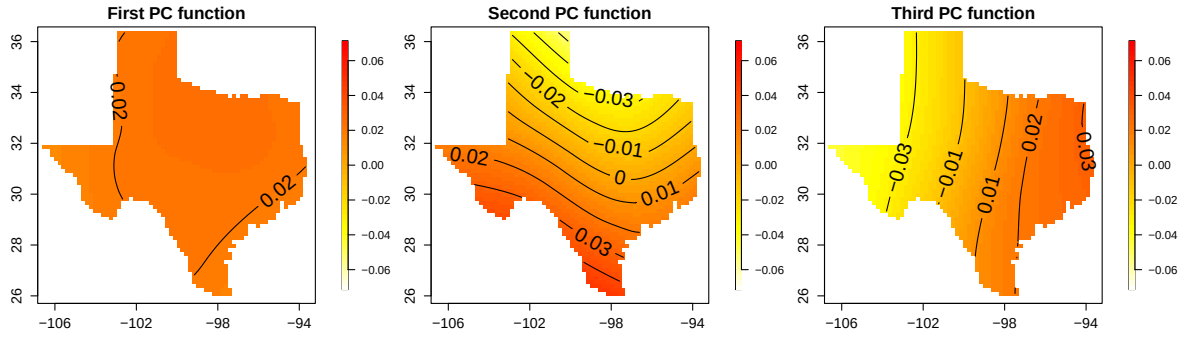

Figure 2.7: The first three principal component functions for the real data analysis.

To compare the proposed method of tFPC with the alternative mFPC model, we left one weather station at Albany (location: 99.27W, 32.72N) out as the predicted target and used the temperature data of the other 48 weather stations as the training data. In other words, the monthly temperature at Albany was predicted along 100 years by tFPC model trained from other weather stations. To apply mFPC model on the training data, note that the change of Texas temperature is more than 20 degrees Celsius from winter to summer, and i.i.d assumption on mFPC model may be violated. For fair comparison, we thus separated the data in each month to train mFPC model. After fitting the models and comparing with the observed values in the testing weather station, we calculated the prediction error (PE). Figure 2.8 depicts boxplot of PE of the proposed tFPC and the alternative mFPC for each month. It shows that tFPC outperforms mFPC in most months such as January to March and October to December, and has the similar results in the other months.

We also compare the performance of short-term forecasting using tFPC and mFPC models. The training set of tFPC model is the first 99 years and the test set is the following three months. In other words, we applied the tFPC model to the data of training set (January 1915–December

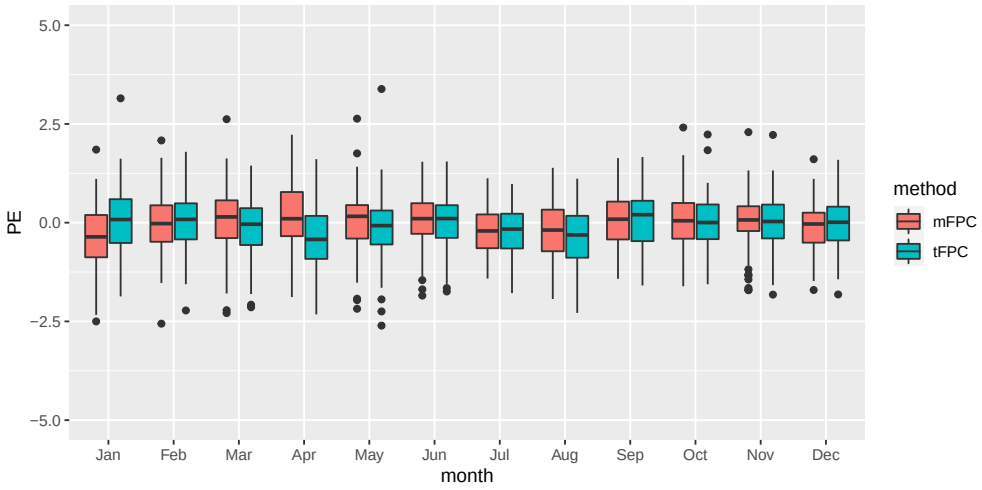

Figure 2.8: Monthly prediction error (PE) at New Brunfels for two methods.

2013), and used the trained model to predict the temperature of 49 weather stations for January 2014–March 2014. To be specific, the temporal FPC scores were forecasted according to the trained AR(p) model (2.5). For the mean effect, we adopted the extrapolation of Fourier basis on the time domain due to the product form (2.4) of the mean function. As for mFPC model, we separated the monthly temperature average data for January, February, and March from the first 99 years to train mFPC models monthly, and predicted the temperatures of the following three months. We compared the predicted values of temperature with the true ones of 49 weather stations. The results of the residual boxplots are presented in Figure 2.9, which confirms the advantage of using tFPC in terms of forecasting.

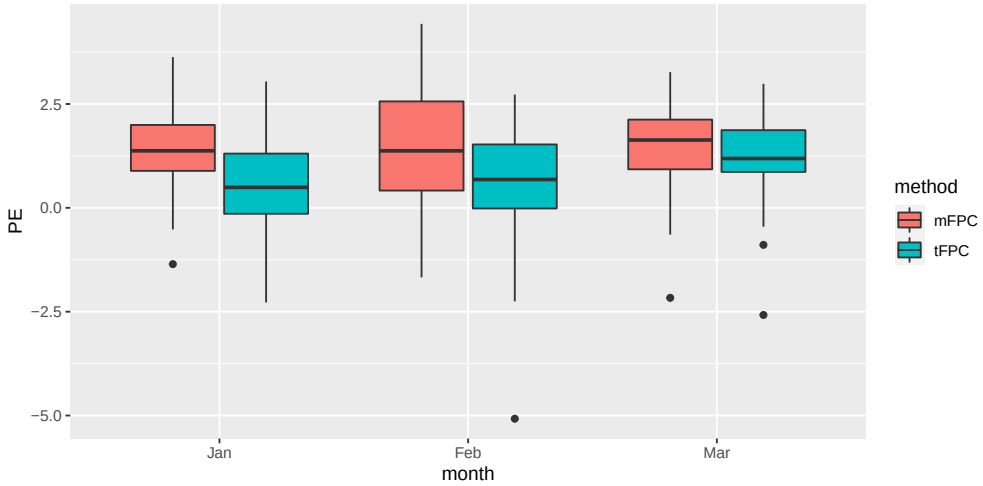

Figure 2.9: The prediction error of two methods for the monthly temperature in January–March, 2014.

3. PRINCIPAL COMPONENT ANALYSIS OF SERIAL CORRELATED TWO DIMENSIONAL FUNCTIONAL DATA WITH A DISTRIBUTION FROM EXPONENTIAL FAMILY

This chapter is organized as follows. In Section 3.1, we propose a principal component model for serial correlated 2-dimensional functional data with a distribution from exponential family. In Section 3.2, we present the details of penalized likelihood, while the variational-EM algorithm are emphasized in Section 3.3 to solve the scalability problem. Section 3.4 introduces the implementation details of model selection. Simulation studies and the Arctic sea-ice-extent data analysis are discussed in Sections 3.5 and 3.6, respectively. Technical details can be found in Appendix A.

3.1 Mixed Effects Model for Serial Correlated 2-d Functional Data with A Distribution from Exponential Family

Denote $(x, y) \in \Omega$ as a 2-dimensional index variable in a compact subset $\Omega \subset \mathbb{R}^2$ and $t \in \{1, \ldots, T\}$ as the time variable. Let $Z_t(x, y)$ be the sequential random variables at (x, y) and time point t. Assume that $Z_t(x, y)$ follows a distribution from exponential family,

$$p\big(Z_t(x,y)|\gamma_t(x,y)\big) = h\big(Z_t(x,y)\big) \exp\big\{Z_t(x,y)\gamma_t(x,y) - g\big(\gamma_t(x,y)\big)\big\}, \tag{3.1}$$

where $\gamma_t(x, y), h(\cdot), g(\cdot)$ are the natural-parameter function, normalization function, and cumulant function, respectively. Some commonly-used distributions such as binary distribution with success probability function $p_t(x, y)$ and Poisson distribution with intensity function $\lambda_t(x, y)$ can be written as the form of exponential-family distributions (3.1). The natural-parameters of binary and Poisson are $\gamma_t(x, y) = \log\{p_t(x, y)/(1 - p_t(x, y))\}$ and $\gamma_t(x, y) = \log \lambda_t(x, y)$, respectively.

Suppose that natural-parameter function $\gamma_t(x, y)$ can be decomposed as

$$\gamma_t(x,y) = \mathbb{E}\{\gamma_t(x,y)\} + \sum_{j=1}^{J} \alpha_{j,t}\phi_j(x,y) + \epsilon_t(x,y),$$

where $\mathbb{E}\{\gamma_t(x,y)\}$, $\alpha_{j,t}$, and $\phi_j(x,y)$ are the mean function, functional principal component (FPC) scores with serial correlation, and smooth FPC functions, respectively. $\epsilon_t(x,y)$ is the white noise with mean 0 and variance σ^2. The FPC functions admit orthonormality, i.e.,

$$\int_\Omega \phi_i(x,y)\phi_j(x,y)\mathrm{d}x\mathrm{d}y = \delta_{ij},$$

where δ_{ij} is the Kronecker delta. We further assume that the mean function $\mathbb{E}\{\gamma_t(x,y)\}$ can be separated into the multiplication of a smooth bivariate function of location $\mu_1(x,y)$ and a continuous function of time $\mu_2(t)$. Thus the model can be rewritten as

$$\gamma_t(x,y) = \mu_1(x,y)\mu_2(t) + \sum_{j=1}^{J} \alpha_{j,t}\phi_j(x,y) + \epsilon_t(x,y). \tag{3.2}$$

As stated in (3.2), the natural-parameter function $\gamma_t(x,y)$ consists of three parts, the mean function, the FPCs, and the measurement error. While the part of mean function is treated as the fixed effect, the FPCs are considered as the random effects whose randomness comes from the FPC scores $\{\alpha_{j,t}\}$. We can view model (3.2) as a mixed-effects model of $\gamma_t(x,y)$.

Assume the FPC scores with serial correlation $\{\alpha_{j,t}\}$ in (3.2) follow the p-th order autoregressive (AR(p)) model

$$\alpha_{j,t} = \sum_{\ell=1}^{p} k_\ell \alpha_{j,t-\ell} + \eta_{j,t}, \quad j = 1,\ldots,J, \quad t = 1,\ldots,T, \tag{3.3}$$

where k_ℓ's are the coefficients of AR(p) models, and $\eta_{j,t}$'s follow a normal distribution with mean 0 and variance $\sigma_j^2, j = 1,\ldots,J$, independently. For the identifiability of the FPC scores, we assume that the coefficients k_ℓ's of AR(p) models are identical with respect to any j-th FPC scores and the variances σ_j^2's are monotonically decreasing, i.e., $\sigma_1^2 > \sigma_2^2 > \cdots > \sigma_J^2$.

One special attention in (3.2) is that the magnitude of mean functions can shift between μ_1 and μ_2. It only can be identified up to a scaling constant. For example, $\{c\mu_1(x,y)\} \times \{\mu_2(t)/c\}$ are equivalent to $\mu_t(x,y) = \mu_1(x,y)\mu_2(t)$ for a nonzero constant c. For the identifiability of the mean

functions μ_1 and μ_2, we impose a L_2-norm constraint on $\mu_1(x, y)$ such that

$$\|\mu_1(x, y)\|_2 = 1. \tag{3.4}$$

Since $\mu_1(x, y), \mu_2(t), \phi_j(x, y)$ are functions which are intrinsically infinite-dimensional, it is impossible to obtain the estimation of these functions directly. Instead, we approximate the functions by basis expansions

$$\mu_1(x, y) = \mathbf{b}(x, y)^\mathsf{T}\boldsymbol{\theta}_b,\ \mu_2(t) = \mathbf{c}(t)^\mathsf{T}\boldsymbol{\theta}_c,$$

and

$$\phi_j(x, y) = \mathbf{b}(x, y)^\mathsf{T}\boldsymbol{\theta}_j,\ j = 1, \ldots, J,$$

where $\boldsymbol{\theta}_b \in \mathbb{R}^{n_b}, \boldsymbol{\theta}_c \in \mathbb{R}^{n_c}$ and $\boldsymbol{\theta}_j \in \mathbb{R}^{n_b}, j = 1, \ldots, J$ are the basis coefficients. $\mathbf{b}(x, y) = (b_1(x, y), \ldots, b_{n_b}(x, y))^\mathsf{T}$ are the n_b-dimensional vectors of bivariate basis functions and $\mathbf{c}(t) = (c_1(t), \ldots, c_{n_c}(t))^\mathsf{T}$ are the n_c-dimensional vectors of univariate basis functions. For the identifiability consideration, we assume that $\mathbf{b}(x, y)$ is orthonormal, i.e.,

$$\int_\Omega \mathbf{b}(x, y)\mathbf{b}^\mathsf{T}(x, y)\mathrm{d}x\mathrm{d}y = \mathbf{I}_{n_b}.$$

In the numerical examples discussed in Sections 3.5 and 3.6, we consider using the triangulated Bernstein polynomial functions (Lai and Schumaker, 2007) as the bivariate basis functions $\mathbf{b}(x, y)$ due to its advantage on irregular domains. Further properties of Bernstein polynomial functions can be referred to Zhou and Pan (2014b). Two examples of triangulations are given in Figures 2.1 and 3.8. As for the univariate basis functions $\mathbf{c}(t)$, the commonly-used Fourier basis functions and polynomial basis functions can be applied to capture the seasonality and trends, respectively. With the basis expansions, we rewrite the model in (3.2) as

$$\gamma_t(x, y) = \mathbf{b}(x, y)^\mathsf{T}\boldsymbol{\theta}_b\boldsymbol{\theta}_c^\mathsf{T}\mathbf{c}(t) + \mathbf{b}(x, y)^\mathsf{T}\boldsymbol{\Theta}\boldsymbol{\alpha}_t + \epsilon_t(x, y),$$

where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)$ and $\boldsymbol{\alpha}_t = (\alpha_{1,t}, \ldots, \alpha_{J,t})^\mathsf{T}$. Meanwhile, vectorizing the p-th order auto-regressive model of FPC scores, (3.3) is rewritten as

$$\boldsymbol{\alpha}_t = \sum_{\ell=1}^{p} k_\ell \boldsymbol{\alpha}_{t-\ell} + \boldsymbol{\eta}_t,$$

where $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \mathbf{H}_J)$ with the covariance matrix $\mathbf{H}_J = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_J^2)$. Note that the constraint in (3.4), the orthonormal constraints of the FPC functions and bivariate basis functions $\mathbf{b}(x, y)$ imply that

$$\|\boldsymbol{\theta}_b\|^2 = 1 \text{ and } \boldsymbol{\Theta}^\mathsf{T}\boldsymbol{\Theta} = \mathbf{I}. \tag{3.5}$$

Suppose we have n_t randomly sampled points $(x_{t1}, y_{t1}), \ldots, (x_{tn_t}, y_{tn_t})$ on the surface at time point $t = 1, \ldots, T$. Denote the observed data and the corresponding latent variables at time t as

$$\mathbf{z}_t \equiv (Z_t(x_1, y_1), \ldots, Z_t(x_{n_t}, y_{n_t}))^\mathsf{T}, \qquad \boldsymbol{\gamma}_t \equiv (\gamma_t(x_1, y_1), \ldots, \gamma_t(x_{n_t}, y_{n_t}))^\mathsf{T}.$$

Write $\mathbf{B}_t = (\mathbf{b}(x_{t1}, y_{t1}), \ldots, \mathbf{b}(x_{tn_t}, y_{tn_t}))^\mathsf{T}$ and $\boldsymbol{\epsilon}_t = (\epsilon_t(x_{t1}, y_{t1}), \ldots, \epsilon_t(x_{tn_t}, y_{tn_t}))^\mathsf{T}$ and let $g(\boldsymbol{\gamma}_t) = \big(g(\gamma_t(x_1, y_1)), \ldots, g(\gamma_t(x_{n_t}, y_{n_t}))\big)^\mathsf{T}$. For notational simplicity, denote $\mathbf{c}_t = \mathbf{c}(t)$ and $\mathbf{k} = (k_1, \ldots, k_p)^\mathsf{T}$. The proposed model given in (3.1)–(3.3) can be rewritten as

$$\begin{aligned}
&p(\mathbf{z}_t|\boldsymbol{\gamma}_t) = \exp\{\mathbf{z}_t^\mathsf{T}\boldsymbol{\gamma}_t - \mathbf{1}^\mathsf{T} g(\boldsymbol{\gamma}_t)\} \prod_{i=1}^{n_t} h(z_{i,t}), \\
&\boldsymbol{\gamma}_t = \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^\mathsf{T}\mathbf{c}_t + \mathbf{B}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t, \;\; \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{n_t}), \\
&\boldsymbol{\alpha}_t = \sum_{\ell=1}^{p} k_\ell \boldsymbol{\alpha}_{t-\ell} + \boldsymbol{\eta}_t, \;\; \boldsymbol{\eta}_t \sim N(\mathbf{0}, \mathbf{H}_J),
\end{aligned} \tag{3.6}$$

subject to the constraints in (3.5). We call the proposed model as temporal-dependent exponential-family functional principal component (tEFPC, for short) model.

The unknown parameters to be estimated are $\Xi = \{\boldsymbol{\theta}_b, \boldsymbol{\theta}_c, \boldsymbol{\Theta}, \sigma^2, \mathbf{H}_J, \mathbf{k}\}$.

3.2 Penalized Complete Data Likelihood

Denote $\mathbf{Z} = \{\mathbf{z}_t\}_{t=1}^T$, $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_t\}_{t=1}^T$, and $\boldsymbol{\alpha} = \{\boldsymbol{\alpha}_t\}_{t=1}^T$. The marginal likelihood $L(\Xi; \mathbf{Z}) = \int\int p(\mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\gamma} \mathrm{d}\boldsymbol{\alpha}$ does not have an analytical expression with respect to $\mathbf{Z}$. Hence it is difficult to obtain the estimates of the parameters Ξ directly by maximizing the marginal likelihood. Instead we propose to use an expectation-maximization (EM) algorithm to estimate Ξ, which will be discussed in Section 3.3. In this section, we first introduce the complete data log-likelihood $l_c(\Xi; \mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = \log p(\mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\alpha})$, which can be separated into the following three parts:

$$l_c(\Xi; \mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = \log p(\mathbf{Z}|\boldsymbol{\gamma}) + \log p(\boldsymbol{\gamma}|\boldsymbol{\alpha}) + \log p(\boldsymbol{\alpha}). \tag{3.7}$$

The first part on the right hand side of Equation (3.7) can be derived as

$$\log p(\mathbf{Z}|\boldsymbol{\gamma}) = \sum_{t=1}^{T} \{\mathbf{z}_t^\mathsf{T} \boldsymbol{\gamma}_t - \mathbf{1}^\mathsf{T} g(\boldsymbol{\gamma}_t)\}.$$

While the second part of (3.7) is

$$\log p(\boldsymbol{\gamma}|\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{t=1}^{T} n_t \log \sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T} \{\boldsymbol{\gamma}_t - \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^\mathsf{T}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t\}^\mathsf{T}\{\boldsymbol{\gamma}_t - \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^\mathsf{T}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t\}.$$

The third part is the distribution of the FPC scores $\boldsymbol{\alpha}$, which can be written as

$$\log p(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{j=1}^{J} \left\{T \log \sigma_j^2 - \log|\mathbf{M}_j| + \frac{1}{\sigma_j^2} S_j(\mathbf{k})\right\},$$

Thus the complete data log likelihood can be derived as

$$\begin{aligned} l_c(\Xi; \mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = &\sum_{t=1}^{T} \left(\mathbf{z}_t^\mathsf{T}\boldsymbol{\gamma}_t - \mathbf{1}^\mathsf{T} g(\boldsymbol{\gamma}_t)\right) - \frac{1}{2}\sum_{t=1}^{T} n_t \log \sigma^2 \\ &- \frac{1}{2\sigma^2}\sum_{t=1}^{T} \left(\boldsymbol{\gamma}_t - \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^\mathsf{T}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t\right)^\mathsf{T}\left(\boldsymbol{\gamma}_t - \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^\mathsf{T}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}\boldsymbol{\alpha}_t\right) \end{aligned}$$

$$-\frac{1}{2}\sum_{j=1}^{J}\Big(T\log\sigma_j^2-\log|\mathbf{M}_j|+\frac{1}{\sigma_j^2}S_j(\mathbf{k})\Big).$$

where $\mathbf{M}_j$ is the precision matrix of $(\alpha_{j,t},\ldots,\alpha_{j,t+p-1})^\mathsf{T}$, and $S_j(\mathbf{k})=(1,\mathbf{k}^\mathsf{T})\mathbf{D}_j(1,\mathbf{k}^\mathsf{T})^\mathsf{T}$ is the sum of squares from the j-th component of $\boldsymbol{\alpha}$. The quadratic matrix $\mathbf{D}_j$ is

$$\mathbf{D}_j=\begin{pmatrix} D_{11,j} & -D_{12,j} & -D_{13,j} & \ldots & -D_{1(p+1),j}\\ -D_{12,j} & D_{22,j} & D_{23,j} & \ldots & D_{2(p+1),j}\\ \vdots & \vdots & \vdots & & \vdots\\ -D_{(p+1)1,j} & D_{(p+1)2,j} & D_{(p+1)3,j} & \ldots & D_{(p+1)(p+1),j}\end{pmatrix},$$

with its components $D_{ik,j}=D_{ki,j}=\alpha_{j,i}\alpha_{j,k}+\alpha_{j,i+1}\alpha_{j,k+1}+\cdots+\alpha_{j,n+1-k}\alpha_{j,n+1-i}$. The details can be found in Box et al. (2015) and Chapter 2 of this dissertation.

When implementing tEFPC, the number of basis functions should be moderately large enough to capture the location/time variations of the natural-parameter function. However, a large number of basis functions may result in the overfitting problem. We introduce the penalized likelihood to avoid the overfitting issue. The objective function of penalized likelihood that we try to minimize is the combination of the complete data log-likelihood and the regularization component, i.e.,

$$\mathrm{Obj}(\Xi,\lambda)=-2l_c(\Xi;\mathbf{Z},\boldsymbol{\gamma},\boldsymbol{\alpha})+\mathrm{Penalty}(\lambda),$$

where $\mathrm{Penalty}(\lambda)$ is the roughness penalty used to regularize the flexibility of functions, with tuning parameters λ to be selected. Specifically, we penalize the integrated squared second derivatives of functions. For the univariate function $\mu_2(t)=\mathbf{c}(t)^\mathsf{T}\boldsymbol{\theta}_c$, the penalty is

$$\int_T\left\{\frac{\partial^2\mu_2(t)}{\partial t^2}\right\}^2\mathrm{d}t=\boldsymbol{\theta}_c^\mathsf{T}\int_T\left\{\frac{\partial^2\mathbf{c}(t)}{\partial t^2}\frac{\partial^2\mathbf{c}(t)^\mathsf{T}}{\partial t^2}\right\}\mathrm{d}t\boldsymbol{\theta}_c=\boldsymbol{\theta}_c^\mathsf{T}\mathbf{P}\boldsymbol{\theta}_c,$$

where $\mathbf{P}$ denote the roughness penalty matrix of the univariate basis, i.e.,

$$\mathbf{P} = \int_T \left\{ \frac{\partial^2 \mathbf{c}(t)}{\partial t^2} \frac{\partial^2 \mathbf{c}(t)^\mathsf{T}}{\partial t^2} \right\} \mathrm{d}t.$$

For the bivariate function $f(x, y)$ (i.e., $\mu_1(x, y)$ and $\phi_j(x, y)$), we use the thin-plate penalization (Ruppert et al., 2003) as the roughness penalty, i.e.,

$$\int_\Omega \left[\left\{ \frac{\partial^2 f(x,y)}{\partial x^2} \right\}^2 + 2 \left\{ \frac{\partial^2 f(x,y)}{\partial x \partial y} \right\}^2 + \left\{ \frac{\partial^2 f(x,y)}{\partial y^2} \right\}^2 \right] \mathrm{d}x\mathrm{d}y.$$

Let $\mathbf{\Gamma}$ denote

$$\int_\Omega \left\{ \frac{\partial^2 \mathbf{b}(x,y)}{\partial x^2} \frac{\partial^2 \mathbf{b}(x,y)^\mathsf{T}}{\partial x^2} + 2 \frac{\partial^2 \mathbf{b}(x,y)}{\partial x \partial y} \frac{\partial^2 \mathbf{b}(x,y)^\mathsf{T}}{\partial x \partial y} + \frac{\partial^2 \mathbf{b}(x,y)}{\partial y^2} \frac{\partial^2 \mathbf{b}(x,y)^\mathsf{T}}{\partial y^2} \right\} \mathrm{d}x\mathrm{d}y;$$

then, the penalties for $\mu_1(x, y)$ and $\phi_j(x, y)$ can be written as $\boldsymbol{\theta}_b \mathbf{\Gamma} \boldsymbol{\theta}_b$ and $\boldsymbol{\theta}_j \mathbf{\Gamma} \boldsymbol{\theta}_j$, $j = 1, \ldots, J$, respectively. Hence, the overall roughness penalty is

$$\mathrm{Penalty}(\lambda) = \lambda_{\mu_s} \boldsymbol{\theta}_b^\mathsf{T} \mathbf{\Gamma} \boldsymbol{\theta}_b + \lambda_{\mu_t} \boldsymbol{\theta}_c^\mathsf{T} \mathbf{P} \boldsymbol{\theta}_c + \lambda_{pc} \sum_{j=1}^{J} \boldsymbol{\theta}_j^\mathsf{T} \mathbf{\Gamma} \boldsymbol{\theta}_j,$$

where $\lambda = (\lambda_{\mu_s}, \lambda_{\mu_t}, \lambda_{pc})$ are the regularization parameters. In summary, the penalized complete data log likelihood is

$$\begin{aligned} \mathrm{Obj}(\Xi, \lambda) = & -2 \sum_{t=1}^{T} \{ \mathbf{z}_t^\mathsf{T} \boldsymbol{\gamma}_t - \mathbf{1}^\mathsf{T} g(\boldsymbol{\gamma}_t) \} + \sum_{t=1}^{T} n_t \log \sigma^2 \\ & + \frac{1}{\sigma^2} \sum_{t=1}^{T} \{ \boldsymbol{\gamma}_t - \mathbf{B}_t \boldsymbol{\theta}_b \boldsymbol{\theta}_c^\mathsf{T} \mathbf{c}_t - \mathbf{B}_t \mathbf{\Theta} \boldsymbol{\alpha}_t \}^\mathsf{T} \{ \boldsymbol{\gamma}_t - \mathbf{B}_t \boldsymbol{\theta}_b \boldsymbol{\theta}_c^\mathsf{T} \mathbf{c}_t - \mathbf{B}_t \mathbf{\Theta} \boldsymbol{\alpha}_t \} \\ & + \sum_{j=1}^{J} \left\{ T \log \sigma_j^2 - \log |\mathbf{M}_j| + \frac{1}{\sigma_j^2} S_j(\mathbf{k}) \right\} \\ & + \lambda_{\mu_s} \boldsymbol{\theta}_b^\mathsf{T} \mathbf{\Gamma} \boldsymbol{\theta}_b + \lambda_{\mu_t} \boldsymbol{\theta}_c^\mathsf{T} \mathbf{P} \boldsymbol{\theta}_c + \lambda_{pc} \sum_{j=1}^{J} \boldsymbol{\theta}_j^\mathsf{T} \mathbf{\Gamma} \boldsymbol{\theta}_j. \end{aligned} \tag{3.8}$$

With the penalized complete data log-likelihood, we therefore utilize the EM algorithm to iteratively update the locally optimal estimated values of parameters. The detailed EM algorithm will be introduced in Section 3.3.

3.3 The EM algorithm

3.3.1 The E-Step

In the E-step, denote $\Xi^{(0)}$ as the values of parameters derived from the previous iteration of the EM algorithm. At the current step, let $Q(\Xi; \Xi^{(0)})$ be the conditional expectation of (3.8), i.e.,

$$Q(\Xi|\Xi^{(0)}) = \mathbb{E}[-2\log p(\mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) + \text{Penalty}(\lambda)|\mathbf{Z}, \Xi^{(0)}].$$

Denote $\widehat{\boldsymbol{\gamma}}_t = \mathbb{E}[\boldsymbol{\gamma}_t|\mathbf{Z}, \Xi^{(0)}]$, $\widehat{\boldsymbol{\Gamma}}_t = \text{var}[\boldsymbol{\gamma}_t|\mathbf{Z}, \Xi^{(0)}]$, $\widehat{\boldsymbol{\alpha}}_t = \mathbb{E}[\boldsymbol{\alpha}_t|\mathbf{Z}, \Xi^{(0)}]$, $\widehat{\boldsymbol{\Sigma}}_t = \text{var}[\boldsymbol{\alpha}_t|\mathbf{Z}, \Xi^{(0)}]$, $\widehat{\boldsymbol{\Lambda}}_t = \text{cov}[\boldsymbol{\gamma}_t, \boldsymbol{\alpha}_t|\mathbf{Z}, \Xi^{(0)}]$. Write $\widehat{S}_j(\mathbf{k}) = \mathbb{E}[S_j(\mathbf{k})|\mathbf{Z}, \Xi^{(0)}] = (1, \mathbf{k}^\mathsf{T})\widehat{\mathbf{D}}_j(1, \mathbf{k}^\mathsf{T})^\mathsf{T}$, where $\widehat{\mathbf{D}}_j = \mathbb{E}[\mathbf{D}_j|\mathbf{Z}, \Xi^{(0)}]$. Using the derivations (3.8) in the last section, it shows that

$$\begin{aligned}
Q(\Xi; \Xi^{(0)}) = &-2\sum_{t=1}^{T}\{\mathbf{z}_t^\mathsf{T}\widehat{\boldsymbol{\gamma}}_t - \mathbf{1}^\mathsf{T}\mathbb{E}[g(\boldsymbol{\gamma}_t)|\mathbf{Z}, \Xi^{(0)}]\} + \sum_{t=1}^{T} n_t \log\sigma^2 \\
&+ \frac{1}{\sigma^2}\sum_{t=1}^{T}\big\{(\widehat{\boldsymbol{\gamma}}_t - \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^\mathsf{T}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}\widehat{\boldsymbol{\alpha}}_t)^\mathsf{T}(\widehat{\boldsymbol{\gamma}}_t - \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^\mathsf{T}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}\widehat{\boldsymbol{\alpha}}_t) \\
&\qquad + \text{tr}(\widehat{\boldsymbol{\Gamma}}_t) + \text{tr}(\mathbf{B}_t\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_t\boldsymbol{\Theta}^\mathsf{T}\mathbf{B}_t^\mathsf{T}) - 2\text{tr}(\mathbf{B}_t\boldsymbol{\Theta}\widehat{\boldsymbol{\Lambda}}_t^\mathsf{T})\big\} \\
&+ \sum_{j=1}^{J}\big\{T\log\sigma_j^2 - \log|\mathbf{M}_j| + \frac{1}{\sigma_j^2}\widehat{S}_j(\mathbf{k})\big\} \\
&+ \lambda_{\mu_s}\boldsymbol{\theta}_b^\mathsf{T}\boldsymbol{\Gamma}\boldsymbol{\theta}_b + \lambda_{\mu_t}\boldsymbol{\theta}_c^\mathsf{T}\mathbf{P}\boldsymbol{\theta}_c + \lambda_{pc}\sum_{j=1}^{J}\boldsymbol{\theta}_j^\mathsf{T}\boldsymbol{\Gamma}\boldsymbol{\theta}_j, \qquad (3.9)
\end{aligned}$$

To obtain the conditional expected values and variances above, it is crucial to derive the conditional distribution $p(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)})$, which is

$$p(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)}) = \frac{p(\mathbf{Z}|\boldsymbol{\gamma}, \Xi^{(0)})p(\boldsymbol{\gamma}|\boldsymbol{\alpha}, \Xi^{(0)})p(\boldsymbol{\alpha}|\Xi^{(0)})}{\int p(\mathbf{Z}|\boldsymbol{\gamma}, \Xi^{(0)})p(\boldsymbol{\gamma}|\boldsymbol{\alpha}, \Xi^{(0)})p(\boldsymbol{\alpha}|\Xi^{(0)})\mathrm{d}\boldsymbol{\alpha}\mathrm{d}\boldsymbol{\gamma}}.$$

However, it is impossible to analytically derive the denominator integral

$$\int p(\mathbf{Z}|\boldsymbol{\gamma}, \Xi^{(0)}) p(\boldsymbol{\gamma}|\boldsymbol{\alpha}, \Xi^{(0)}) p(\boldsymbol{\alpha}|\Xi^{(0)}) \mathrm{d}\boldsymbol{\alpha} \mathrm{d}\boldsymbol{\gamma},$$

due to the high-dimensional $\mathbf{Z}$. It requires us to find other convenient ways to approximate the conditional distribution $p(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)})$. In the following sections, we discuss two approaches that can be applied to achieve the target.

3.3.1.1 Laplace approximation - Kalman filter and smoother approach

In this section, we discuss the Laplace approximation–Kalman Filter and Smoother method (shortly, LapKFS). Laplace approximation (Zhang and Cressie, 2019; Durbin and Koopman, 2012) is a method to approximate the conditional distribution $p(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)})$ by a Gaussian distribution $\widetilde{p}(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)})$, i.e.,

$$p(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)}) \approx \widetilde{p}(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)}),$$

where the mean of the Gaussian distribution $\widetilde{p}(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)})$ is the mode of complete data log likelihood $\log p(\mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\alpha}; \Xi^{(0)})$ with respect to $(\boldsymbol{\gamma}, \boldsymbol{\alpha})$, and the variance is approximated by the Hessian matrix of $\log p(\mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\alpha}; \Xi^{(0)})$.

Following the suggestion from Chapter 10 of Durbin and Koopman (2012), we further incorporate Laplace approximation into the intrinsic state-space structure in (3.6). We first vectorize the unobserved variables $\boldsymbol{\gamma}_t$ and $\boldsymbol{\beta}_t = (\boldsymbol{\alpha}_{t+p}^{\mathsf{T}}, \ldots, \boldsymbol{\alpha}_t^{\mathsf{T}})^{\mathsf{T}}$ as $\boldsymbol{\xi}_t$. After rewriting the model (3.6) with respect to $\boldsymbol{\xi}_t$, we replace the exponential-family distribution $p(\mathbf{z}_t|\boldsymbol{\xi}_t, \Xi^{(0)})$ by its Gaussian approximation. The Kalman filter and smoother can be applied to obtain the approximating conditional distribution $\widetilde{p}(\boldsymbol{\xi}_t|\mathbf{Z}, \Xi^{(0)}), t = 1, \ldots, T$. Finally, the approximating conditional distribution $\widetilde{p}(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)})$ is derived. The details of LapKFS approach can be found in Section A.1 of Appendix.

3.3.1.2 Variational inference approach

The LapKFS approach, however, is still computationally expensive when $\boldsymbol{\gamma}_t$ and $\boldsymbol{\alpha}_t$ are high-dimensional. Specifically, in the procedures of LapKFS, the latent variables $\boldsymbol{\gamma}_t$ and $\boldsymbol{\alpha}_t$ are integrated as a $(n_t+(p+1)J)$-dimensional vector $\boldsymbol{\xi}_t$. The inversion of the covariance matrix associated with $\boldsymbol{\xi}_t$ leads to the $\mathcal{O}((n_t+(p+1)J)^3)$ computational costs, which is not scalable when n_t is very large in the application of sea-ice-extent data analysis.

To overcome the scalability issue, we propose a variational inference approach to obtain the approximating conditional distribution $\widetilde{p}(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)})$. Variational inference (Palmer et al., 2006; Chiquet et al., 2018; Blei et al., 2017) assumes variables come from a given variational distribution with unknown parameters, for example, the Gaussian distribution with the mean and variance parameters to be determined. It approximates the conditional distribution $p(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)})$ by minimizing the Kullback-Leibler (KL) divergence between this conditional distribution and the given variational distribution. By assuming specific structure (e.g., independence) of the given variational distribution, variational inference can avoid the problem of the large-scale matrix inversion. Thus it greatly reduce the computational costs. In the rest of this section, we introduce the mean-field variational family (Blei et al., 2017) and apply it to approximate $p(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)})$.

We first rewrite the likelihood as the scaled version with respect to $\{\gamma_{ti}\}_{t,i}$ and $\{\alpha_{j,t}\}_{j,t}$, i.e.,

$$
\begin{aligned}
\log p(\mathbf{Z}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) \equiv & \sum_{t=1}^{T}\sum_{i=1}^{n_t}\{z_{ti}\gamma_{ti} - g(\gamma_{ti})\} \\
& - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\sum_{i=1}^{n_t}(\gamma_{ti} - \mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t - \sum_{j=1}^{J}\mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_j\alpha_{j,t})^2 - \frac{1}{2}\sum_{t=1}^{T} n_t \log\sigma^2 \\
& \quad - \frac{T-p}{2}\sum_{j=1}^{J}\log\sigma_j^2 - \sum_{t=p+1}^{T}\sum_{j=1}^{J}\frac{1}{2\sigma_j^2}(\alpha_{j,t} - \sum_{\ell=1}^{p}k_\ell\alpha_{j,t-\ell})^2 + \sum_{j=1}^{J}\log p(\alpha_{j,1}, \ldots, \alpha_{j,p}),
\end{aligned}
$$

where $z_{ti} = z_t(x_i, y_i)$, $\gamma_{ti} = \gamma_t(x_i, y_i)$, and $\mathbf{b}_{ti} = \mathbf{b}_t(x_i, y_i)$. Since $\sum_{j=1}^{J}\log p(\alpha_{j,1}, \ldots, \alpha_{j,p})$ is invariant with respect to T, its likelihood contribution can be omitted when $T \gg p$. Letting $q(\boldsymbol{\gamma}, \boldsymbol{\alpha})$ be the variational distribution, we aim to find the optimal distribution that minimizes the KL diver-

gence between the conditional distribution $p(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z})$ and the variational distribution $q(\boldsymbol{\gamma}, \boldsymbol{\alpha})$. We further assume the variational distribution $q(\boldsymbol{\gamma}, \boldsymbol{\alpha})$ admits the mean-field assumption such that

$$p(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}) \approx q(\boldsymbol{\gamma}, \boldsymbol{\alpha}) \triangleq \prod_{t=1}^{T} \prod_{i=1}^{n_t} q(\gamma_{ti}) \prod_{\tau=1}^{T} \prod_{j=1}^{J} q(\alpha_{j,\tau}),$$

where the components γ_{ti} and $\alpha_{j,\tau}$ follow the Gaussian distributions with specific unknown mean and variance parameters, i.e.,

$$q(\gamma_{ti}) = N(\gamma_{ti}; \mu_{ti}, \phi_{ti}^2),\ i = 1, \ldots, n_t; t = 1, \ldots, T$$
$$q(\alpha_{j,\tau}) = N(\alpha_{j,\tau}; \nu_{\tau j}, \varphi_{\tau j}^2),\ j = 1, \ldots, J; \tau = 1, \ldots, T.$$

The variational parameters $\{\mu_{ti}, \phi_{ti}^2, \nu_{\tau j}, \varphi_{\tau j}^2\}$ can be learnt by minimizing the KL divergence between $p(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z})$ and $q(\boldsymbol{\gamma}, \boldsymbol{\alpha})$,

$$\{\widehat{\mu}_{ti}, \widehat{\phi}_{ti}^2, \widehat{\nu}_{\tau j}, \widehat{\varphi}_{\tau j}^2\} = \underset{\mu_{ti}, \phi_{ti}^2, \nu_{\tau j}, \varphi_{\tau j}^2}{\operatorname{argmin}} \mathrm{KL}\big\{p(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}) \| q(\boldsymbol{\gamma}, \boldsymbol{\alpha})\big\},$$

which is equivalent to maximizing the Evidence Lower Bound (ELBO, Blei et al., 2017), i.e.,

$$\begin{aligned}\{\widehat{\mu}_{ti}, \widehat{\phi}_{ti}^2, \widehat{\nu}_{\tau j}, \widehat{\varphi}_{\tau j}^2\} &= \underset{\mu_{ti}, \phi_{ti}^2, \nu_{\tau j}, \varphi_{\tau j}^2}{\operatorname{argmax}} \mathrm{ELBO}(q) \\ &= \underset{\mu_{ti}, \phi_{ti}^2, \nu_{\tau j}, \varphi_{\tau j}^2}{\operatorname{argmax}} \big\{\mathbb{E}[\log p(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \mathbf{Z})] - \mathbb{E}[\log q(\boldsymbol{\gamma}, \boldsymbol{\alpha})]\big\}.\end{aligned}$$

The maximum of ELBO can be obtained via coordinate ascent algorithm (Blei et al., 2017). Assuming we have the variational parameters $\{\mu_{ti}^{(0)}, \phi_{ti}^{2(0)}, \nu_{\tau j}^{(0)}, \varphi_{\tau j}^{2(0)}\}$ obtained from the previous step of coordinate ascent algorithm, the updating formulas for the parameters $\{\mu_{ti}, \phi_{ti}^2, \nu_{\tau j}, \varphi_{\tau j}^2\}$ are presented as below.

In the following part of this section, we introduce the Coordinate Ascent Variational Inference (CAVI, Blei et al., 2017) method to obtain the optimal estimate of the unknown variational parameters $\{\mu_{ti}, \phi_{ti}^2, \nu_{\tau j}, \varphi_{\tau j}^2\}$. Assuming we have the updating results in the previous step

$\{\mu_{ti}^{(0)}, \phi_{ti}^{2(0)}, \nu_{\tau j}^{(0)}, \varphi_{\tau j}^{2(0)}\}$, we then update the parameters of the current step sequentially. Denote $\log p(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \mathbf{Z})$ the complete data log likelihood, we will derive the updating formulas with respect to $\{\gamma_{ti}\}$ and $\{\alpha_{j,\tau}\}$ separately.

Firstly, we will show that the distribution $p(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \mathbf{Z})$ can be approximated by a Gaussian distribution $\widetilde{p}(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \mathbf{Z})$, which is conjugate with

$$q(\boldsymbol{\gamma}, \boldsymbol{\alpha}) = \prod_{t=1}^{T}\prod_{i=1}^{n_t} q(\gamma_{ti}) \prod_{\tau=1}^{T}\prod_{j=1}^{J} q(\alpha_{j,\tau}).$$

Thus the updating formulas of the variational parameters are proportional to the following parts

$$q(\gamma_{ti}) \propto \exp\{\mathbb{E}_{q_{\boldsymbol{\gamma}_{-ti},\boldsymbol{\alpha}}}[\log \widetilde{p}(\gamma_{ti}, \boldsymbol{\gamma}_{-ti}, \boldsymbol{\alpha}, \mathbf{Z})]\}, \tag{3.10}$$

$$q(\alpha_{j,\tau}) \propto \exp\{\mathbb{E}_{q_{\boldsymbol{\alpha}_{-\{j,\tau\}},\boldsymbol{\gamma}}}[\log \widetilde{p}(\alpha_{j,\tau}, \boldsymbol{\alpha}_{-\{j,\tau\}}, \boldsymbol{\gamma}, \mathbf{Z})]\}, \tag{3.11}$$

where $q_{\boldsymbol{\gamma}_{-ti},\boldsymbol{\alpha}} = q_{\boldsymbol{\gamma},\boldsymbol{\alpha}}/q(\gamma_{ti})$, and $q_{\boldsymbol{\alpha}_{-\{j,\tau\}},\boldsymbol{\gamma}} = q_{\boldsymbol{\gamma},\boldsymbol{\alpha}}/q(\alpha_{j,\tau})$, respectively.

Updating formula for variational parameters μ_{ti}, ϕ_{ti}^2 For the updating formula of γ_{ti} part, we first derive the log likelihood partially corresponding to γ_{ti}, i.e.,

$$\log p(\gamma_{ti}, \boldsymbol{\gamma}_{-ti}, \boldsymbol{\alpha}, \mathbf{Z}) \equiv z_{ti}\gamma_{ti} - g(\gamma_{ti}) - \frac{1}{2\sigma^2}(\gamma_{ti} - \mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t - \sum_{j=1}^{J}\mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_j\alpha_{tj})^2. \tag{3.12}$$

Since $g(\gamma_{ti})$ is not linear or quadratic terms of γ_{ti}, we approximate it by the Taylor approximation of the second order around the previous mean estimation $\mathbb{E}[\gamma_{ti}]$,

$$g(\gamma_{ti}) \approx g(\mathbb{E}[\gamma_{ti}]) + g'(\mathbb{E}[\gamma_{ti}])(\gamma_{ti} - \mathbb{E}[\gamma_{ti}]) + \frac{g''(\mathbb{E}[\gamma_{ti}])}{2}(\gamma_{ti} - \mathbb{E}[\gamma_{ti}])^2.$$

By utilizing $\mathbb{E}[\gamma_{ti}] = \mu_{ti}^{(0)}$, we can rewrite the formula (3.12) above as

$$
\begin{aligned}
&\log p(\gamma_{ti}, \boldsymbol{\gamma}_{-ti}, \boldsymbol{\alpha}, \mathbf{Z}) \\
\equiv\ & z_{ti}\gamma_{ti} - \Big\{ g(\mu_{ti}^{(0)}) + g'(\mu_{ti}^{(0)})(\gamma_{ti} - \mu_{ti}^{(0)}) + \frac{g''(\mu_{ti}^{(0)})}{2}(\gamma_{ti} - \mu_{ti}^{(0)})^2 \Big\} \\
& - \frac{1}{2\sigma^2}\big\{\gamma_{ti}^2 - 2(\mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t + \sum_{j=1}^{J}\mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_j\alpha_{tj})\gamma_{ti}\big\} \\
\equiv\ & -\frac{1}{2}\big\{\frac{1}{\sigma^2} + g''(\mu_{ti}^{(0)})\big\}\gamma_{ti}^2 + \big\{\frac{1}{\sigma^2}(\mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t + \sum_{j=1}^{J}\mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_j\alpha_{tj}) \\
& + g''(\mu_{ti}^{(0)})[\mu_{ti}^{(0)} + g''(\mu_{ti}^{(0)})^{-1}(z_{ti} - g'(\mu_{ti}^{(0)}))]\big\}\gamma_{ti} \\
\triangleq\ & \log \widetilde{p}(\gamma_{ti}, \boldsymbol{\gamma}_{-ti}, \boldsymbol{\alpha}, \mathbf{Z}),
\end{aligned}
$$

where $\widetilde{p}(\gamma_{ti}, \boldsymbol{\gamma}_{-ti}, \boldsymbol{\alpha}, \mathbf{Z})$ is the approximating Gaussian distribution with respect to γ_{ti}. The variance of $\widetilde{p}(\gamma_{ti}, \boldsymbol{\gamma}_{-ti}, \boldsymbol{\alpha}, \mathbf{Z})$ is

$$
1/\big\{\frac{1}{\sigma^2} + g''(\mu_{ti}^{(0)})\big\},
$$

and the mean is

$$
\big\{\frac{1}{\sigma^2}(\mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t + \sum_{j=1}^{J}\mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_j\alpha_{tj}) + g''(\mu_{ti}^{(0)})[\mu_{ti}^{(0)} + g''(\mu_{ti}^{(0)})^{-1}(z_{ti} - g'(\mu_{ti}^{(0)}))]\big\}/\big\{\frac{1}{\sigma^2} + g''(\mu_{ti}^{(0)})\big\}.
$$

Since the approximating distribution $\widetilde{p}(\gamma_{ti}, \boldsymbol{\gamma}_{-ti}, \boldsymbol{\alpha}, \mathbf{Z})$ is conjugate to the variational distribution $q(\gamma_{ti})$, the variational distribution of γ_{ti} is proportional to

$$
q(\gamma_{ti}) \propto \exp\{\mathbb{E}_{q_{\boldsymbol{\gamma}_{-ti}}, \boldsymbol{\alpha}}[\log \widetilde{p}(\gamma_{ti}, \boldsymbol{\gamma}_{-ti}, \boldsymbol{\alpha}, \mathbf{Z})]\}.
$$

Notice that $\mathbb{E}[\alpha_{ti}] = \nu_{ti}^{(0)}$, we then have

$$
\begin{aligned}
&\mathbb{E}_{q_{\boldsymbol{\gamma}_{-ti}}, \boldsymbol{\alpha}}[\log \widetilde{p}(\gamma_{ti}, \boldsymbol{\gamma}_{-ti}, \boldsymbol{\alpha}, \mathbf{Z})] \\
=\ & -\frac{1}{2}\big\{\frac{1}{\sigma^2} + g''(\mu_{ti}^{(0)})\big\}\gamma_{ti}^2 + \big\{\frac{1}{\sigma^2}(\mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t + \sum_{j=1}^{J}\mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_j\nu_{tj}^{(0)}) \\
& + g''(\mu_{ti}^{(0)})[\mu_{ti}^{(0)} + g''(\mu_{ti}^{(0)})^{-1}(z_{ti} - g'(\mu_{ti}^{(0)}))]\big\}\gamma_{ti}
\end{aligned}
$$

$$\equiv -\frac{1}{2\phi_{ti}^{2(1)}}\left\{\gamma_{ti}-\mu_{ti}^{(1)}\right\}^2$$

where,

$$\phi_{ti}^{2(1)}=\left\{\frac{1}{\sigma^2}+g''(\mu_{ti}^{(0)})\right\}^{-1},$$

$$\mu_{ti}^{(1)}=\phi_{ti}^{2(1)}\left[\frac{1}{\sigma^2}(\mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t+\sum_{j=1}^{J}\mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_j\nu_{tj}^{(0)})+g''(\mu_{ti}^{(0)})\{\mu_{ti}^{(0)}+g''(\mu_{ti}^{(0)})^{-1}(z_{ti}-g'(\mu_{ti}^{(0)}))\}\right].$$

We have

$$q(\gamma_{ti})\propto\exp\{\mathbb{E}_{q_{\boldsymbol{\gamma}_{-ti},\boldsymbol{\alpha}}}[\log\widetilde{p}(\gamma_{ti},\boldsymbol{\gamma}_{-ti},\boldsymbol{\alpha},\mathbf{Z})]\}=\exp\{-\frac{1}{2\phi_{ti}^{(1)}}(\gamma_{ti}-\mu_{ti}^{(1)})^2\},$$

and it follows a Gaussian distribution. Thus we obtain the updating formulas for ϕ_{ti}^2 and μ_{ti} as $\phi_{ti}^{2(1)}$ and $\mu_{ti}^{(1)}$.

Updating formula for variational parameters $\nu_{\tau j},\varphi_{\tau j}^2$ Similarly, we can obtain the updating formulas for variational parameters $\nu_{\tau j}$ and $\varphi_{\tau j}^2$. We first derive the likelihood partially corresponding to $\alpha_{j,\tau}$. Here we separately consider the following two scenarios, i) $\tau\geq p+1$; ii) $1\leq\tau\leq p$. For scenario i) $\tau\geq p+1$, we assume $\alpha_{j,T+1},\alpha_{j,T+2},\ldots,\alpha_{j,T+p}=0$. The log likelihood is

$$\begin{aligned}
&\log p(\alpha_{j,\tau},\boldsymbol{\alpha}_{-\{j,\tau\}},\boldsymbol{\gamma},\mathbf{Z})\\
&\equiv -\frac{1}{2\sigma^2}\sum_{i=1}^{n_\tau}(\gamma_{\tau i}-\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_\tau-\sum_{l\neq j}^{J}\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_l\alpha_{l,\tau}-\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_j\alpha_{j,\tau})^2 &(*)\\
&\quad-\frac{1}{2\sigma_j^2}(\alpha_{j,\tau}-\sum_{\ell=1}^{p}k_\ell\alpha_{j,\tau-\ell})^2 &(**)\\
&\quad-\frac{1}{2\sigma_j^2}(\alpha_{j,\tau+1}-k_1\alpha_{j,\tau}-k_2\alpha_{j,\tau-1}-\cdots-k_p\alpha_{j,\tau-p+1})^2 &(3.13)\\
&\quad-\frac{1}{2\sigma_j^2}(\alpha_{j,\tau+2}-k_1\alpha_{j,\tau+1}-k_2\alpha_{j,\tau}-\cdots-k_p\alpha_{j,\tau-p+2})^2\\
&\quad\cdots\\
&\quad-\frac{1}{2\sigma_j^2}(\alpha_{j,\tau+p}-k_1\alpha_{j,\tau+p-1}-k_2\alpha_{j,\tau+p-2}-\cdots-k_p\alpha_{j,\tau})^2 &(***)
\end{aligned}$$

$$\triangleq \log \widetilde{p}(\alpha_{j,\tau}, \boldsymbol{\alpha}_{-\{j,\tau\}}, \boldsymbol{\gamma}, \mathbf{Z}),$$

where $\widetilde{p}(\alpha_{j,\tau}, \boldsymbol{\alpha}_{-\{j,\tau\}}, \boldsymbol{\gamma}, \mathbf{Z})$ is the Gaussian distribution with respect to $\alpha_{j,\tau}$. Since the distribution $\widetilde{p}(\alpha_{j,\tau}, \boldsymbol{\alpha}_{-\{j,\tau\}}, \boldsymbol{\gamma}, \mathbf{Z})$ is conjugate to the variational distribution $q(\alpha_{j,\tau})$, the variational distribution of $\alpha_{j,\tau}$ is proportional to

$$q(\alpha_{j,\tau}) \propto \exp\{\mathbb{E}_{q_{\boldsymbol{\alpha}_{-\{j,\tau\}},\boldsymbol{\gamma}}}[\log \widetilde{p}(\alpha_{j,\tau}, \boldsymbol{\alpha}_{-\{j,\tau\}}, \boldsymbol{\gamma}, \mathbf{Z})]\}.$$

Note that part $(*)$ of (3.13) is proportional to

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n_\tau}\big\{(\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_j)^2\alpha_{\tau j}^2 - 2\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_j(\gamma_{\tau i} - \mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_\tau - \sum_{l\neq j}^{J}\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_l\alpha_{\tau l})\alpha_{\tau j}\big\},$$

while part $(**)$ is proportional to

$$-\frac{1}{2\sigma_j^2}(\alpha_{\tau j}^2 - 2\sum_{\ell=1}^{p}k_\ell\alpha_{\tau-\ell,j}\alpha_{\tau j}).$$

As for part $(***)$:

$$\begin{aligned}
&-\frac{1}{2\sigma_j^2}[k_1^2\alpha_{\tau,j}^2 - 2k_1(\alpha_{\tau+1,j} - k_2\alpha_{\tau-1,j} - \cdots - k_p\alpha_{\tau-p+1,j})\alpha_{\tau,j}]\\
&-\frac{1}{2\sigma_j^2}[k_2^2\alpha_{\tau,j}^2 - 2k_2(\alpha_{\tau+2,j} - k_1\alpha_{\tau+1,j} - k_3\alpha_{\tau-1,j} - \cdots - k_p\alpha_{\tau-p+2,j})\alpha_{\tau,j}]\\
&-\ldots\\
&-\frac{1}{2\sigma_j^2}[k_p^2\alpha_{\tau,j}^2 - 2k_p(\alpha_{\tau+p,j} - k_1\alpha_{\tau+p-1,j} - k_2\alpha_{\tau+p-2,j} - \cdots - k_{p-1}\alpha_{\tau+1,j})\alpha_{\tau,j}]\\
=&-\frac{1}{2\sigma_j^2}\{\sum_{\ell=1}^{p}k_\ell^2\alpha_{\tau,j}^2 - 2[\sum_{\ell=1}^{p}k_\ell\alpha_{\tau+\ell,j} - \sum_{\ell=1}^{p}\sum_{i\neq\ell}^{p}k_\ell k_i\alpha_{\tau-i+\ell,j}]\alpha_{\tau,j}\}.
\end{aligned}$$

Therefore, take the summation of the parts $(*)$, $(**)$ and $(***)$ together, and take the expectation

$\mathbb{E}_{q_{\boldsymbol{\alpha}_{-\tau j},\boldsymbol{\gamma}}}[\log \widetilde{p}(\alpha_{\tau j}, \boldsymbol{\alpha}_{-\tau j}, \boldsymbol{\gamma}, \mathbf{Z})]$, we have

$$
\begin{aligned}
&-\frac{1}{2}\{\frac{1}{\sigma^2}\sum_{i=1}^{n_\tau}(\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_j)^2+\frac{1}{\sigma_j^2}(1+\sum_{\ell=1}^{p}k_\ell^2)\}\alpha_{j,\tau}^2 \quad (\Delta)\\
&+\Bigg\{\frac{1}{\sigma^2}\sum_{i=1}^{n_\tau}\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_j(\mu_{\tau i}^{(0)}-\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_\tau-\sum_{l\neq j}^{J}\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_l\nu_{\tau l}^{(0)})\\
&\quad+\frac{1}{\sigma_j^2}(\sum_{\ell=1}^{p}k_\ell\nu_{\tau-i,j}^{(0)}+\sum_{\ell=1}^{p}k_\ell\nu_{\tau+i,j}^{(0)}-\sum_{\ell=1}^{p}\sum_{i\neq\ell}^{p}k_\ell k_i\nu_{\tau-i+\ell,j}^{(0)})\Bigg\}\alpha_{j,\tau} \quad (\Delta\Delta)
\end{aligned}
$$

So the updating formula for scenario i) is

$$
\varphi_{\tau,j}^{2(1)}=\{\frac{1}{\sigma^2}\sum_{i=1}^{n_\tau}(\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_j)^2+\frac{1}{\sigma_j^2}(1+\sum_{\ell=1}^{p}k_\ell^2)\}^{-1}
$$

$$
\begin{aligned}
\nu_{\tau,j}^{(1)}=\varphi_{\tau,j}^{2(1)}\Bigg\{&\frac{1}{\sigma^2}\sum_{i=1}^{n_\tau}\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_j(\mu_{\tau i}^{(0)}-\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_\tau-\sum_{l\neq j}^{J}\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_l\nu_{\tau l}^{(0)})\\
&+\frac{1}{\sigma_j^2}(\sum_{\ell=1}^{p}k_\ell\nu_{\tau-\ell,j}^{(0)}+\sum_{\ell=1}^{p}k_\ell\nu_{\tau+\ell,j}^{(0)}-\sum_{i=1}^{p}\sum_{l\neq i}^{p}k_i k_l\nu_{\tau-l+i,j}^{(0)})\Bigg\}.
\end{aligned}
$$

where $\nu_{T+1,j}^{(0)},\nu_{T+2,j}^{(0)},\ldots,\nu_{T+p,j}^{(0)}=0$.

For scenario ii) $1\leq\tau\leq p$, we will not have part $(**)$ in (3.13). Denote $\nu_{0,j}^{(0)},\nu_{-1,j}^{(0)},\ldots,\nu_{-p,j}^{(0)}=0$, then we have the similar results, i.e.,

$$
\varphi_{\tau,j}^{2(1)}=\{\frac{1}{\sigma^2}\sum_{i=1}^{n_\tau}(\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_j)^2+\frac{1}{\sigma_j^2}\sum_{\ell=1}^{p}k_\ell^2\}^{-1}
$$

$$
\begin{aligned}
\nu_{\tau,j}^{(1)}=\varphi_{\tau,j}^{2(1)}\Bigg\{&\frac{1}{\sigma^2}\sum_{i=1}^{n_\tau}\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_j(\mu_{\tau i}^{(0)}-\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_\tau-\sum_{l\neq j}^{J}\mathbf{b}_{\tau i}^{\mathsf{T}}\boldsymbol{\theta}_l\nu_{\tau l}^{(0)})\\
&+\frac{1}{\sigma_j^2}(\sum_{\ell=1}^{p}k_\ell\nu_{\tau+\ell,j}^{(0)}-\sum_{i=1}^{p}\sum_{l\neq i}^{p}k_i k_l\nu_{\tau-l+i,j}^{(0)})\Bigg\}.
\end{aligned}
$$

Once we update all the parameters in a loop, we need to calculate the ELBO as the stopping criterion. If $|\mathrm{ELBO}^{(l)}-\mathrm{ELBO}^{(l-1)}|<\xi$, then we stop the iteration of updating procedures in E

step. The empirical ELBO can be obtained by

$$
\begin{aligned}
&\mathrm{ELBO}(q) = \mathbb{E}[\log p(\boldsymbol{\gamma}, \boldsymbol{\alpha}, \mathbf{Z})] - \mathbb{E}[\log q(\boldsymbol{\gamma})] - \mathbb{E}[\log q(\boldsymbol{\alpha})] \\
&\approx \sum_{t=1}^{T}\sum_{i=1}^{n_t} z_{ti}\mu_{ti}^{(l)} - \sum_{t=1}^{T}\sum_{i=1}^{n_t}\big[g(0) + g'(0)(\mu_{ti}^{(l)} - 0) + \frac{g''(0)}{2}(\mu_{ti}^{2(l)} + \phi_{ti}^{2(l)})\big] - \frac{1}{2}\sum_{t=1}^{T} n_t \log\sigma^2 \\
&\quad - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\sum_{i=1}^{n_t}\big\{(\mu_{ti}^{(l)} - \mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t - \sum_{j=1}^{J}\mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_j\nu_{tj}^{(l)})^2 + \phi_{ti}^{2(l)} + \sum_{j=1}^{J}(\mathbf{b}_{ti}^{\mathsf{T}}\boldsymbol{\theta}_j)^2\varphi_{tj}^{2(l)}\big\} \\
&\quad - \frac{T-p}{2}\sum_{j=1}^{J}\sum_{j=1}^{J}\log\sigma_j^2 - \sum_{t=p+1}^{T}\sum_{j=1}^{J}\frac{1}{2\sigma_j^2}\big\{(\nu_{tj}^{(l)} - \sum_{i=1}^{p}k_i\nu_{t-i,j}^{(l)})^2 + \varphi_{tj}^{2(l)} + \sum_{i=1}^{p}k_i^2\varphi_{t-i,j}^{2(l)}\big\}
\end{aligned}
$$

After the convergence of iteratively updating formulas for $\{\mu_{ti}, \phi_{ti}^2, \nu_{\tau j}, \varphi_{\tau j}^2\}$, we can approximate the conditional distribution $p(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)})$ by

$$
q(\boldsymbol{\gamma}, \boldsymbol{\alpha}) = \prod_{t=1}^{T}\prod_{i=1}^{n_t} q(\gamma_{ti}) \prod_{\tau=1}^{T}\prod_{j=1}^{J} q(\alpha_{j,\tau}),
$$

where

$$
\begin{aligned}
q(\gamma_{ti}) &= N(\gamma_{ti}; \widehat{\mu}_{ti}, \widehat{\phi}_{ti}^2),\ i = 1, \ldots, n_t; t = 1, \ldots, T \\
q(\alpha_{j,\tau}) &= N(\alpha_{j,\tau}; \widehat{\nu}_{\tau j}, \widehat{\varphi}_{\tau j}^2),\ j = 1, \ldots, J; \tau = 1, \ldots, T,
\end{aligned}
$$

with the final updated parameters $\{\widehat{\mu}_{ti}, \widehat{\phi}_{ti}^2, \widehat{\nu}_{\tau j}, \widehat{\varphi}_{\tau j}^2\}$.

So we can obtain the required values $\widehat{\boldsymbol{\gamma}}_t \approx \mathbb{E}_{q(\boldsymbol{\gamma},\boldsymbol{\alpha})}[\boldsymbol{\gamma}_t|\mathbf{Z}, \Xi^{(0)}] = \widehat{\boldsymbol{\mu}}_t$, $\widehat{\boldsymbol{\alpha}}_t \approx \mathbb{E}_{q(\boldsymbol{\gamma},\boldsymbol{\alpha})}[\boldsymbol{\alpha}_t|\mathbf{Z}, \Xi^{(0)}] = \widehat{\boldsymbol{\nu}}_t$, $\widehat{\boldsymbol{\Gamma}}_t \approx \mathrm{var}_{q(\boldsymbol{\gamma},\boldsymbol{\alpha})}(\boldsymbol{\gamma}_t|\mathbf{Z}, \Xi^{(0)}) = \mathrm{diag}(\widehat{\phi}_{t1}^2, \ldots, \widehat{\phi}_{tn_t}^2)$, $\widehat{\boldsymbol{\Sigma}}_t \approx \mathrm{var}_{q(\boldsymbol{\gamma},\boldsymbol{\alpha})}(\boldsymbol{\alpha}_t|\mathbf{Z}, \Xi^{(0)}) = \mathrm{diag}(\widehat{\varphi}_{t1}^2, \ldots, \widehat{\varphi}_{tJ}^2)$, and $\widehat{\boldsymbol{\Lambda}}_t \approx \mathrm{cov}_{q(\boldsymbol{\gamma},\boldsymbol{\alpha})}[\boldsymbol{\gamma}_t, \boldsymbol{\alpha}_t|\mathbf{Z}, \Xi^{(0)}] = \mathbf{0}$.

3.3.2 The M-Step

In the M step, given the estimates $\Xi^{(0)} = \{\boldsymbol{\theta}_b^{(0)}, \boldsymbol{\theta}_c^{(0)}, \boldsymbol{\Theta}^{(0)}, \sigma^{2(0)}, \{\sigma_j^{2(0)}\}_{j=1}^{J}, \mathbf{k}^{(0)}\}$ at the previous step, we aim to minimize $Q(\Xi|\Xi^{(0)})$ in (3.9). However, the analytical minimizer is difficult to derive. Instead, we use block-wise optimization to update parameters Ξ. The updating formula for

σ^2 has the analytic form

$$\widehat{\sigma}^2 = \frac{1}{\sum_{t=1}^{T} n_t} \sum_{t=1}^{T} \{(\widehat{\boldsymbol{\gamma}}_t - \mathbf{B}_t \boldsymbol{\theta}_b^{(0)} \boldsymbol{\theta}_c^{(0)\mathsf{T}} \mathbf{c}_t - \mathbf{B}_t \boldsymbol{\Theta}^{(0)} \widehat{\boldsymbol{\alpha}}_t)^{\mathsf{T}} (\widehat{\boldsymbol{\gamma}}_t - \mathbf{B}_t \boldsymbol{\theta}_b^{(0)} \boldsymbol{\theta}_c^{(0)\mathsf{T}} \mathbf{c}_t - \mathbf{B}_t \boldsymbol{\Theta}^{(0)} \widehat{\boldsymbol{\alpha}}_t)$$
$$+ \mathrm{tr}(\widehat{\boldsymbol{\Gamma}}_t) + \mathrm{tr}(\mathbf{B}_t \boldsymbol{\Theta}^{(0)} \widehat{\boldsymbol{\Sigma}}_t \boldsymbol{\Theta}^{(0)\mathsf{T}} \mathbf{B}_t^{\mathsf{T}}) - 2\mathrm{tr}(\mathbf{B}_t \boldsymbol{\Theta}^{(0)} \widehat{\boldsymbol{\Lambda}}_t^{\mathsf{T}})\}.$$

Remark: in the binary case, $\boldsymbol{\gamma}_t$ controls the outcomes $\{0, 1\}$ up to a scaling constant. To avoid the identifiability issue, we follow the sugggestion from Heagerty and Lele (1998) to assume the default value $\sigma^2 = 1$.

The updating formula for σ_j^2 is

$$\widehat{\sigma}_j^2 = \frac{\widehat{S}_j(\mathbf{k}^{(0)})}{n},\ j = 1, \ldots, J.$$

The optimization problem with respect to $\boldsymbol{\theta}_b$ can be simplified as minimizing the following $f(\boldsymbol{\theta}_b)$,

$$f(\boldsymbol{\theta}_b) = (\boldsymbol{\theta}_b - \mathbf{m})^{\mathsf{T}} \mathbf{A} (\boldsymbol{\theta}_b - \mathbf{m}),\ \text{such that } \boldsymbol{\theta}_b^{\mathsf{T}} \boldsymbol{\theta}_b = 1, \tag{3.14}$$

where

$$\mathbf{m} = \left\{ \sum_{t=1}^{T} (\boldsymbol{\theta}_c^{(0)\mathsf{T}} \mathbf{c}_t)^2 \mathbf{B}_t^{\mathsf{T}} \mathbf{B}_t + \widehat{\sigma}^2 \lambda_{\mu_s} \boldsymbol{\Gamma} \right\}^{-1} \sum_{t=1}^{T} (\boldsymbol{\theta}_c^{(0)\mathsf{T}} \mathbf{c}_t) \mathbf{B}_t^{\mathsf{T}} (\widehat{\boldsymbol{\gamma}}_t - \mathbf{B}_t \boldsymbol{\Theta}^{(0)} \widehat{\boldsymbol{\alpha}}_t),$$

$$\mathbf{A} = \sum_{t=1}^{T} (\boldsymbol{\theta}_c^{(0)\mathsf{T}} \mathbf{c}_t)^2 \mathbf{B}_t^{\mathsf{T}} \mathbf{B}_t + \widehat{\sigma}^2 \lambda_{\mu_s} \boldsymbol{\Gamma}.$$

Note that (3.14) is a manifold optimization problem of Rayleigh quotient on the sphere, which can be solved using the gradient descent algorithm on the sphere (Absil et al., 2009). The stepwise solutions for updating $\boldsymbol{\theta}_b$ can be referred to Chapter 2 of this dissertation.

The coefficients of the univariate mean $\boldsymbol{\theta}_c$ can be updated analytically by solving the roots from

the derivative of the objective function

$$\frac{1}{\sigma^2}\sum_{t=1}^{T}(\widehat{\boldsymbol{\gamma}}_t - \mathbf{B}_t\widehat{\boldsymbol{\theta}}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}\widehat{\boldsymbol{\alpha}}_t)^{\mathsf{T}}(\widehat{\boldsymbol{\gamma}}_t - \mathbf{B}_t\widehat{\boldsymbol{\theta}}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t - \mathbf{B}_t\boldsymbol{\Theta}\widehat{\boldsymbol{\alpha}}_t) + \lambda_{\mu_t}\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{P}\boldsymbol{\theta}_c.$$

Thus the updating formula of $\boldsymbol{\theta}_c$ can be written as

$$\widehat{\boldsymbol{\theta}}_c = \Big\{\sum_{t=1}^{T}(\mathbf{B}_t\widehat{\boldsymbol{\theta}}_b\mathbf{c}_t^{\mathsf{T}})^{\mathsf{T}}(\mathbf{B}_t\widehat{\boldsymbol{\theta}}_b\mathbf{c}_t^{\mathsf{T}}) + \widehat{\sigma}^2\lambda_t\mathbf{P}\Big\}^{-1}\sum_{t=1}^{T}(\mathbf{B}_t\widehat{\boldsymbol{\theta}}_b\mathbf{c}_t^{\mathsf{T}})^{\mathsf{T}}(\widehat{\boldsymbol{\gamma}}_t - \mathbf{B}_t\boldsymbol{\Theta}^{(0)}\widehat{\boldsymbol{\alpha}}_t).$$

For $\boldsymbol{\Theta}$, we first update $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_J)$ sequentially. The objective function for $\boldsymbol{\theta}_j$ is

$$\begin{aligned}\sum_{t=1}^{T}\|\widehat{\boldsymbol{\gamma}}_t - \mathbf{B}_t\widehat{\boldsymbol{\theta}}_b\widehat{\boldsymbol{\theta}}_c^{\mathsf{T}}\mathbf{c}_t - \sum_{j'\neq j}\mathbf{B}_t\boldsymbol{\theta}_{j'}\widehat{\alpha}_{j',t} - \mathbf{B}_t\boldsymbol{\theta}_j\widehat{\alpha}_{j,t}\|^2 \\ + \widehat{\sigma}^2\lambda_{pc}\boldsymbol{\theta}_j^{\mathsf{T}}\boldsymbol{\Gamma}\boldsymbol{\theta}_j + \sum_{t=1}^{T}\mathrm{tr}(\mathbf{B}_t\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_t\boldsymbol{\Theta}^{\mathsf{T}}\mathbf{B}_t^{\mathsf{T}}) - 2\sum_{t=1}^{T}\mathrm{tr}(\mathbf{B}_t\boldsymbol{\Theta}\widehat{\boldsymbol{\Lambda}}_t^{\mathsf{T}}),\end{aligned}$$

which has an analytic form

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_j = \{&\sum_{t=1}^{T}(\widehat{\alpha}_{tj}^2 + \widehat{\boldsymbol{\Sigma}}_{t,jj})\mathbf{B}_t^{\mathsf{T}}\mathbf{B}_t + \widehat{\sigma}^2\lambda_{pc}\boldsymbol{\Gamma}\}^{-1} \\ &\times\sum_{t=1}^{T}\mathbf{B}_t^{\mathsf{T}}\{(\widehat{\boldsymbol{\gamma}}_t - \mathbf{B}_t\widehat{\boldsymbol{\theta}}_b\widehat{\boldsymbol{\theta}}_c^{\mathsf{T}}\mathbf{c}_t)\widehat{\alpha}_{j,t} - \sum_{j'\neq j}(\widehat{\alpha}_{j',t}\widehat{\alpha}_{j,t} + \widehat{\boldsymbol{\Sigma}}_{t,jj'})\mathbf{B}_t\widehat{\boldsymbol{\theta}}_{j'} - \widehat{\boldsymbol{\Lambda}}_{t,j}\},\end{aligned}$$

where $\widehat{\boldsymbol{\Lambda}}_{t,j}$ is the j-th column of $\widehat{\boldsymbol{\Lambda}}_t$. To guarantee the orthonormality of $\widehat{\boldsymbol{\Theta}}$, we utilize the spectral decomposition $\widehat{\boldsymbol{\Theta}}\widehat{\mathbf{H}}_J\widehat{\boldsymbol{\Theta}}^{\mathsf{T}} = \widetilde{\mathbf{Q}}\widetilde{\mathbf{D}}\widetilde{\mathbf{Q}}^{\mathsf{T}}$, where $\widehat{\mathbf{H}_J} = \mathrm{diag}(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_J^2)$. As a result, $\widetilde{\mathbf{Q}}$ has orthonormal columns and $\widetilde{\mathbf{D}}$ is a diagonal matrix with decreasing diagonal elements. Afterwards, we replace $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\mathbf{H}}_J$ by $\widetilde{\mathbf{Q}}$ and $\widetilde{\mathbf{D}}$, respectively. Also, we transform the estimated $\widehat{\boldsymbol{\alpha}}_t$ as $\widehat{\boldsymbol{\alpha}}_t \leftarrow \widetilde{\mathbf{Q}}^T\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\alpha}}_t$. This orthogonalization procedure preserves the variance of $\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\alpha}}_t$ and realize the orthonormality of $\widehat{\boldsymbol{\Theta}}$.

Finally, to update the coefficients $\mathbf{k}$ of AR(p) model, it is equivalent to minimizing

$$\sum_{j=1}^{J} \Big\{ -\log|\mathbf{M}_j| + \frac{1}{\widehat{\sigma}_j^2}\widehat{S}_j(\mathbf{k}) \Big\}. \tag{3.15}$$

Note that the contribution of $-\log|\mathbf{M}_j|$ can be ignored when the sample size T is relatively large. We simply minimize the objective function $\sum_{j=1}^{J} \left\{ \widehat{S}_j(\mathbf{k}) / \widehat{\sigma}_j^2 \right\}$ to approximate the exact solution of (3.15). It leads to the analytic form

$$\widehat{\mathbf{k}} = \Big(\sum_{j=1}^{J} \frac{1}{\widehat{\sigma}_j^2} \widehat{\mathbf{D}}_{pj} \Big)^{-1} \sum_{j=1}^{J} \frac{1}{\widehat{\sigma}_j^2} \widehat{\mathbf{d}}_j,$$

where $\widehat{\mathbf{d}}_j = (\widehat{D}_{12,j}, \ldots, \widehat{D}_{1(p+1),j})^{\mathsf{T}}$ and $\widehat{\mathbf{D}}_{pj}$ is the right-bottom $p \times p$ major submatrix of $\widehat{\mathbf{D}}_j$.

3.4 Model Selection

When applying the penalized spline method for estimating underlying functions, we usually use moderately large number of basis functions. This is reasonable since the roughness penalty regularizes the estimation and prevents overfitting. Furthermore, the number of bivariate basis functions n_b is determined by the order d of polynomials and the order r of the smoothness parameter on the connected edges of triangulations. In practice, we set $d = 3$ as cubic order splines, and $r = 1$ for the continuous first derivative across the connected edges, which is good enough to accurately estimate the functions. For the number of univariate basis functions n_c, it will not greatly affect the model performance in the following simulation studies and Arctic sea-ice-extent data analysis when we choose a moderate one.

The regularization parameters $(\lambda_{\mu_s}, \lambda_{\mu_t}, \lambda_{pc})$ are determined by K-fold cross-validation (CV). The number of principal component functions will be determined by the empirical proportion of variances of FPC scores. The order p of autoregressive model for the latent variables will be selected using Akaike information criteria (AIC, Akaike, 1974).

3.5 Simulation Studies

In this section, we present the results of two simulation cases to assess the performance of the proposed tEFPC model with Variational-EM (VEM) algorithm and LapKFS-EM (LapEM) approach.

The domain Ω was set to be a 2×2 square with a hole in the middle as shown in Figure 2.1. The length of time points was $T = 200$. We randomly generated m locations $(x_i, y_i), i = 1, \ldots, m$ within the domain Ω and set the locations at different time points were the same. To compare the effects of different sample size, we specified the number of locations $m = 300, 400, 500, 600$. Two simulated datasets of binary case and Poisson case were generated on Ω according to the model (3.1)–(3.3). In the binary case, the cumulant function at time point t is $g(\gamma_t(x, y)) = \log\{1 + \exp(\gamma_t(x, y))\}$, while in the Poisson case, the data at time point t were generated from the Poisson distribution, $p\{z_t(x, y) = n\} = \lambda_t(x, y)^n \exp\{-\lambda_t(x, y)\}/(n!)$, with the intensity parameter function $\lambda_t(x, y) = \exp\{\gamma_t(x, y)\}$.

The t-th surface $\gamma_t(x, y)$ was generated by the combination of mean functions and $J = 2$ FPC functions. Specifically, functions $\mu_1(x, y)$, $\mu_2(t)$, $\phi_j(x, y)$'s were generated as

$$\begin{aligned}
\mu_1(x, y) &= \{\exp(\sqrt{0.1x^2 + 0.2y}) + \exp(-\sqrt{0.1x^2 + 0.2y})\}/2, \\
\mu_2(t) &= \cos(2\pi t/12), \quad \text{for binary distribution and,} \\
\mu_2(t) &= 1 + \cos(2\pi t/12), \quad \text{for Poisson distribution,} \\
\phi_1(x, y) &= 0.8578 \sin(x^2 + 0.5y^2), \\
\phi_2(x, y) &= 0.8721 \sin(0.3x^2 + 0.6y^2) - 0.2988 \sin(x^2 + 0.5y^2),
\end{aligned} \tag{3.16}$$

where the fractional numbers of $\phi_1(x, y)$ and $\phi_2(x, y)$ were assigned to make them orthornormal on the domain. We considered the noised surface $\gamma_t(x, y)$ with white noise $\epsilon_t(x, y) \sim N(0, 1)$. Moreover, the FPC scores $\alpha_{1,t}$ and $\alpha_{2,t}$ were set to be the AR(2) model

$$\alpha_{j,t} = k_1\alpha_{j,t-1} + k_2\alpha_{j,t-2} + \eta_{j,t}, \ j = 1, 2,$$

m	Methods	PA	MIAE mean	MIAE indiv
300	VEM	18.899 (22.703)	0.2227 (0.0755)	0.3819 (0.0777)
	LapEM	15.904 (12.075)	3.0286 (2.3364)	0.5417 (0.2519)
400	VEM	11.199 (10.339)	0.1950 (0.0938)	0.3286 (0.0411)
	LapEM	15.958 (11.558)	3.6025 (2.8154)	0.5910 (0.4247)
500	VEM	7.9704 (5.5771)	0.1784 (0.0701)	0.2944 (0.0322)
	LapEM	14.081 (7.6816)	3.9495 (2.6635)	0.5406 (0.2743)
600	VEM	6.6017 (3.3290)	0.1821 (0.1590)	0.2755 (0.0396)
	LapEM	15.928 (11.453)	3.7494 (2.3416)	0.5539 (0.2566)

Table 3.1: The averages with standard deviations (in parenthesis) of different criteria over 100 repeated simulations with different sample size in the case of binary distribution.

m	Methods	PA	MIAE mean	MIAE indiv
300	VEM	5.8883 (4.4304)	0.3392 (0.1274)	0.2121 (0.0516)
	LapEM	11.155 (3.1671)	0.7675 (0.0971)	0.4170 (0.0538)
400	VEM	5.3047 (5.3436)	0.3326 (0.1203)	0.1983 (0.0214)
	LapEM	10.384 (4.2993)	0.7825 (0.0926)	0.4459 (0.0290)
500	VEM	4.9288 (2.9607)	0.3124 (0.1155)	0.1910 (0.0162)
	LapEM	9.3949 (2.0503)	0.7632 (0.0928)	0.4371 (0.0277)
600	VEM	3.7108 (1.2559)	0.3005 (0.1217)	0.1863 (0.0161)
	LapEM	8.5924 (2.1931)	0.8137 (0.1170)	0.4922 (0.0375)

Table 3.2: The averages with standard deviations (in parenthesis) of different criteria over 100 repeated simulations with different sample size in the case of Poisson distribution.

with the coefficients $k_1 = 0.8, k_2 = 0.1$, and the noises $\eta_{1,t} \sim N(0,1), \eta_{2,t} \sim N(0,0.25)$.

We applied the proposed model in Section 3.1 to the simulated data. The triangulations were presented the same in Figure 2.1, on which the bivariate splines basis functions were constructed. For the function of time, we used Fourier basis functions with dimension $n_c = 11$ to approximate the unknown functions. The number of PCs and the order of AR in the simulation study were set to be the same as the true ones for simplicity. The VEM and LapEM approaches were both implemented for the comparison. The penalty parameters $(\lambda_{\mu_s}, \lambda_{\mu_t}, \lambda_{pc})$ were selected by 5-fold CV in Section 3.4 with sequential grid search suggested in Li et al. (2018). The penalty parameters of LapEM approach were set to be the same as VEM approach due to the barricade of higher

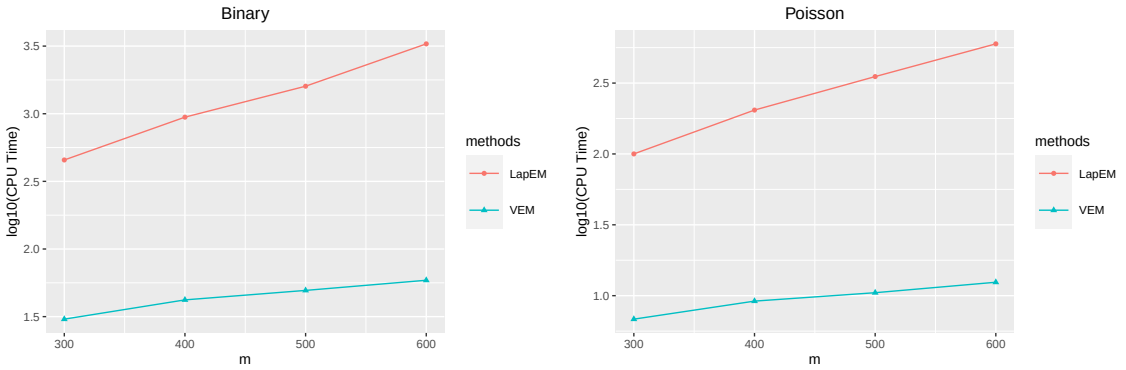

Figure 3.1: The CPU time comparison between VEM and LapEM in both binary and Poisson distributions. The CPU time is counted in seconds.

computational costs in LapEM approach.

To quantitatively measure the performance of estimating the mean $\mu_1(x,y)\mu_2(t)$ and individual surfaces $\gamma_t(x,y)$, we calculated the mean integrated abosolute errors (MIAE) through

$$\int_{\mathcal{T}}\int_{\Omega}|f(x,y,t)-\widehat{f}(x,y,t)|dxdydt,$$

where the integration was evaluated as a scaled sum over a collection of 1976 dense grid points distributed evenly on the domain. We used the principal angle (PA) to evaluate the performance of the FPC function estimations. PA is defined as $\text{angle} = \cos^{-1}(\rho)\times 180/\pi$ between the linear space spanned by the true principal component functions $\mathbf{V}$ and its estimation $\widehat{\mathbf{V}}$, where ρ is the minimum singular value of the matrix $\mathbf{Q}_{\widehat{\mathbf{V}}}^{\mathsf{T}}\mathbf{Q}_{\mathbf{V}}$. $\mathbf{Q}_{\widehat{\mathbf{V}}}$ and $\mathbf{Q}_{\mathbf{V}}$ are denoted as the orthonormal matrices of the QR decompositions of $\widehat{\mathbf{V}}$ and $\mathbf{V}$ respectively. The computational costs of VEM and LapEM approaches were assessed from the average CPU time of running one EM algorithm. Our simulation studies were run on the same computation platform of 2.40 GHz Intel(R) Xeon(R) E5-2680 CPU without implementing any parallel acceleration techniques.

The average of PAs, MIAEs of the mean and each individual function for binary and Poisson distributions are summarized in Tables 3.1 and 3.2, respectively. In both binary and Poisson cases, the VEM approach has smaller values than the LapEM approach in PAs, MIAE of mean and

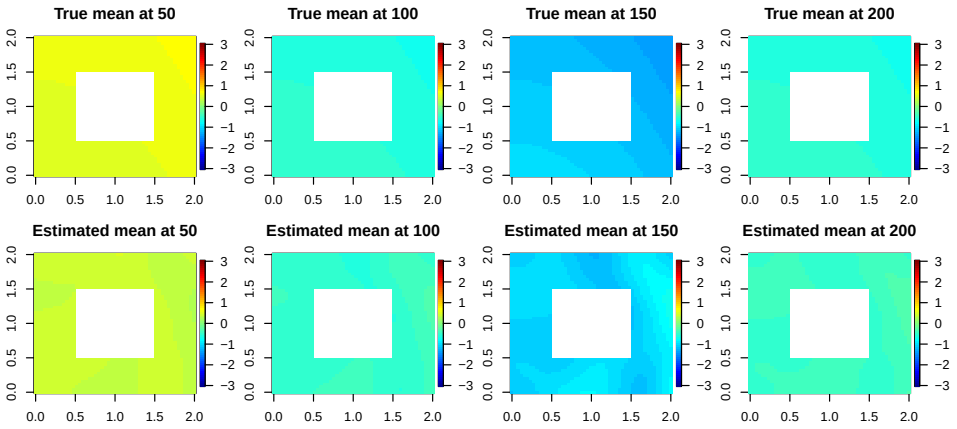

Figure 3.2: The mean functions in the binary case of simulation study. From left to right are respectively the functions at time points $t = 50, 100, 150,$ and 200. The first row represents the true mean functions while the second row represents the estimated mean functions by VEM approach.

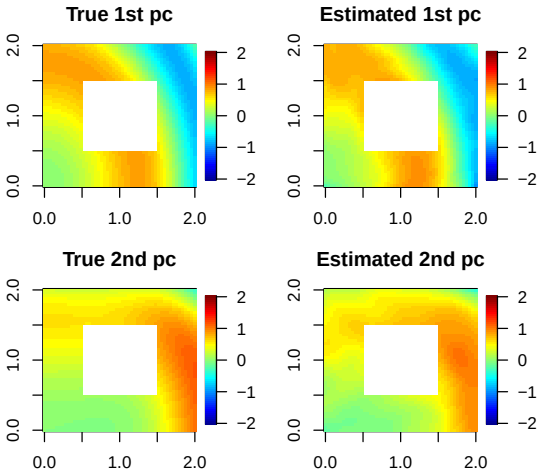

Figure 3.3: The principal component functions of binary case in simulation study. The first and second row are the first and second principal component functions respectively. The left column represents the true functions, while the right column represents the estimated functions.

individual functions. Besides, the logarithms of CPU time (in seconds) against different number of locations m by two approaches are presented in Figure 3.1, where the red line represents the logarithm of CPU time compared with LapEM approach and blue line represents its counterpart of VEM approach. It shows that VEM approach reduced more than 90% of CPU time of LapEM

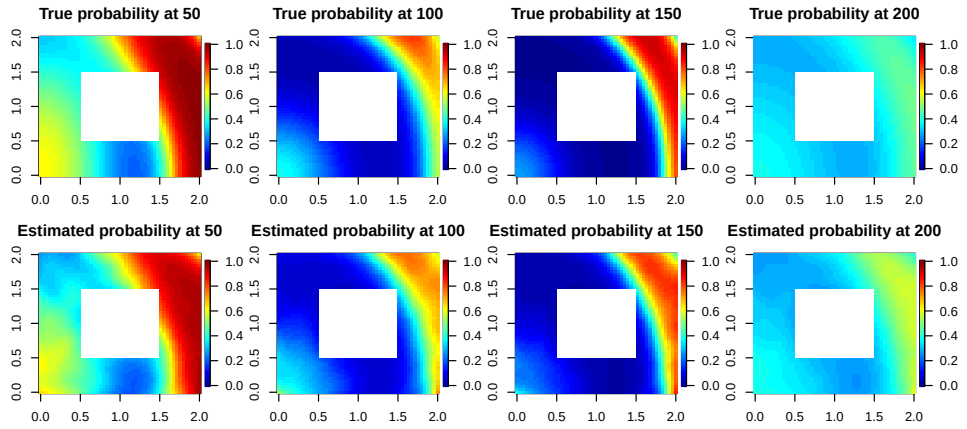

Figure 3.4: The probability surfaces of binary case in simulation study. From left to right are respectively the probability functions at time points $t = 50, 100, 150,$ and 200. The first row represents the true probability functions while the second row represents the estimated functions by VEM approach.

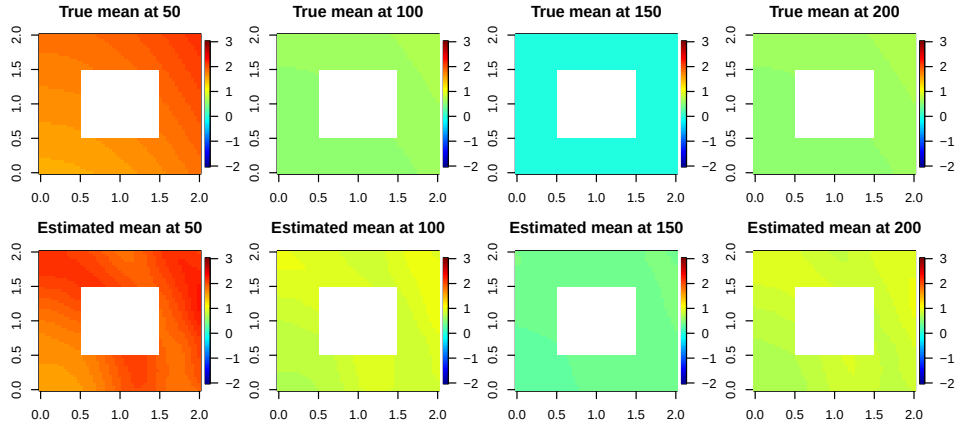

Figure 3.5: The mean functions in the Poisson case of simulation study. From left to right are respectively the functions at time points $t = 50, 100, 150,$ and 200. The first row represents the true mean functions while the second row represents the estimated mean functions by VEM approach.

approach in both binary and Poisson cases when $m = 300$. As the number of sample points m increased, the more CPU time was reduced by VEM approach compared with LapEM approach.

For the case of binary distributions, the heat maps of the true and estimated mean functions of one random replicate at time $t = 50, 100, 150, 200$ with $m = 600$ by VEM approach are depicted

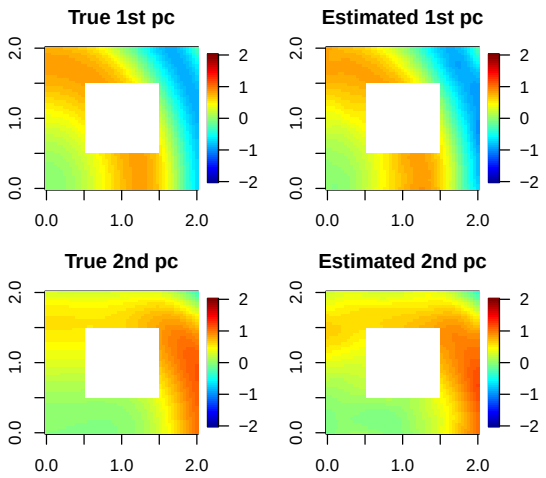

Figure 3.6: The principal component functions of Poisson case in simulation study. The first and second row are the first and second principal component functions respectively. The left column represents the true functions, while the right column represents the estimated functions.

in Figure 3.2. Figure 3.3 presents the true PC functions and estimated PC functions in the case of binary distributions. We also present the true probability surface with its estimated counterpart by VEM approach in Figure 3.4 for the case of binary distribution.

The heat maps of the true and estimated mean function by VEM approach of one random replicate in $m = 100$ are depicted in Figure 3.5. Figure 3.6 presents the true PC functions and the estimated PC functions. We also present the true natural parameter $\gamma_t(x, y)$ with its estimated $\widehat{\gamma}_t(x, y)$ by VEM approach in Figure 3.7 for the confirmation of the good estimation of functions.

3.6 Arctic Sea-ice-extent Data Analysis

In this section, we apply the proposed tEFPC model on the Arctic sea-ice-extent (SIE) monthly dataset (Meier et al., 2021, version 4, `https://nsidc.org/data/g02202`). The data was collected by the National Oceanic and Atmospheric Administration (NOAA) and National Snow & Ice Data Center (NSIDC) and analyzed in Peng et al. (2013) and Meier et al. (2014b). We use the commonly-used 15% cut-off criterion (Peng et al., 2013; Zhang and Cressie, 2019) to create a binary variable indicating water or ice. To be specific, if the numeric values are smaller than 15% and denote as $Z = 0$, we treat them as water, otherwise we treat them as ice and denote as $Z = 1$.

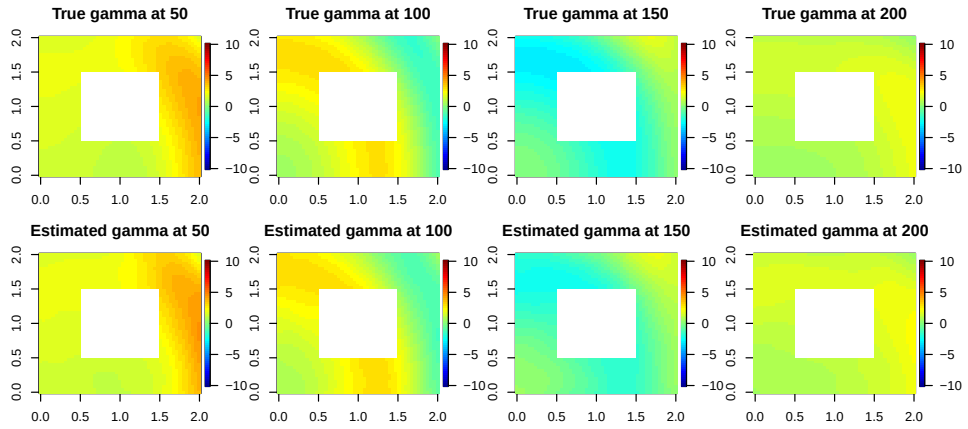

Figure 3.7: The natural-parameter functions of Poisson case in simulation study. From left to right are respectively the natural-parameter functions at time points $t = 50, 100, 150,$ and 200. The first row represents the true natural-parameter functions while the second row represents the estimated functions by VEM approach.

We consider the binary observations located in the region whose latitude is between $66.5°\mathrm{N}$ and $87°\mathrm{N}$. There are more than 4.8×10^6 observations within 20 years.

We applied the proposed model with binary distribution to analyze the sea-ice-extent data. The bivariate basis functions were constructed by the Bernstein polynomial splines on the triangulations covering the irregular domains, which are visualized as the blue triangles in Figure 3.8. We used the combination of 5-dimensional Fourier basis $(1, \cos(2\pi t/12), \sin(2\pi t/12), \cos(2 \times 2\pi t/12), \sin(2 \times 2\pi t/12))$, linear polynomials (t), and interaction term $t\cos(2\pi t/12)$ as the univariate basis functions for $\mu_2(t)$ to model the periodicity and trend of data, inflation of period amplitude, respectively. The number of principal components was selected as $J = 2$ via the scree plot. For the order of autoregressive model for FPC scores, we selected $p = 2$ through AIC. The penalty parameters $\lambda = (\lambda_{\mu_s}, \lambda_{\mu_t}, \lambda_{pc})$ were selected through the 5-fold CV with parallel grid-search. Due to the higher computational costs of LapEM approach, it was intractable to obtain the parameter estimation via LapEM in the limited computation resource. We applied the VEM algorithm, which could be implemented in a reasonable period of time, to estimate the unknown parameters.

In the bottom panel of Figure 3.9, we visualize the probability of ice $\widehat{p}(Z_t(x, y) = 1) =$

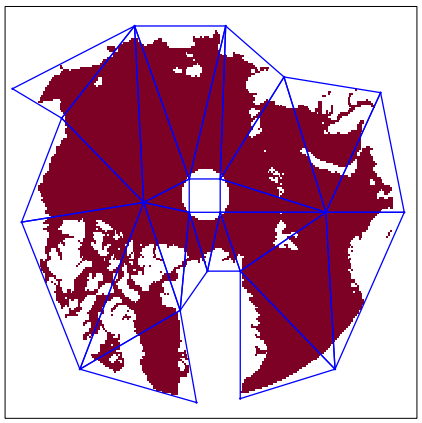

Figure 3.8: The location visualization of the sea-ice-extent monthly data, where the red irregular region represents the data domain in Arctic Circle. The blue triangles are constructed for bivariate basis functions.

$1/\{1+\exp(-\widehat{\gamma}_t(x,y))\}$ over the regions of March, September, in 2010, 2020 respectively, where $\widehat{\gamma}_t(x,y)$ were obtained from (3.6) with the parameters estimated by VEM algorithm. Compared with the observed data in the top panel of Figure 3.9, the tEFPC model provided the well-fitted heat maps to indicate the probabilities of ice in different months.

We also present the estimated bivariate function $\widehat{\mu}_1(x,y)$ and univariate function $\widehat{\mu}_2(t)$ of the mean in Figure 3.10. The left panel of Figure 3.10 presents the univariate function $\widehat{\mu}_2(t)$, which depicts the seasonality and trend of the sea ice dataset. As it shows, the lowest point in one period of $\widehat{\mu}_2(t)$ gradually decreases, which indicates the probability of ice is decreasing. Besides, the amplitude of periods is enlarged, showing that the sea ice cover are changing more and more intensely. The orthonormal function $\widehat{\mu}_1(x,y)$ in the right panel of Figure 3.10 depicts the different patterns of different regions: if the numeric values of the function are larger than 0, the region will be more likely to be ice-covered (e.g. the Arctic pole region), otherwise, it would be more likely to be water (e.g. the Greenland sea region).

The estimated PC surface functions are visualized in Figure 3.11. The first PC surface describes the major variation patterns of sea-ice data. In the left panel of Figure 3.11, the PC values are increasing gradually from the north pole (the central point) to the rims of Arctic Circle, which

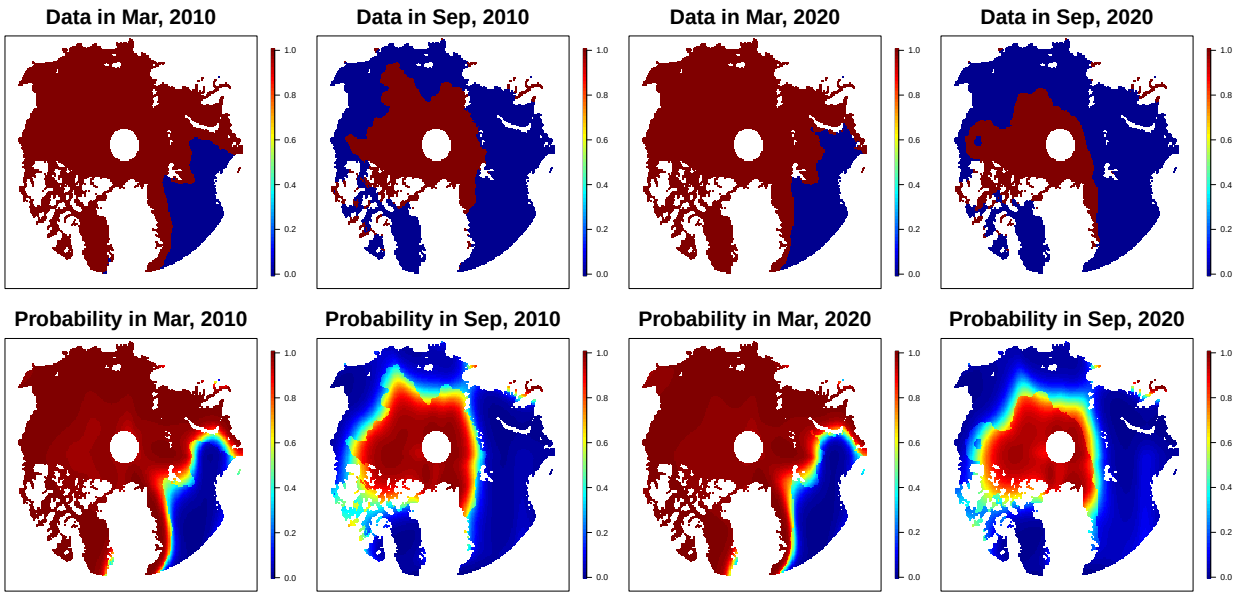

Figure 3.9: Top: observations of the sea-ice extent data on March, September in 2010, 2020; Bottom: the corresponding probability surface estimated by the proposed model.

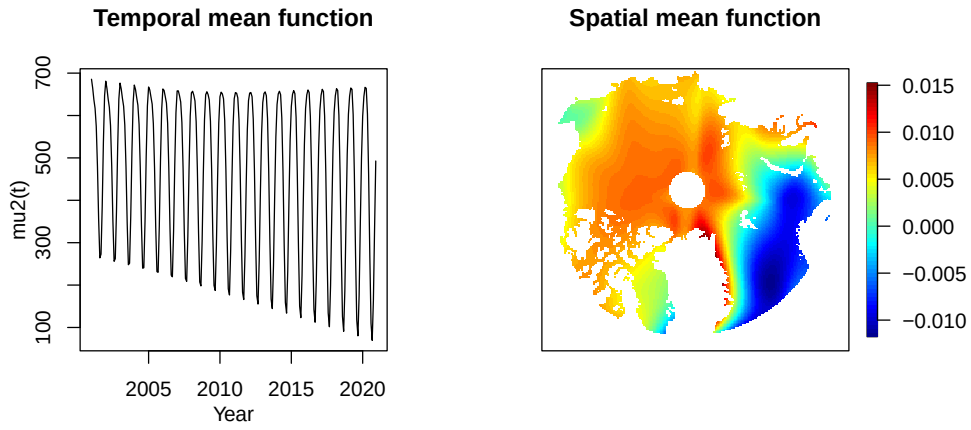

Figure 3.10: The estimated univariate function $\widehat{\mu}_2(t)$ and bivariate function $\widehat{\mu}_1(x, y)$.

indicates the global location effects. While the second PC surface in the right panel of Figure 3.11 shows the regional difference of location variation patterns. The right-bottom regions (i.e., the Greenland sea region) show negative PC values, while the central regions show the positive values.

One application of our proposed model is to forecast the sea-ice-extent. We used data from Jan-

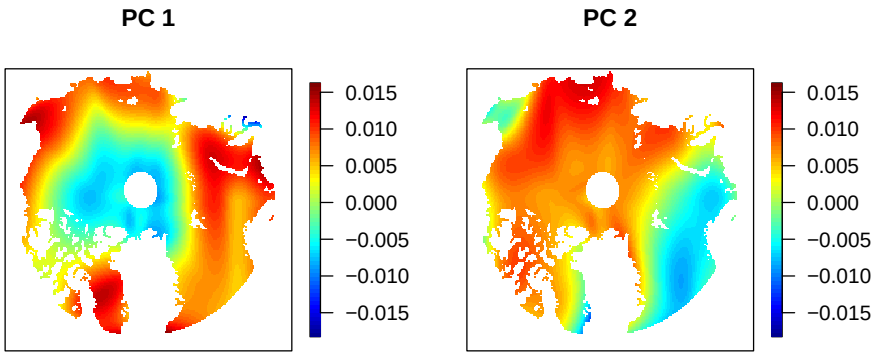

Figure 3.11: The estimated surface principal components.

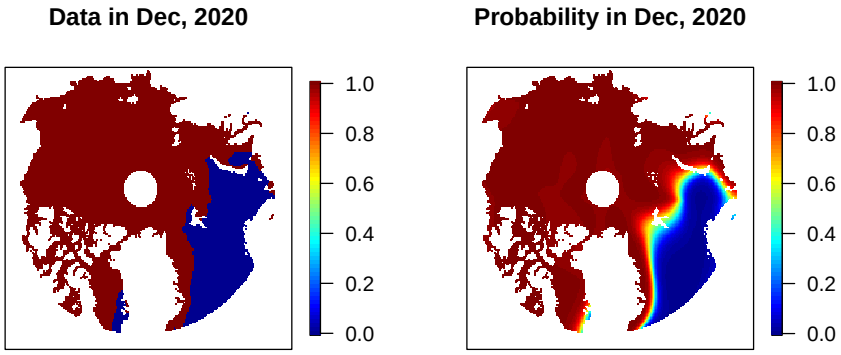

Figure 3.12: The observations and forecasted probability surface on December, 2020.

uary 2001 to November 2020 to forecast the sea-ice-extent in December 2020. After training the model with the same procedures above, we can forecast the probability surface $p(z_{T+1}(x,y)=1)$ by obtaining the individual function $\gamma_{T+1}(x,y)$. That is $p(z_{T+1}(x,y)=1)=g(\widehat{\gamma}_{T+1}(x,y))$, where $\widehat{\gamma}_{T+1}(x_i,y_i)=\mathbf{b}(x_i,y_i)\widehat{\boldsymbol{\theta}}_b\widehat{\boldsymbol{\theta}}_c\mathbf{c}_{T+1}+\mathbf{b}(x_i,y_i)\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\alpha}}_{T+1}, i=1,\ldots,m$ and $\widehat{\boldsymbol{\alpha}}_{T+1}=\sum_{\ell=1}^{p}\widehat{k}_\ell\widehat{\boldsymbol{\alpha}}_{T+1-\ell}$. Using the cutoff at 0.5, the classification rate is 95.105%. The observations and forecasted probability in December 2020 are shown in Figure 3.12.

4. SUMMARY AND DISCUSSIONS

This dissertation discussed the results of dynamic functional principal component analysis in 2-dimensional data with serial correlation. To be specific, we have proposed a mixed-effects model with functional principal component analysis to analyze the serial correlated Gaussian data in two-dimensional surfaces in Chapter 2. The autoregression was incorporated into the functional principal components for the serial correlation. We implemented the EM algorithm with Kalman filter and smoother in model fitting. We also extended the first model to deal with the binary or count data in Chapter 3 by incorporating the distributions of exponential family into the model and assuming the natural-parameters forms a decomposition similar to the first model.

There are many possible extensions for the proposed models in this dissertation. A natural extension is to consider the serial-correlated paired 2-dimensional functional data, for example, the temperature and precipitation, the sea-ice-extent and albedo. Moreover, borrowing the information of covariates and developing temporal-dependent supervised PCA for two-dimensional functional data is also of interest, and needs further investigation. Last but not least, as the outliers often exist in the real datasets, the robust dynamic functional principal component analysis with serial correlation should also be further explored.

REFERENCES

Absil, P. A., Mahony, R., and Sepulchre, R. *Optimization algorithms on matrix manifolds*. Princeton: Princeton University Press (2009).

Akaike, H. "A new look at the statistical model identification." *IEEE Transactions on Automatic Control*, 19(6):716–723 (1974).

Aue, A., Norinho, D. D., and Hörmann, S. "On the prediction of stationary functional time series." *Journal of the American Statistical Association*, 110(509):378–392 (2015).

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. "Variational inference: A review for statisticians." *Journal of the American statistical Association*, 112(518):859–877 (2017).

Bosq, D. *Linear processes in function spaces: theory and applications*, volume 149. Springer (2000).

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. *Time series analysis: Forecasting and control*. Hoboken: John Wiley & Sons (2015).

Cavalieri, D. J. and Parkinson, C. L. "Arctic sea ice variability and trends, 1979–2010." *The Cryosphere*, 6(4):881–889 (2012).

Chen, L.-H. and Jiang, C.-R. "Multi-dimensional functional principal component analysis." *Statistics and Computing*, 27(5):1181–1192 (2017).

Chiou, J.-M., Chen, Y.-T., and Yang, Y.-F. "Multivariate functional principal component analysis: A normalization approach." *Statistica Sinica*, 1571–1596 (2014).

Chiquet, J., Mariadassou, M., and Robin, S. "Variational inference for probabilistic Poisson PCA." *The Annals of Applied Statistics*, 12(4):2674–2698 (2018).

Cohen, J., Screen, J. A., Furtado, J. C., Barlow, M., Whittleston, D., Coumou, D., Francis, J., Dethloff, K., Entekhabi, D., Overland, J., et al. "Recent Arctic amplification and extreme mid-latitude weather." *Nature geoscience*, 7(9):627–637 (2014).

Cressie, N. and Wikle, C. K. *Statistics for spatio-temporal data*. John Wiley & Sons (2015).

Dai, X. and Müller, H.-G. "Principal component analysis for functional data on Riemannian man-

ifolds and spheres." *The Annals of Statistics*, 46(6B):3334–3361 (2018).

de Boor, C. *A practical guide to splines*. New York: Springer-Verlag (1978).

Delicado, P., Giraldo, R., Comas, C., and Mateu, J. "Statistics for spatial functional data: some recent contributions." *Environmetrics: The official journal of the International Environmetrics Society*, 21(3-4):224–239 (2010).

Dempster, A. P., Laird, N. M., and Rubin, D. B. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22 (1977).

Di, C.-Z., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. M. "Multilevel functional principal component analysis." *The annals of applied statistics*, 3(1):458 (2009).

Ding, F., He, S., Jones, D. E., and Huang, J. Z. "Functional PCA With Covariate-Dependent Mean and Covariance Structure." *Technometrics*, 0(0):1–11 (2022).
URL https://doi.org/10.1080/00401706.2021.2008502

Durbin, J. and Koopman, S. J. *Time series analysis by state space methods*. Oxford: Oxford University Press, 2nd edition (2012).

Gao, Y., Shang, H. L., and Yang, Y. "High-dimensional functional time series forecasting: An application to age-specific mortality rates." *Journal of Multivariate Analysis*, 170:232–243 (2019).

Giraldo, R., Delicado, P., and Mateu, J. "Hierarchical clustering of spatially correlated functional data." *Statistica Neerlandica*, 66(4):403–421 (2012).

Goldsmith, J., Zipunnikov, V., and Schrack, J. "Generalized multilevel function-on-scalar regression and principal component analysis." *Biometrics*, 71(2):344–353 (2015).

Golub, G. H. and Van Loan, C. F. *Matrix computations*. JHU press (2013).

Hall, P., Müller, H.-G., and Wang, J.-L. "Properties of principal component methods for functional and longitudinal data analysis." *The Annals of Statistics*, 34(3):1493–1517 (2006).

Hall, P., Müller, H.-G., and Yao, F. "Modelling sparse generalized longitudinal observations with latent Gaussian processes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):703–723 (2008).

Hansen, J., Sato, M., Ruedy, R., Lo, K., Lea, D. W., and Medina-Elizade, M. "Global temperature change." *Proceedings of the National Academy of Sciences*, 103(39):14288–14293 (2006).

He, K., Shen, S., and Zhou, L. "Temporal-dependent principal component analysis of two-dimensional functional data." Technical report, Renmin University of China (2022).

Heagerty, P. J. and Lele, S. R. "A composite likelihood approach to binary spatial data." *Journal of the American Statistical Association*, 93(443):1099–1111 (1998).

Hörmann, S. and Kokoszka, P. "Functional time series." In *Handbook of statistics*, volume 30, 157–186. Elsevier (2012).

Huang, H., Li, Y., and Guan, Y. "Joint modeling and clustering paired generalized longitudinal trajectories with application to cocaine abuse treatment data." *Journal of the American Statistical Association*, 109(508):1412–1424 (2014).

Hyndman, R. J. and Shang, H. L. "Forecasting functional time series." *Journal of the Korean Statistical Society*, 38(3):199–211 (2009).

Hyndman, R. J. and Ullah, M. S. "Robust forecasting of mortality and fertility rates: a functional data approach." *Computational Statistics & Data Analysis*, 51(10):4942–4956 (2007).

James, G. M., Hastie, T. J., and Sugar, C. A. "Principal component models for sparse functional data." *Biometrika*, 87(3):587–602 (2000).

James, G. M. and Sugar, C. A. "Clustering for sparsely sampled functional data." *Journal of the American Statistical Association*, 98(462):397–408 (2003).

Jones, P. D., Wigley, T. M., and Wright, P. B. "Global temperature variations between 1861 and 1984." *Nature*, 322(6078):430–434 (1986).

Karhunen, K. "Zur spektraltheorie stochastischer prozesse." *Annales Academiae Scientiarum Fennicae. Series A. I, Mathematica*, 34:1–7 (1946).

Kokoszka, P. and Reimherr, M. "Determining the order of the functional autoregressive model." *Journal of Time Series Analysis*, 34(1):116–129 (2013).

Kuenzer, T., Hörmann, S., and Kokoszka, P. "Principal component analysis of spatially indexed functions." *Journal of the American Statistical Association*, 116(535):1444–1456 (2021).

Lai, M.-J. and Schumaker, L. L. *Spline functions on triangulations*. Cambridge: Cambridge University Press (2007).

Li, G., Huang, J. Z., and Shen, H. "Exponential Family Functional data analysis via a low-rank model." *Biometrics*, 74(4):1301–1310 (2018).

Li, G., Shen, H., and Huang, J. Z. "Supervised sparse and functional principal component analysis." *Journal of Computational and Graphical Statistics*, 25(3):859–878 (2016).

Li, Y. and Guan, Y. "Functional principal component analysis of spatiotemporal point processes with applications in disease surveillance." *Journal of the American Statistical Association*, 109(507):1205–1215 (2014).

Li, Y., Qiu, Y., and Xu, Y. "From multivariate to functional data analysis: Fundamentals, recent developments, and emerging areas." *Journal of Multivariate Analysis*, 188:104806 (2022).

Lila, E., Aston, J. A., and Sangalli, L. M. "Smooth principal component analysis over two-dimensional manifolds with an application to neuroimaging." *The Annals of Applied Statistics*, 10(4):1854–1879 (2016).

Lin, Z., Wang, L., and Cao, J. "Interpretable functional principal component analysis." *Biometrics*, 72(3):846–854 (2016).

Lin, Z. and Zhu, H. "MFPCA: multiscale functional principal component analysis." In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 4320–4327 (2019).

Loève, M. "Fonctions aléatoires à décomposition orthogonale exponentielle." *La Revue Scientifique*, 84:159–162 (1946).

Martínez-Hernández, I. and Genton, M. G. "Recent developments in complex and spatially correlated functional data." *Brazilian Journal of Probability and Statistics*, 34(2):204–229 (2020).

Meier, W., Fetterer, F., Savoie, M., Mallory, S., Duerr, R., and Stroeve, J. "NOAA/NSIDC climate data record of passive microwave sea ice concentration." *Version*, 4 (2021).

Meier, W. N., Hovelsrud, G. K., Van Oort, B. E., Key, J. R., Kovacs, K. M., Michel, C., Haas, C., Granskog, M. A., Gerland, S., Perovich, D. K., et al. "Arctic sea ice in transformation: A review of recent observed changes and impacts on biology and human activity." *Reviews of Geophysics*,

52(3):185–217 (2014a).

Meier, W. N., Peng, G., Scott, D. J., and Savoie, M. H. "Verification of a new NOAA/NSIDC passive microwave sea-ice concentration climate record." *Polar Research*, 33(1):21004 (2014b).

Meier, W. N., Stroeve, J., and Fetterer, F. "Whither Arctic sea ice? A clear signal of decline regionally, seasonally and extending beyond the satellite record." *Annals of Glaciology*, 46:428–434 (2007).

Menne, M. J., Williams Jr, C. N., and Vose, R. S. "The US Historical Climatology Network monthly temperature data, version 2." *Bulletin of the American Meteorological Society*, 90(7):993–1008 (2009).

Mercer, J. "Functions of positive and negative type, and their connection the theory of integral equations." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209:415–446 (1909).

Nelder, J. A. and Mead, R. "A simplex method for function minimization." *The Computer Journal*, 7(4):308–313 (1965).

Palmer, J., Wipf, D., Kreutz-Delgado, K., and Rao, B. "Variational EM algorithms for non-Gaussian latent variable models." *Advances in neural information processing systems*, 18:1059 (2006).

Parkinson, C. L. "Global sea ice coverage from satellite data: Annual cycle and 35-yr trends." *Journal of Climate*, 27(24):9377–9382 (2014).

Parkinson, C. L., Cavalieri, D. J., Gloersen, P., Zwally, H. J., and Comiso, J. C. "Arctic sea ice extents, areas, and trends, 1978–1996." *Journal of Geophysical Research: Oceans*, 104(C9):20837–20856 (1999).

Peng, G. and Meier, W. N. "Temporal and regional variability of Arctic sea-ice coverage from satellite data." *Annals of Glaciology*, 59(76pt2):191–200 (2018).

Peng, G., Meier, W. N., Scott, D., and Savoie, M. "A long-term and reproducible passive microwave sea ice concentration data record for climate studies and monitoring." *Earth System Science Data*, 5(2):311–318 (2013).

Ramsay, J. O. and Silverman, B. W. *Functional data analysis*. New York: Springer Science & Business Media, 2nd edition (2005).

Rice, J. A. and Wu, C. O. "Nonparametric mixed effects models for unequally sampled noisy curves." *Biometrics*, 57(1):253–259 (2001).

Ruiz-Medina, M. "New challenges in spatial and spatiotemporal functional statistics for high-dimensional data." *Spatial Statistics*, 1:82–91 (2012).

Ruppert, D., Wand, M. P., and Carroll, R. J. *Semiparametric regression*. Cambridge: Cambridge University Press (2003).

Schwarz, G. "Estimating the dimension of a model." *The annals of statistics*, 461–464 (1978).

Screen, J. A., Simmonds, I., Deser, C., and Tomas, R. "The atmospheric response to three decades of observed Arctic sea ice loss." *Journal of climate*, 26(4):1230–1248 (2013).

Serban, N., Staicu, A.-M., and Carroll, R. J. "Multilevel cross-dependent binary longitudinal data." *Biometrics*, 69(4):903–913 (2013).

Sewall, J. O. and Sloan, L. C. "Disappearing Arctic sea ice reduces available water in the American west." *Geophysical Research Letters*, 31(6) (2004).

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

Shang, H. L. and Hyndman, R. J. "Grouped functional time series forecasting: An application to age-specific mortality rates." *Journal of Computational and Graphical Statistics*, 26(2):330–343 (2017).

Shen, H. "On modeling and forecasting time series of smooth curves." *Technometrics*, 51(3):227–238 (2009).

Shen, H. and Huang, J. Z. "Interday forecasting and intraday updating of call center arrivals." *Manufacturing & Service Operations Management*, 10(3):391–410 (2008).

Shi, H., Yang, Y., Wang, L., Ma, D., Beg, M. F., Pei, J., and Cao, J. "Two-Dimensional Functional Principal Component Analysis for Image Feature Extraction." *Journal of Computational and Graphical Statistics*, (just-accepted):1–26 (2022).

Staniswalis, J. G. and Lee, J. J. "Nonparametric regression analysis of longitudinal data." *Journal of the American Statistical Association*, 93(444):1403–1418 (1998).

Stroeve, J., Serreze, M., Drobot, S., Gearheard, S., Holland, M., Maslanik, J., Meier, W., and Scambos, T. "Arctic sea ice extent plummets in 2007." *Eos, Transactions American Geophysical Union*, 89(2):13–14 (2008).

Vavrus, S. and Harrison, S. P. "The impact of sea-ice dynamics on the Arctic climate system." *Climate Dynamics*, 20(7-8):741–757 (2003).

Wang, J.-L., Chiou, J.-M., and Müller, H.-G. "Functional data analysis." *Annual Review of Statistics and Its Application*, 3:257–295 (2016).

Wang, L., Wang, G., Lai, M.-J., and Gao, L. "Efficient estimation of partially linear models for data on complicated domains by bivariate penalized splines over triangulations." *Statistica Sinica*, 30:347–369 (2020).

Yao, F., Müller, H.-G., and Wang, J.-L. "Functional data analysis for sparse longitudinal data." *Journal of the American Statistical Association*, 100:577–590 (2005a).

—. "Functional linear regression analysis for longitudinal data." *The Annals of Statistics*, 33(6):2873–2903 (2005b).

Yu, S., Wang, G., Wang, L., Liu, C., and Yang, L. "Estimation and inference for generalized geoadditive models." *Journal of the American Statistical Association*, 115(530):761–774 (2020).

Zhang, B. and Cressie, N. "Estimating spatial changes over time of Arctic Sea ice using hidden $2\times$ 2 tables." *Journal of Time Series Analysis*, 40(3):288–311 (2019).

—. "Bayesian inference of spatio-temporal changes of Arctic sea ice." *Bayesian Analysis*, 15(2):605–631 (2020).

Zhang, H. and Li, Y. "Unified principal component analysis for sparse and dense functional data under spatial dependency." *Journal of Business & Economic Statistics*, 1–15 (2021).

Zhang, L., Baladandayuthapani, V., Zhu, H., Baggerly, K. A., Majewski, T., Czerniak, B. A., and Morris, J. S. "Functional CAR models for large spatially correlated functional datasets." *Journal of the American Statistical Association*, 111(514):772–786 (2016).

Zhou, L., Huang, J., and Carroll, R. "Joint modelling of paired sparse functional data using principal components." *Biometrika*, 95(3):601–619 (2008).

Zhou, L., Huang, J. Z., Martinez, J. G., Maity, A., Baladandayuthapani, V., and Carroll, R. J. "Reduced rank mixed effects models for spatially correlated hierarchical functional data." *Journal of the American Statistical Association*, 105(489):390–400 (2010).

Zhou, L. and Pan, H. "Principal component analysis of two-dimensional functional data." *Journal of Computational and Graphical Statistics*, 23(3):779–801 (2014a).

—. "Smoothing noisy data for irregular regions using penalized bivariate splines on triangulations." *Computational Statistics*, 29(1):263–281 (2014b).

APPENDIX A

APPENDIX OF CHAPTER 3

A.1 The Details of LapKFS Approach in the E Step

To obtain the conditional distribution $p(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)})$, we utilize the Bayes' formula such that

$$p(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)}) = \frac{p(\mathbf{Z}|\boldsymbol{\gamma}, \Xi^{(0)})p(\boldsymbol{\gamma}|\boldsymbol{\alpha}, \Xi^{(0)})p(\boldsymbol{\alpha}|\Xi^{(0)})}{\int p(\mathbf{Z}|\boldsymbol{\gamma}, \Xi^{(0)})p(\boldsymbol{\gamma}|\boldsymbol{\alpha}, \Xi^{(0)})p(\boldsymbol{\alpha}|\Xi^{(0)})d\mathbf{Z}}.$$

However, the integral $\int p(\mathbf{Z}|\boldsymbol{\gamma}, \Xi^{(0)})p(\boldsymbol{\gamma}|\boldsymbol{\alpha}, \Xi^{(0)})p(\boldsymbol{\alpha}|\Xi^{(0)})d\mathbf{Z}$ is intractable since the likelihood $p(\mathbf{Z}|\boldsymbol{\gamma}, \Xi^{(0)})$ is not conjugate to the prior normal distribution $p(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\Xi^{(0)})$. Instead, we utilize the Laplace approximation around the mode techniques to obtain the approximating conditional distribution $\widetilde{p}(\boldsymbol{\gamma}, \boldsymbol{\alpha}|\mathbf{Z}, \Xi^{(0)})$.

We first rewrite the model by vectorizing $\boldsymbol{\beta}_t = (\boldsymbol{\alpha}_{t+p}^{\mathsf{T}}, \boldsymbol{\alpha}_{t+p-1}^{\mathsf{T}}, \ldots, \boldsymbol{\alpha}_t^{\mathsf{T}})^{\mathsf{T}}$, such that

$$\begin{aligned} p(\mathbf{z}_t|\boldsymbol{\gamma}_t) &= \exp(\mathbf{z}_t^{\mathsf{T}}\boldsymbol{\gamma}_t - \mathbf{1}^{\mathsf{T}}g(\boldsymbol{\gamma}_t)) \prod_{i=1}^{n_t} h(z_{i,t}) \\ \boldsymbol{\gamma}_t &= \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t + \mathbf{B}_t\boldsymbol{\Theta}\mathbf{S}\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t \\ \boldsymbol{\beta}_t &= \mathbf{T}\boldsymbol{\beta}_{t-1} + \boldsymbol{\psi}_t \end{aligned} \tag{A.1}$$

where

$$\mathbf{T} = \begin{pmatrix} k_1\mathbf{I} & k_2\mathbf{I} & \ldots & k_p\mathbf{I} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \ldots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{I} & \mathbf{0}, \end{pmatrix}$$

is the coefficient matrix of the state equations and $\mathbf{S} = (\mathbf{0}, \ldots, \mathbf{0}, \mathbf{I}) \in \mathbb{R}^{J\times(p+1)J}$ is a matrix used

to indicate the component $\boldsymbol{\alpha}_t$ in $\boldsymbol{\beta}_t$. The residual error $\boldsymbol{\psi}_t$ is Gaussian distributed

$$\boldsymbol{\psi}_t = (\boldsymbol{\eta}_{t+p}^{\mathsf{T}}, \mathbf{0}, \ldots, \mathbf{0})^{\mathsf{T}} \sim N(\mathbf{0}, \widetilde{\mathbf{H}}),$$

where the variance matrix $\widetilde{\mathbf{H}} = \mathrm{diag}(\mathbf{H}_J, \mathbf{0}, \ldots, \mathbf{0}) \in \mathbb{R}^{J(p+1)\times J(p+1)}$. Besides, the initial state $\boldsymbol{\beta}_1$ is also multivariate Gaussian distributed

$$\boldsymbol{\beta}_1 = (\boldsymbol{\alpha}_{1+p}^{\mathsf{T}}, \boldsymbol{\alpha}_p^{\mathsf{T}}, \ldots, \boldsymbol{\alpha}_1^{\mathsf{T}})^{\mathsf{T}} \sim N(\mathbf{0}, \mathbf{Q}_1),$$

with variance $\mathbf{Q}_1$. In practice, $\mathbf{Q}_1$ was always assigned as the identity matrix $\mathbf{I}_{J(p+1)}$. Afterwards, we combine the joint vectors $\boldsymbol{\xi}_t = (\boldsymbol{\gamma}_t^{\mathsf{T}}, \boldsymbol{\beta}_t^{\mathsf{T}})^{\mathsf{T}}$. The distribution of $\boldsymbol{\xi}_t$ is

$$\begin{aligned}
\boldsymbol{\xi}_t = \begin{pmatrix} \boldsymbol{\gamma}_t \\ \boldsymbol{\beta}_t \end{pmatrix} &= \begin{pmatrix} \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t + \mathbf{B}_t\boldsymbol{\Theta}\mathbf{S}\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t \\ \mathbf{T}\boldsymbol{\beta}_{t-1} + \boldsymbol{\psi}_t \end{pmatrix} \\
&\sim N\Bigg(\begin{pmatrix} \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t + \mathbf{B}_t\boldsymbol{\Theta}\mathbf{S}\mathbf{T}\boldsymbol{\beta}_{t-1} \\ \mathbf{T}\boldsymbol{\beta}_{t-1} \end{pmatrix}, \begin{pmatrix} \mathbf{B}_t\boldsymbol{\Theta}\mathbf{S}\widetilde{\mathbf{H}}(\mathbf{B}_t\boldsymbol{\Theta}\mathbf{S})^{\mathsf{T}} + \sigma^2\mathbf{I} & \mathbf{B}_t\boldsymbol{\Theta}\mathbf{S}\widetilde{\mathbf{H}} \\ \widetilde{\mathbf{H}}(\mathbf{B}_t\boldsymbol{\Theta}\mathbf{S})^{\mathsf{T}} & \widetilde{\mathbf{H}} \end{pmatrix} \Bigg) \\
&= \begin{pmatrix} \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \mathbf{0} & \mathbf{B}_t\boldsymbol{\Theta}\mathbf{S}\mathbf{T} \\ \mathbf{0} & \mathbf{T} \end{pmatrix} \begin{pmatrix} \boldsymbol{\gamma}_{t-1} \\ \boldsymbol{\beta}_{t-1} \end{pmatrix} + \widetilde{\boldsymbol{\eta}}_t,
\end{aligned} \tag{A.2}$$

with $\widetilde{\boldsymbol{\eta}}_t \sim N(\mathbf{0}, \overline{\mathbf{H}})$, where

$$\overline{\mathbf{H}} = \begin{pmatrix} \mathbf{B}_t\boldsymbol{\Theta}\mathbf{S}\widetilde{\mathbf{H}}(\mathbf{B}_t\boldsymbol{\Theta}\mathbf{S})^{\mathsf{T}} + \sigma^2\mathbf{I} & \mathbf{B}_t\boldsymbol{\Theta}\mathbf{S}\widetilde{\mathbf{H}} \\ \widetilde{\mathbf{H}}(\mathbf{B}_t\boldsymbol{\Theta}\mathbf{S})^{\mathsf{T}} & \widetilde{\mathbf{H}} \end{pmatrix} = \begin{pmatrix} \sigma^2\mathbf{I}_{n_t} & \mathbf{0} \\ \mathbf{0} & \widetilde{\mathbf{H}} \end{pmatrix}.$$

Denote $\widetilde{\boldsymbol{\mu}}_t = \begin{pmatrix} \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t \\ \mathbf{0} \end{pmatrix}$ and $\mathbf{T}_t = \begin{pmatrix} \mathbf{0} & \mathbf{B}_t\boldsymbol{\Theta}\mathbf{S}\mathbf{T} \\ \mathbf{0} & \mathbf{T} \end{pmatrix}$, we rewrite the above formula of $\boldsymbol{\xi}_t$ as

$$\boldsymbol{\xi}_t = \widetilde{\boldsymbol{\mu}}_t + \mathbf{T}_t\boldsymbol{\xi}_{t-1} + \widetilde{\boldsymbol{\eta}}_t. \tag{A.3}$$

Recall,

$$p(\mathbf{z}_t|\boldsymbol{\gamma}_t) = \exp(\mathbf{z}_t^\mathsf{T}\boldsymbol{\gamma}_t - \mathbf{1}^\mathsf{T} g(\boldsymbol{\gamma}_t)) \prod_{i=1}^{n_t} h(z_{i,t}), \tag{A.4}$$

and $\boldsymbol{\gamma}_t = \widetilde{\mathbf{S}}\boldsymbol{\xi}_t$ with $\widetilde{\mathbf{S}} = (\mathbf{I}_{n_t}, \mathbf{0}_{n_t \times (p+1)J})$. Now we aim to derive the conditional distribution $p(\boldsymbol{\xi}|\mathbf{Z}, \Xi^{(0)})$. Firstly, the likelihood (A.4) can be rewritten as

$$p(\mathbf{z}_t|\boldsymbol{\xi}_t) = \exp\{\mathbf{z}_t^\mathsf{T}\widetilde{\mathbf{S}}\boldsymbol{\xi}_t - \mathbf{1}^\mathsf{T} g(\widetilde{\mathbf{S}}\boldsymbol{\xi}_t)\} \prod_{i=1}^{n_t} h(z_{i,t}).$$

We can obtain the first- and second-order derivatives of $p(\mathbf{z}_t|\boldsymbol{\xi}_t)$ with respect to $\boldsymbol{\xi}_t$, i.e.,

$$\frac{\partial \log p(\mathbf{z}_t|\boldsymbol{\xi}_t)}{\partial \boldsymbol{\xi}_t} = (\mathbf{z}_t^\mathsf{T}\widetilde{\mathbf{S}})^\mathsf{T} - \frac{\partial g(\widetilde{\mathbf{S}}\boldsymbol{\xi}_t)}{\partial \boldsymbol{\xi}_t},$$

$$\frac{\partial^2 \log p(\mathbf{z}_t|\boldsymbol{\xi}_t)}{\partial \boldsymbol{\xi}_t^2} = \frac{\partial^2 g(\widetilde{\mathbf{S}}\boldsymbol{\xi}_t)}{\partial \boldsymbol{\xi}_t^2}.$$

The Laplace approximation approach is to approximate the conditional distribution $p(\boldsymbol{\xi}|\mathbf{Z}, \Xi^{(0)})$ around its mode. We utilize the Newton-Raphson algorithm to iteratively update the estimated mode in the following discussions.

Assume we have one guess $\boldsymbol{\xi}^{(0)}$, then the new update of $\boldsymbol{\xi}^+$ can be obtained through

$$\boldsymbol{\xi}^+ = \boldsymbol{\xi}^{(0)} - \left\{ \frac{\partial^2 \log p(\boldsymbol{\xi}|\mathbf{Z})}{\partial \boldsymbol{\xi}^2}\bigg|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} \right\}^{-1} \frac{\partial \log p(\boldsymbol{\xi}|\mathbf{Z})}{\partial \boldsymbol{\xi}}\bigg|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}},$$

where the second derivative is

$$\frac{\partial^2 \log p(\boldsymbol{\xi}|\mathbf{Z})}{\partial \boldsymbol{\xi}^2}\bigg|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} = \frac{\partial^2 \log p(\mathbf{Z}|\boldsymbol{\xi})}{\partial \boldsymbol{\xi}^2} + \frac{\partial^2 \log p(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}^2}\bigg|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}},$$

and the first derivative is

$$\frac{\partial \log p(\boldsymbol{\xi}|\mathbf{Z})}{\partial \boldsymbol{\xi}}\bigg|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} = \frac{\partial \log p(\mathbf{Z}|\boldsymbol{\xi})}{\partial \boldsymbol{\xi}} + \frac{\partial \log p(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}\bigg|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}}.$$

Note that $\frac{\partial^2 \log p(\mathbf{Z}|\boldsymbol{\xi})}{\partial \boldsymbol{\xi}^2}\Big|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}}$ is a block-wise diagonal matrix with each block $\frac{\partial^2 \log p(\mathbf{Z}_t|\boldsymbol{\xi}_t)}{\partial \boldsymbol{\xi}_t^2}\Big|_{\boldsymbol{\xi}_t=\boldsymbol{\xi}_t^{(0)}}$. We then consider the update of $\boldsymbol{\xi}_t$ one by one for $t = 1, \ldots, T$. Assume the initial guess of $\boldsymbol{\xi}_t$ is $\boldsymbol{\xi}_t^{(0)}$. Borrowing the idea from Durbin and Koopman (2012), the updating procedure to approximate the conditional distribution $p(\boldsymbol{\xi}|\mathbf{Z}, \Xi^{(0)})$ around its mode can be treated as an approximating linear dynamic system. The observed equation is

$$\mathbf{x}_t = \widetilde{\mathbf{S}}\boldsymbol{\xi}_t + \widetilde{\boldsymbol{\epsilon}}_t, \tag{A.5}$$

where

$$\mathbf{x}_t = \boldsymbol{\gamma}_t^{(0)} - \big(\frac{\partial^2 \log p(\mathbf{z}_t|\boldsymbol{\gamma}_t)}{\partial \boldsymbol{\gamma}_t^2}\big|_{\boldsymbol{\gamma}_t=\boldsymbol{\gamma}_t^{(0)}}\big)^{-1} \frac{\partial \log p(\mathbf{z}_t|\boldsymbol{\gamma}_t)}{\partial \boldsymbol{\gamma}_t}\big|_{\boldsymbol{\gamma}_t=\boldsymbol{\gamma}_t^{(0)}},$$

and the innovation term $\widetilde{\boldsymbol{\epsilon}}_t \sim N(\mathbf{0}, \widetilde{\boldsymbol{\Sigma}}_t)$ with $\widetilde{\boldsymbol{\Sigma}}_t = -\big\{\frac{\partial^2 \log p(\mathbf{z}_t|\boldsymbol{\gamma}_t)}{\partial \boldsymbol{\gamma}_t^2}\big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(0)}}\big\}^{-1}$. We hence formulate a linear dynamic system (A.3) and (A.5). The Kalman filter and smoother algorithms can be applied to obtain the conditional mean $\widehat{\boldsymbol{\xi}}_t = \mathbb{E}(\boldsymbol{\xi}_t|\mathbf{Z}, \Xi^{(0)})$ and variance $\mathrm{Var}(\boldsymbol{\xi}_t|\mathbf{Z}, \Xi^{(0)})$ for $t = 1, \ldots, T$.

The Kalman filter procedure is, for $t = 1, \ldots, T$,

$$\begin{aligned}
\mathbf{v}_t &= \boldsymbol{x}_t - \widetilde{\mathbf{S}}\mathbf{b}_t, & \mathbf{F}_t &= \widetilde{\mathbf{S}}\mathbf{Q}_t\widetilde{\mathbf{S}}^{\mathsf{T}} + \widetilde{\boldsymbol{\Sigma}}_t, \\
\mathbf{b}_{t|t} &= \mathbf{b}_t + \mathbf{Q}_t\widetilde{\mathbf{S}}^{\mathsf{T}}\mathbf{F}_t^{-1}\mathbf{v}_t, & \mathbf{Q}_{t|t} &= \mathbf{Q}_t - \mathbf{Q}_t\widetilde{\mathbf{S}}^{\mathsf{T}}\mathbf{F}_t^{-1}\widetilde{\mathbf{S}}\mathbf{Q}_t, \\
\mathbf{b}_{t+1} &= \mathbf{T}_t\mathbf{b}_{t|t} + \widetilde{\boldsymbol{\mu}}_t, & \mathbf{Q}_{t+1} &= \mathbf{T}_t\mathbf{Q}_{t|t}\mathbf{T}_t^{\mathsf{T}} + \overline{\mathbf{H}}.
\end{aligned} \tag{A.6}$$

While we assign the expectation and variance of initial state $\mathbf{b}_1 = \begin{pmatrix} \mathbf{B}_t\boldsymbol{\theta}_b\boldsymbol{\theta}_c^{\mathsf{T}}\mathbf{c}_t|_{t=1} \\ \mathbf{0} \end{pmatrix}$ and $\mathbf{Q}_1 = \mathrm{diag}(\sigma^2\mathbf{I}, \mathbf{I})$.

The Kalman smoother procedure is a backshifting algorithm. Denote $\mathbf{r}_T = \mathbf{0}, \mathbf{N}_T = \mathbf{0}$. For

$t = T, \ldots, 1$, the updating formulas are listed below

$$\begin{aligned} \mathbf{r}_{t-1} &= \widetilde{\mathbf{S}}^{\mathsf{T}}\mathbf{F}_t^{-1}\mathbf{v}_t + \mathbf{L}_t^{\mathsf{T}}\mathbf{r}_t \qquad & \mathbf{N}_{t-1} &= \widetilde{\mathbf{S}}^{\mathsf{T}}\mathbf{F}_t^{-1}\widetilde{\mathbf{S}} + \mathbf{L}_t^{\mathsf{T}}\mathbf{N}_t\mathbf{L}_t \\ \widehat{\mathbf{b}}_t &= \mathbf{b}_t + \mathbf{Q}_t\mathbf{r}_{t-1} & \mathbf{V}_t &= \mathbf{Q}_t - \mathbf{Q}_t\mathbf{N}_{t-1}\mathbf{Q}_t, \end{aligned} \tag{A.7}$$

where $\mathbf{L}_t = \mathbf{T}_t - \mathbf{T}_t\mathbf{Q}_t\widetilde{\mathbf{S}}^{\mathsf{T}}\mathbf{F}_t^{-1}\widetilde{\mathbf{S}}$. The details of Kalman filter and smoother can be found in He et al. (2022).

Once we obtain $\widehat{\boldsymbol{\xi}}_t$ as the updated guess of $\boldsymbol{\xi}_t$, we repeat our Newton-Raphson procedures again until convergence. In summary, the general algorithm for finding the mode of conditional distribution $p(\boldsymbol{\xi}|\mathbf{Z}, \Xi^{(0)})$ can be listed below.

Given an initial guess of $\boldsymbol{\xi}^{(0)} = (\boldsymbol{\xi}_1^{(0)T} \ \ldots \ \boldsymbol{\xi}_T^{(0)T})^{\mathsf{T}}$,

1) implement Kalman filter and smoother based on (A.3) and (A.5) to obtain $\widehat{\boldsymbol{\xi}}_t$.

2) replace $\boldsymbol{\xi}_t^{(0)}$ by $\boldsymbol{\xi}_t$, for $t = 1, \ldots, T$

3) iteratively do 1) and 2) until the algorithm converged.

Therefore, the approximating Gaussian distribution $\widetilde{p}(\boldsymbol{\xi}_t|\mathbf{Z}, \Xi^{(0)})$ can be derived, where the mean is the mode $\boldsymbol{\xi}_t$ and the variance is $\mathbf{V}_t$ of the final iteration in Kalman filter and smoother algorithm.