

Modelling and Forecasting High-dimensional Functional Data

Yuan Gao

A thesis submitted for the degree of
Doctor of Philosophy
The Australian National University

December 2020
Draft Copy – 1 December 2020

To my family

Declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma in any university or equivalent institution, and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Yuan Gao

Acknowledgments

This thesis is the result of my Ph.D. program, funded by the Research School of Finance, Actuarial Studies and Statistics (RSFAS) at the ANU. During my Ph.D. program, I have been helped by many individuals and institutions. I would like to thank the following people, without whom I would not have been able to complete this research, and without whom I would not have made it through my Ph.D. degree.

First of all, I would like to thank all of my supervisors, Professor Han Lin Shang, Dr. Yanrong Yang, Professor Michael Martin, and Professor Alan Welsh, whose continuous assistance and encouragement throughout this period have been paramount to the completion of this thesis. I am extremely grateful to Professor Han Lin Shang for his constant guidance throughout all the stages of my Ph.D. study. His expertise in the field has been valuable to me; his patience and encouragement have made my experience rewarding. I sincerely thank Dr. Yanrong Yang for her insightful suggestions and the effort she has devoted to my work. The knowledge I have gained from her has been instrumental in completing my thesis.

I would like to express special thanks to the RSFAS for providing me with an excellent academic environment and generous funding throughout my Ph.D. program. In particular, I would like to thank Associate Professor Timothy Higgins, the Ph.D. Convenor in Statistics and the Director of Higher Degree Research at RSFAS, for providing me with considerable assistance in Ph.D-related issues. Moreover, I am grateful to Associate Professor Stephen Sault and Ms Tracy Skinner for their effort to arrange my tutorials.

Doctoral research at the ANU has been a memorable experience for me, and I would

like to thank the exceptional RSFAS faculty staff for their academic support and general help. Additionally, thanks to my fellow Ph.D. students at RSFAS, who helped make my Ph.D. study enjoyable as well as fruitful.

Finally, I owe a great debt to my father Huien Gao and mother Qingfeng Zhang. They are always there for me through all the ups and downs.

This research was supported by the Australian Government Research Training Program.

Publications

Gao, Y., & Shang, H. L. (2017). Multivariate functional time series forecasting: Application to age-specific mortality rates. *Risks*, 5(2), 21.

Gao, Y., Shang, H. L., & Yang, Y. (2017). High-dimensional functional time series forecasting. In *Functional Statistics and Related Fields* (pp. 131-136). Springer, Cham.

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

Gao, Y., Shang, H. L., & Yang, Y. (2020, June). Modelling Functional Data with High-dimensional Error Structure. *International Workshop on Functional and Operatorial Statistics* (pp. 99-106). Springer, Cham.

Abstract

This thesis summarizes the research developed along this Ph.D. trajectory. The aim of this thesis is to develop new techniques for modeling and forecasting high-dimensional functional data.

The first contribution of this thesis is to propose a functional error correction model (VECM) for the forecast of multivariate functional time series data. The model utilizes functional principal component analysis to reduce the infinite-dimensional functions to low-order principal component scores; the VECM is then applied to produce the forecast. An algorithm to generate bootstrap prediction intervals is also provided. The advantage of this model is that it not only takes into account the covariance between different groups but also can cope with data for which the assumption of stationarity does not hold. The usefulness of this model is demonstrated through a series of simulation studies and applications to the age-and sex-specific mortality rates in Switzerland and the Czech Republic.

Extending from the multivariate functional time series, the second contribution of this thesis is to address the problem of forecasting high-dimensional functional time series. We propose a twofold dimension reduction model, where dynamic functional principal component analysis is first applied to reduce each functional time series to a vector; we then use the factor model as a further dimension reduction technique so that only a small number of latent factors is preserved. Classic time series models can be used to forecast the factors, and conditional forecasts of the functions can be constructed. Asymptotic properties of the approximated functions are established, including both estimation error

and forecast error. The proposed method avoids the curse of dimensionality problem and is easy to implement. We show the superiority of our approach via both simulation studies and an application to Japanese age- and sex- specific mortality rates.

Finally, we develop a factor-augmented smoothing model for the raw functional data contaminated by high-dimensional measurement errors. The high dimensionality here concerns the dimension of the measurement error, which is in a different sense from that in the second contribution. The proposed model reduces the dimension of the measurement error with a factor model, while smoothing the functional component. We provide strong motivations from three aspects with examples. Asymptotic theorems are also established to demonstrate the effects of including factor structures on the smoothing results. As a byproduct of independent interest, an estimator for the population covariance matrix of the raw data is presented based on the proposed model. Extensive simulation studies as well as an application to Australian weather data demonstrate that these factor adjustments are extremely important in improving estimation accuracy and avoiding the curse of dimensionality.

Contents

Declaration	v
Acknowledgments	vii
Publications	ix
Abstract	xi
1 Introduction	1
2 Multivariate Functional Time Series Forecasting	9
2.1 Introduction	9
2.2 Forecasting models	12
2.2.1 Univariate autoregressive integrated moving average model	13
2.2.2 Vector autoregressive model	14
2.2.3 Vector error correction model	15
2.2.4 Product-ratio model	19
2.2.5 Bootstrap prediction interval	20
2.3 Forecast evaluation	21
2.4 Simulation studies	23
2.4.1 With co-integration	25
2.4.2 Without co-integration	25
2.4.3 Results	26

2.5	Empirical studies	27
2.5.1	Swiss age-specific mortality rates	28
2.5.2	Czech Republic age-specific mortality rates	30
2.6	Conclusions	36
3	High-dimensional Functional Time Series Forecasting	41
3.1	Introduction	41
3.2	Twofold dimension reduction	45
3.2.1	Dynamic functional principal component analysis	46
3.2.2	Factor model	47
3.2.3	Estimation	50
3.2.4	Forecasting	52
3.3	Asymptotic properties	54
3.4	Simulation Studies	58
3.4.1	Data generation	58
3.4.2	Model fitting	59
3.4.3	Results	61
3.5	Mortality rate forecast	62
3.6	Conclusion	67
3.7	Appendix	67
4	Factor-Augmented Smoothing Model for Functional data	89
4.1	Introduction	89
4.2	Motivation	93
4.2.1	Functional data with measurement error	93
4.2.2	Mis-identification of the basis function	95

4.2.3	Functional data with step jumps in the mean level	95
4.3	Model specification and estimation	98
4.3.1	Model statement	98
4.3.2	Estimation	99
4.4	Asymptotic theory	107
4.4.1	Assumptions	107
4.4.2	Asymptotic properties	110
4.5	Statistical inference on covariance matrix estimation	113
4.6	Simulation studies	115
4.6.1	Data generation	116
4.6.2	Estimation	116
4.6.3	Results	117
4.6.4	Covariance matrix estimation	118
4.6.5	Mis-identification of the basis function	120
4.6.6	Functional data with step jumps	121
4.7	Application to weather data	126
4.7.1	Canadian weather data	126
4.7.2	Australian temperature data	130
4.8	Conclusion	130
4.9	Appendices	132
4.9.1	Appendix A	132
4.9.2	Appendix B	145
4.9.3	Appendix C	149
5	Conclusions and Future Work	171

List of Figures

2.1	Simulated basis functions for the first and second populations	24
2.2	The first row presents MSPE and the mean interval scores for the two populations in the co-integration setting. The second row presents MSPE and the mean interval scores for the two populations without the co-integration setting	27
2.3	Smoothed log mortality rates in Switzerland from 1950 to 2014	28
2.4	Czech Republic: forecast errors for female and male mortality rates (MSPE and Interval scores are presented)	34
2.5	Czech Republic: P -values for the three tests comparing a functional VECM to univariate, VAR and the product-ratio models respectively (the horizontal line is the default level of significance $\alpha = 0.05$)	35
3.1	Log smoothed female mortality in the Tokyo prefecture from 1975 to 2015 .	63
3.2	Mean interval score for one- to ten-step-ahead forecast. The green solid line represents the mean interval score for the high-dimensional functional time series model. The red colored dotted line represents the mean interval score for independent functional time series forecast	66
4.1	Average daily temperature and log precipitation in 35 Canadian weather stations averaged over 1960 to 1994	94
4.2	Simulated sample of functional data with changing basis functions	96

4.3	Simulated sample of functional data with step jump	97
4.4	Applying FASM on the same data	122
4.5	The mean function $\mu(u)$	124
4.6	Residual plots of the two models	125
4.7	Comparison between the smoothed curves	127
4.8	Comparison between the smoothed curves	128
4.9	Comparison between the Residuals	129
4.10	Half-hourly Friday temperature data at Adelaide airport	131

List of Tables

2.1	MSPE for Swiss female and male rates (the smallest values are highlighted in bold)	31
2.2	Mean interval score (80%) for Swiss female and male rates (the smallest values are highlighted in bold)	32
3.1	The MNR under different settings	61
3.2	The mean MAFE and MSFE values when fitting the independent functional time series model and the proposed high-dimensional functional time series model two models for one-, two-, and three-step-ahead forecasts	62
3.3	MAFE and MSFE for the Japanese female and male rates	65
4.1	The aMSE of the function estimates with different sample sizes and dimensions. The size of η is controlled by σ	117
4.2	The MSE of the two covariance estimators with different sample sizes and dimensions. The size of η is controlled by σ	119
4.3	The trade-off between model fit and flexibility	124

Introduction

With the increasing capability to collect and store data, functional data analysis (FDA) has received growing attention over the last 20 years. Functional data are considered realizations of smooth random objects in graphical representations of curves, images, and shapes. The monographs of Ramsay & Silverman (2002, 2005) and Ramsay & Hooker (2017) provide a comprehensive account of the methodology and applications of FDA; other relevant monographs include Ferraty & Vieu (2006), Horváth & Kokoszka (2012). More recent advances in this field can be found in many survey papers (see, e.g., Shang 2014, Cuevas 2014, Febrero-Bande et al. 2017, Goia & Vieu 2016, Reiss et al. 2017, Wang et al. 2016).

When the sample of curves is collected sequentially, they form a functional time series (FTS). We denote a sequence of FTS as $\mathcal{X}_t(u)$, $t = 1, \dots, T$, and $u \in \mathcal{I} \subset \mathbb{R}$, where \mathcal{I} is a compact interval on the real line \mathbb{R} . With the increasing popularity of functional time series has come a rapidly growing body of research on functional time series modeling and forecasting. From a parametric aspect, Bosq (2012) and Bosq & Blanke (2007) proposed the functional autoregressive of order 1 (FAR(1)) and derived one-step-ahead forecasts that are based on a regularized form of the Yule-Walker equations. Later, FAR(1) was extended to FAR(p), under which the order p can be determined via the sequential hypothesis testing procedure of Kokoszka & Reimherr (2013). Klepsch & Klüppelberg (2017) proposed the

functional moving average (FMA) process and introduced an innovations algorithm to obtain the best linear predictor. Klepsch et al. (2017) extended the vector autoregressive (VAR) model to the vector autoregressive moving average model. Recently, Li et al. (2020) considered long-range dependent curve time series and proposed a functional autoregressive fractionally integrated moving average model. From a nonparametric perspective, Besse et al. (2000) and Ferraty & Vieu (2006) proposed functional kernel regression to model the temporal dependence via a similarity measure defined by semi-metric, bandwidth, and kernel functions. From a semi-parametric viewpoint, Aneiros-Pérez & Vieu (2008) put forward a semi-functional partial linear model that combines both parametric and nonparametric models; this model allows us to consider additive covariates and to use a continuous path in the past to predict future values of a stochastic process. Apart from estimating a conditional mean, Hörmann et al. (2013) considered a functional autoregressive conditional heteroskedasticity model for modeling conditional variance, while Aue et al. (2017) considered a functional generalized autoregressive conditional heteroskedasticity model. Kokoszka et al. (2017) considered a portmanteau test for testing autocorrelation under a functional analog of the generalized autoregressive conditional heteroskedasticity model.

In Chapter 2, we consider modeling and forecasting multivariate FTS, denoted $\mathcal{X}_t^{(i)}(u)$, $i = 1, \dots, N$. It is believed that, in some cases, FTS from N multiple populations are correlated and thus require simultaneous modeling. For instance, the yield curves in different economies can be modeled as multivariate FTS (Kowal et al. 2017), or female and male mortality rates in a given population can be modeled and predicted together (Hyndman et al. 2013). When generalizing scalar FTS models to multivariate FTS models, many challenges arise. To our knowledge, theory and estimation approaches have not been developed for the extension of vector autoregressive models to functional

data settings. We address this problem by proposing an alternative forecast algorithm based on functional principal component analysis (FPCA). Models based on FPCA are easy to implement, and many statistical software packages are ready to use.

Functional principal component analysis is a powerful tool in the field of FDA. It is considered an extension of the multivariate principal component analysis, and has proven to be of much more value than its multivariate counterpart. For a functional time series, functional principal component analysis is applied to reduce the infinite-dimensional functional data into the finite-dimensional principal component scores. Scalar or multiple time series models can then be utilized to model and forecast the principal component scores. For single-population FTS, Hyndman & Ullah (2007) and Hyndman & Shang (2009) suggested a curve prediction approach based on modeling functional principal component scores by scalar time series. Aue et al. (2015) claimed that there is lag autocorrelation between the principal component scores, and fitted a VAR model instead. However, there is little extant research on multivariate FTS. Hyndman et al. (2013) proposed a product-ratio model that insures coherent forecast between female and male mortality rates. Shang & Hyndman (2017) considered reconciling the forecasts of a group of hierarchical-structured FTS. In Chapter 2, we propose a functional vector error correction model (VECM) for the prediction of female and male mortality rates. After FPCA is performed, we fit a VECM to the paired principal component scores from the two sub-populations. The advantage of using functional VECM lies in the fact that it considers cointegration between multiple time series data and can cope with data for which the assumption of stationarity does not hold. Lütkepohl (2005) provided a comprehensive introduction on VECM. Yang & Wang (2013) and Zhou et al. (2014) also used VECM in forecasting multi-group demography data but not under a functional setting.

The models mentioned in the literature above are usually difficult to apply to data

where the number of FTS N is large; for instance, fitting a VECM when the dimension is larger than 10 is almost infeasible. In Chapter 3, we consider extending the idea of modeling multivariate FTS to high-dimensional FTS. By high dimension, we mean allowing the dimension of the FTS N to grow with the sample size T . When dealing with large data, dimension reduction techniques are called for. We propose using a twofold dimension reduction model where we first conduct dynamic FPCA (Hörmann et al. 2015) on each sub-population, then extract the resulting principal components from each sub-population, and last use factor models to further reduce the N -dimensional principal component scores into low dimensional time series. An appropriate scalar time series model can later be used for prediction.

We choose the factor model because it is one of the most commonly used methods to achieve dimension reduction in modeling time series of large dimensions. It is assumed that the covariance structure of the time series can be captured with a smaller number of common factors. Early attempts in this direction include Anderson (1963), Priestley et al. (1974), Brillinger (1981), Peña & Box (1987). Research that focuses on inference when $N \rightarrow \infty$ together with T includes Chamberlain (1983), Chamberlain & Rothschild (1983), Forni et al. (2000), Bai (2003). Principal component analysis that relies on the variance-covariance matrix is frequently used to estimate the latent factors (Bai & Ng 2002). However, in time series settings, the static variance-covariance matrix can not capture the lagged covariance. We adopt an estimation approach that depends on the autocovariance matrices summed at different non-zero lags (Lam et al. 2011); such a dynamic approach is also considered in papers such as Peña & Poncela (2006), Pan & Yao (2008).

In practice, within both dimension reduction steps, the number of the retained functional principal components and the number of extracted common factors need to be

determined. There are numerous ways of determining the optimal number of principal components, such as the bootstrap approach proposed by Hall & Vial (2006) and Bathia et al. (2010), the description length approach proposed by Poskitt & Sengarapillai (2013), pseudo-AIC (Shibata 1981), the scree plot (Cattell 1966), and an eigenvector variability plot (Tu et al. 2009). Under the factor model framework, Connor & Korajczyk (1993) developed a test for the number of factors in asset returns. The method based on information criteria proposed by Bai & Ng (2002) does not impose any restrictions in N or T and received much attention in econometrics. Bai & Ng (2007) and Amengual & Watson (2007) extended the work of Bai & Ng (2002), which is based on the static factor model to a restricted dynamic case. M. Hallin & Liška (2007) further extended to the general dynamic case.

In both Chapters 2 and 3, we not only propose models for h -step-ahead curve forecasts, but also introduce algorithms for constructing bootstrap prediction intervals. Bootstrap procedures are frequently used in the domain of functional time series forecasts because of their robust features and simplicity of implementation. Castro et al. (2005) used an approach based on resampling pairs of functional observations employing kernel-driven resampling probabilities. The same authors also applied a parametric, residual-based bootstrap approach using an estimated first-order functional autoregression with i.i.d. resampling of appropriately defined functional residuals. For the same prediction problem, Hyndman & Shang (2009) applied different bootstrap approaches, including bootstrapping the functional curves by randomly disturbing the forecast scores using residuals obtained from univariate autoregressive fits. Aneiros-Pérez et al. (2011) considered the nonparametric functional autoregressive models. In this thesis, we adopt a sieve bootstrap approach, which is general and easy to implement. The idea is similar to Paparoditis (2018), in which the residuals of the time series models fitted on the functional principal

component scores were resampled. In Chapter 2, we also include the uncertainty from functional smoothing and in Chapter 3, we include the uncertainty from factor modeling as well.

In practice, we do not observe the smooth curves $\mathcal{X}_i(u)$ directly; rather, the observed data are discrete points that are often contaminated by noise or measurement error. In Chapter 4, we consider the modeling of a mixture of smooth functions and high-dimensional measurement error.

$$Y_{ij} = \mathcal{X}_i(u_j) + \eta_{ij}, \quad j = 1, \dots, p; \quad i = 1, \dots, n,$$

where Y_{ij} represents the j th observation on the i th subject, and η_{ij} denotes the measurement error or the deviation of the observed data from the true underlying functions.

To retrieve the smooth curve from raw data, we consider functional smoothing. There is a vast literature on functional smoothing. Under parametric settings, \mathcal{X}_i can be written as a linear combination of an appropriate set of basis functions, and least squares can be used to find the estimate (Ramsay & Silverman 2005). Under nonparametric settings, classic smoothing techniques include kernel methods (Wand & Jones 1995, Boente & Fraiman 2000), local polynomial smoothing (Fan & Gijbels 1996, Zhang & Chen 2007), and spline smoothing (Wahba 1990, Eubank 1999). The approaches mentioned above apply smoothing to each subject and are suitable for data that can be viewed as curves observed at fine time grids with a measurement error negligible relative to the size of the smooth data. When data are only observed at sparse grids, which, for example, is often the case in the longitudinal data field, these approaches are not suitable. Recent research projects dealing with sparsely collected functional data include Yao et al. (2005a,b), Zhang & Wang (2016), and even partially observed functional data (Delaigle et al. 2020). The essential idea behind this is to apply a smoothing model to all observations collectively

rather than individual trajectories. Within the aforementioned literature, the measurement error term η_{ij} is usually assumed to be white noise; that is, i.i.d. with mean 0 and some variance. However, this may not be realistic when the data are very noisy. When the signal-to-noise level is low, the measurement error can hide the underlying smooth curves and should not be neglected. It is reasonable to impose a model on the large measurement error. To our knowledge, no research has been conducted on the case when the measurement errors are large and follow a non-zero covariance structure. In Chapter 4, we propose a factor-augmented smoothing model that imposes a factor model on the measurement error term η_{ij} while recovering the functional component. Moreover, an iterative numerical estimation approach is implemented in practice. This model avoids the problem of the "curse of dimensionality" and enables us to easily deduce further inference. As an example, we propose a covariance estimator for the observed data based on this model. Such type of covariance estimator for high-dimensional data is also constructed in papers such as Fan et al. (2008, 2011).

The remainder of this thesis is structured as follows. In Chapter 2, we review the research in the area of mortality rate modeling and introduce the functional vector error correction model for forecasting multiple functional time series. In Chapter 3, we extend multiple functional time series forecasts to high-dimensional functional time series forecasts and apply it to the Japanese prefecture-specific mortality rate. In Chapter 4, we study the deviance between the observed data and the smooth function and propose a factor-augmented smoothing model for functional data with large measurement error. Finally, we draw conclusions in Chapter 5 and provide ideas on future research directions. Each chapter in this thesis is self-contained; notations are defined in each chapter.

Multivariate Functional Time Series

Forecasting

2.1 Introduction

Most countries around the world have seen steady decrease in mortality rates in recent years, which also comes with aging populations. Policy makers from both insurance companies and government departments seek more accurate modeling and forecasting of the mortality rates. The renowned Lee-Carter model (Lee & Carter 1992) lays as a benchmark in mortality modeling. Their model is the first to decompose mortality rates into one component regarding age and the other component regarding time using singular value decomposition. Since then, many extensions are made based on Lee-Carter model. For instance, Booth et al. (2002) address the non-linearity problem in the time component. Koissi et al. (2006) propose a bootstrapped confidence interval for forecasts. Renshaw & Haberman (2006) introduce the age-period-cohort model which incorporates cohort effect in mortality modeling. Other than the Lee-Carter model, Cairns et al. (2006) propose the Cairns-Blake-Dowd (CBD) model that satisfies the new-data-invariant property. Chan et al. (2014) use a VARIMA model for the joint forecast of CBD model parameters.

It is believed that mortality trends in two or more populations could be correlated with

each other, which happens especially between sub-populations, such as females and males in a given population. This calls for a model that makes predictions in several populations simultaneously. We would also expect the forecasts of similar populations do not diverge over the long run, so coherence between the forecasts is a desired property. Carter & Lee (1992) examine how mortality rates of female and male populations could be forecast together using only one time-varying component. Li & Lee (2005) propose a model with a common factor and a population-specific factor to achieve coherency. Yang & Wang (2013) use a vector error correction model (VECM) to model the time-varying factors in multiple populations. Zhou et al. (2014) argue that the VECM performs better than the original Lee-Carter and the VAR models, and that the assumption of a dominant population is not needed. Danesi et al. (2015) compare several multi-population forecasting models and show that the preferable models are those providing a balance between model parsimony and flexibility. These mentioned approaches model mortality rates using raw data without smoothing techniques. In this chapter, we propose a model under functional data analysis (FDA) framework.

In functional data settings (see Ramsay & Silverman (2005), for a comprehensive introduction on FDA), it is assumed that there is an underlying smooth function of age as the mortality rate in each year. Since mortality rates are collected sequentially over time, we use the term functional time series for the data. Let $Y_{t,j}$ denote the log of observed mortality rate of age u_j at year t . Suppose $\mathcal{X}(u)$ is an underlying smooth function, where $u \in \mathcal{I}$ represents the age continuum defined on a finite interval. In practice, we can only observe functional data at a set of grid points and the data are often contaminated by random noise:

$$Y_{t,j} = \mathcal{X}_t(u_j) + \eta_{t,j}, \quad t = 1, \dots, n, \quad j = 1, \dots, p,$$

where n denotes the number of years and p denotes the number of discrete data points of age observed for each function. The errors $\{\eta_{t,j}\}$ are independent random variables with mean zero and variances $\sigma_{t,j}^2$. Smoothing techniques are thus needed to obtain each function $\mathcal{X}_t(u)$ from a set of realizations. Among many others, localized least squares and spline-based smoothing are some of the approaches frequently used (see for example Wahba (1975), Rice & Silverman (1991)). We are not the first to use functional data approach to model mortality rates. Hyndman & Ullah (2007) propose a model under FDA framework, and is robust to outlying years. Chiou & Müller (2014) introduce a time-varying eigenfunction to address the cohort effect. Hyndman et al. (2013) propose a product-ratio model to achieve coherency in the forecasts of multiple populations.

Our proposed method is illustrated in Section 2.2 and Appendix. It can be summarized in four steps:

- 1) Smooth the observed data in each population;
- 2) Reduce dimension of the functions in each population using functional principal component analysis (FPCA) separately;
- 3) Fit the first set of principal component scores from all populations with VECM. Then fit the second set of principal component scores with another VECM and so on. Produce forecasts using the fitted VECMs;
- 4) Produce forecasts of mortality curves.

In the papers of Yang & Wang (2013) and Zhou et al. (2014), they also use VECM to model the time-varying factor, namely the first set of principal component scores. Our model is different in the following three ways. First, the studied object is in a FDA setting. Nonparametric smoothing techniques are used to eliminate extraneous variations or noises in the observed data. Second, as with other Lee-Carter based models, only the

first set of principal component scores are used for prediction in Yang & Wang (2013) and Zhou et al. (2014). For most countries, the fraction of variance explained is not high enough for one time-varying factor to adequately explain the mortality change. Our approach uses more than one set of principal component scores, and we review some of the ways to choose the optimal number of principal component scores. Third, in their previous papers, only point forecasts are calculated, while we propose a bootstrap algorithm for constructing interval forecasts. Point and interval forecast accuracies are both considered.

This chapter is organized as follows: In Section 2.2, we revisit the existing functional time series models and puts forward a new functional time series method using a VECM. In Section 2.3, we illustrates how the forecast results are evaluated. Simulation experiments are shown in Section 2.4. In Section 2.5, real data analyses are conducted using age-and sex-specific mortality rates in Switzerland and the Czech Republic. Concluding remarks are given in Section 2.6, along with some reflections on how the methods presented here can be further extended.

2.2 Forecasting models

Let us consider predicting multivariate functional time series simultaneously. Consider two populations as an example: $\mathcal{X}_t^{(\omega)}(u)$, $\omega = 1, 2$ are the smoothed log mortality rates of each population. According to (2.8) in Appendix of this chapter, for a sequence of functional time series $\{\mathcal{X}_t^{(\omega)}(u)\}$, each element can be decomposed as:

$$\begin{aligned}\mathcal{X}_t^{(\omega)}(u) &= \mu^{(\omega)}(u) + \sum_{k=1}^{\infty} \zeta_{t,k}^{(\omega)} \phi_k^{(\omega)}(u) \\ &= \mu^{(\omega)}(u) + \sum_{k=1}^K \zeta_{t,k}^{(\omega)} \phi_k^{(\omega)}(u) + e_t^{(\omega)}(u),\end{aligned}$$

where $e_t^{(\omega)}(u)$ denotes the model truncation error function that captures the remaining terms. Thus with functional principal component (FPC) regression, each series of functions are projected onto a K -dimension space.

The functional time series curves are characterized by the corresponding principal component scores that form a time series of vectors with the dimension K : $\boldsymbol{\zeta}_t^{(\omega)} = (\zeta_{t,1}^{(\omega)}, \dots, \zeta_{t,K}^{(\omega)})^\top$. To construct h -step-ahead predictions $\widehat{\mathcal{X}}_{n+h|n}^{(\omega)}$ of the curve, we need to construct predictions for the K -dimension vectors of the principal component scores, namely $\widehat{\boldsymbol{\zeta}}_{n+h|n}^{(\omega)} = (\widehat{\zeta}_{(n+h|n),1}^{(\omega)}, \dots, \widehat{\zeta}_{(n+h|n),K}^{(\omega)})^\top$ with techniques from multivariate time series using covariance structures between multiple populations (see also Aue et al. 2015). The h -step-ahead prediction for $\mathcal{X}_{n+h|n}^{(\omega)}$ can then be constructed by forward projection

$$\begin{aligned} \widehat{\mathcal{X}}_{n+h|n}^{(\omega)} &= \mathbb{E} \left[\mathcal{X}_{n+h}^{(\omega)} | \mathcal{X}_1^{(\omega)}(u), \dots, \mathcal{X}_n^{(\omega)}(u) \right] \\ &= \widehat{\mu}^{(\omega)}(u) + \widehat{\zeta}_{(n+h|n),1}^{(\omega)} \widehat{\phi}_1^{(\omega)}(u) + \dots + \widehat{\zeta}_{(n+h|n),K}^{(\omega)} \widehat{\phi}_K^{(\omega)}(u), \quad \omega = 1, 2. \end{aligned}$$

In what follows, we consider four methods for modeling and predicting the principal component scores $\boldsymbol{\zeta}_{n+h}$, where h denotes a forecast horizon.

2.2.1 Univariate autoregressive integrated moving average model

The FPC scores can be modeled separately as univariate time series using the autoregressive integrated moving average (ARIMA(p, d, q)) model:

$$\Phi(B)(1 - B)^d \zeta_{t,k}^{(\omega)} = \Theta(B) w_{t,k}^{(\omega)}, \quad k = 1, \dots, K, \quad \omega = 1, 2,$$

where B denotes the lag operator, and $w_{t,k}$ is the white noise. $\Phi(B)$ denotes the autoregressive part and $\Theta(B)$ denotes the moving average part. The orders p, d, q can be determined automatically according to either the Akaike information criterion or the

Bayesian information criterion value (Hyndman & Khandakar 2008). Then, the maximum likelihood method can be used to estimate the parameters.

This prediction model can be quick and efficient in some cases. However, Aue et al. (2015) argue that, although the FPC scores have no instantaneous correlation, there may be autocovariance at lags greater than zero. The following model addresses this problem by using a vector time series model for the prediction of each series of FPC scores.

2.2.2 Vector autoregressive model

Model structure

Now that each function $\mathcal{X}_t^{(\omega)}(u)$ is characterized by a K -dimension vector $\boldsymbol{\zeta}_t^{(\omega)}$, we can model the $\boldsymbol{\zeta}_t^{(\omega)}$ s using a vector autoregressive (VAR)(p) model:

$$\boldsymbol{\zeta}_t^{(\omega)} = \boldsymbol{v}^{(\omega)} + \boldsymbol{A}_1^{(\omega)} \boldsymbol{\zeta}_{t-1}^{(\omega)} + \cdots + \boldsymbol{A}_p^{(\omega)} \boldsymbol{\zeta}_{t-p}^{(\omega)} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{A}^{(\omega)} = \{\boldsymbol{A}_1^{(\omega)}, \dots, \boldsymbol{A}_p^{(\omega)}\}$ are fixed $K \times K$ coefficient matrices and $\{\boldsymbol{\epsilon}_t\}$ form a sequence of i.i.d. random K -vectors with a zero mean vector. There is a whole collection of approaches to estimating VAR model parameters in Lütkepohl (2005) including multivariate least squares estimation, Yule-Walker estimation and maximum likelihood estimation.

The VAR model seeks to make use of the valuable information hidden in the data that may have been lost by depending only on univariate models. However, the model does not fully take into account the common covariance structures between the populations.

Relationship between the FAR and VAR models

As elucidated in the introduction, Bosq (2012) propose functional autoregressive models for functional time series data. Although computations for FAR(p) models are challenging if not infeasible, one exception is FAR(1), which takes the form of:

$$\mathcal{X}_t = \Psi(\mathcal{X}_{t-1}) + \epsilon_t, \quad (2.1)$$

where $\Psi : \mathcal{H} \rightarrow \mathcal{H}$ is a bounded linear operator. However, it can be proved that if an FAR(p) structure is indeed imposed on $(\mathcal{X}_t : t \in \mathbb{Z})$, then the empirical principal component scores ξ_t should approximately follow a VAR(p) model. Let us consider FAR(1) as an example. Apply $\langle \cdot, \hat{\phi}_k \rangle$ to both sides of (2.1) to obtain:

$$\begin{aligned} \langle \mathcal{X}_t, \hat{\phi}_k \rangle &= \langle \Psi(\mathcal{X}_{t-1}), \hat{\phi}_k \rangle + \langle \epsilon_t, \hat{\phi}_k \rangle \\ &= \sum_{k'=1}^{\infty} \langle \mathcal{X}_{t-1}, \hat{\phi}_{k'} \rangle \langle \Psi(\hat{\phi}_{k'}), \hat{\phi}_k \rangle + \langle \epsilon_t, \hat{\phi}_k \rangle \\ &= \sum_{k'=1}^d \langle \mathcal{X}_{t-1}, \hat{\phi}_{k'} \rangle \langle \Psi(\hat{\phi}_{k'}), \hat{\phi}_k \rangle + \delta_{t,k}, \end{aligned}$$

with remainder terms $\delta_{t,k} = d_{t,k} + \langle \epsilon_t, \hat{\phi}_k \rangle$, where $d_{t,k} = \sum_{k'=d+1}^{\infty} \langle \mathcal{X}_{t-1}, \hat{\phi}_{k'} \rangle \langle \Psi(\hat{\phi}_{k'}), \hat{\phi}_k \rangle$.

With matrix notation, we get $\xi_t = \mathbf{B}\xi_{t-1} + \delta_t$, for $t = 2, \dots, n$ where $\mathbf{B} \in \mathbb{R}^{d \times d}$. This is a VAR(1) model for the estimated principal component scores. In fact, it can be proved that the two models make asymptotically equivalent predictions (Aue et al. 2015).

2.2.3 Vector error correction model

The VAR model relies on the assumption of stationarity, however, in many cases, that assumption does not stand. For instance, age- and sex-specific mortality rates over a number of years show persistently varying mean functions. The extension we suggest here

uses the VECMs to fit pairs of principal component scores of the two populations. In a VECM, each variable in the vector is non-stationary, but there is some linear combination between the variables that are stationary in the long run. Integrated variables with this property are called co-integrated variables and the process involving co-integrated variables is called a co-integration process. For more details on VECMs, consult Lütkepohl (2005).

Fitting a VECM to principal component scores

For the k^{th} principal component score in the two populations, suppose the two are both first integrated and have a relationship of long-term equilibrium:

$$\zeta_{t,k}^{(1)} - \beta \zeta_{t,k}^{(2)} = \delta_{t,k},$$

where β is a constant and $\delta_{t,k}$ is a stable process. According to Granger's Representation Theorem, there exists the following VECM specifications exist for $\zeta_{t,k}^{(1)}$ and $\zeta_{t,k}^{(2)}$:

$$\begin{aligned} \Delta \zeta_{t,k}^{(1)} &= \alpha_1 (\zeta_{t-1,k}^{(1)} - \beta \zeta_{t-1,k}^{(2)}) + \gamma_{1,1} \Delta \zeta_{t-1,k}^{(1)} + \gamma_{1,2} \Delta \zeta_{t-1,k}^{(2)} + \epsilon_{t,k}^{(1)}, \\ \Delta \zeta_{t,k}^{(2)} &= \alpha_2 (\zeta_{t-1,k}^{(1)} - \beta \zeta_{t-1,k}^{(2)}) + \gamma_{2,1} \Delta \zeta_{t-1,k}^{(1)} + \gamma_{2,2} \Delta \zeta_{t-1,k}^{(2)} + \epsilon_{t,k}^{(2)}, \end{aligned} \quad (2.2)$$

where $\Delta \zeta_{t,k} = \zeta_{t,k} - \zeta_{t-1,k}$, $k = 1, \dots, K$, and $\alpha_1, \alpha_2, \gamma_{1,1}, \gamma_{1,2}, \gamma_{2,1}, \gamma_{2,2}$ are the coefficients, $\epsilon_{t,k}^{(1)}$ and $\epsilon_{t,k}^{(2)}$ are innovations. Note that further lags of $\Delta \zeta_{t,k}$'s may also be included. We combine the scores from two populations to vectors: $\zeta_{t,k} = (\zeta_{t,k}^{(1)}, \zeta_{t,k}^{(2)})^\top$, and $\Delta \zeta_{t,k} = (\Delta \zeta_{t,k}^{(1)}, \Delta \zeta_{t,k}^{(2)})^\top$.

For simplicity, we consider lag 1 VECM model in this chapter. In practice there are various ways of selecting the appropriate order. For example, Lütkepohl (2005) introduces using Akaike information criterion (AIC), Hannan-Quinn (HQ) criterion and Schwarz

(SC) criterion to select the order p . It is shown that HQ and SC provide consistent estimate of the true order. However, if forecasting is the objective, the forecast mean squared error is justifiable.

Estimation

In a vector notation, the model in (2.2) can be written as:

$$\zeta_{t,k} - \zeta_{t-1,k} = \alpha \beta^\top \zeta_{t-1,k} + \Gamma_1 (\zeta_{t-1,k} - \zeta_{t-2,k}) + \epsilon_{t,k},$$

or

$$\Delta \zeta_{t,k} = \alpha \beta^\top \zeta_{t-1,k} + \Gamma_1 \Delta \zeta_{t-1,k} + \epsilon_{t,k}, \quad (2.3)$$

where

$$\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}, \quad \beta^\top = \begin{pmatrix} 1 & \beta \end{pmatrix}, \quad \Gamma_1 = \begin{bmatrix} \gamma_{1,1} & \gamma_{1,2} \\ \gamma_{2,1} & \gamma_{2,2} \end{bmatrix}.$$

We further write the model in a more compact matrix form:

$$\Delta \zeta_k = \Pi_k \zeta_{-1,k} + \Gamma_1 \Delta \zeta_{-1,k} + \epsilon_k,$$

where

$$\begin{aligned}\Delta\tilde{\boldsymbol{\zeta}}_k &= [\Delta\tilde{\boldsymbol{\zeta}}_{1,k}, \dots, \Delta\tilde{\boldsymbol{\zeta}}_{n,k}], \\ \tilde{\boldsymbol{\zeta}}_{-1,k} &= [\tilde{\boldsymbol{\zeta}}_{0,k}, \dots, \tilde{\boldsymbol{\zeta}}_{n-1,k}], \\ \Delta\tilde{\boldsymbol{\zeta}}_{-1,k} &= [\Delta\tilde{\boldsymbol{\zeta}}_{0,k}, \dots, \Delta\tilde{\boldsymbol{\zeta}}_{n-1,k}], \\ \boldsymbol{\epsilon}_k &= [\boldsymbol{\epsilon}_{1,k}, \dots, \boldsymbol{\epsilon}_{t,k}].\end{aligned}$$

With this simple form, the least squares estimator is

$$[\hat{\boldsymbol{\Pi}}_k : \hat{\boldsymbol{\Gamma}}_1] = [\Delta\tilde{\boldsymbol{\zeta}}_k \tilde{\boldsymbol{\zeta}}'_{-1,k} : \Delta\tilde{\boldsymbol{\zeta}}_k \Delta\tilde{\boldsymbol{\zeta}}'_{-1,k}] \begin{bmatrix} \tilde{\boldsymbol{\zeta}}_{-1,k} \tilde{\boldsymbol{\zeta}}'_{-1,k} & \tilde{\boldsymbol{\zeta}}_{-1,k} \Delta\tilde{\boldsymbol{\zeta}}'_{-1,k} \\ \Delta\tilde{\boldsymbol{\zeta}}_{-1,k} \tilde{\boldsymbol{\zeta}}'_{-1,k} & \Delta\tilde{\boldsymbol{\zeta}}_{-1,k} \Delta\tilde{\boldsymbol{\zeta}}'_{-1,k} \end{bmatrix}^{-1}$$

For further details, refer to Lütkepohl (2005). There is a sequence of tests to determine the lag order, such as the likelihood ratio test.

Forecasting

When forecasting is the objective, the VAR form is convenient. Rearranging the terms in (2.3) gives the VAR representation:

$$\boldsymbol{\zeta}_{t,k} = (\mathbf{I}_K + \boldsymbol{\Gamma}_1 + \boldsymbol{\alpha}\boldsymbol{\beta}^\top) \boldsymbol{\zeta}_{t-1,k} - \boldsymbol{\Gamma}_1 \boldsymbol{\zeta}_{t-2,k} + \boldsymbol{\epsilon}_{t,k}.$$

With this VAR form, the optimal h -step-ahead forecast can be obtained in the same way in VAR models. The forecast with a minimal mean squared error is given by the conditional expectation.

2.2.4 Product-ratio model

Coherent forecasting refers to nondivergent forecasting for related populations (Li & Lee 2005). It aims to maintain certain structural relationships between the forecasts of related populations. When we model two or more populations, joint modeling plays a very important role in terms of achieving coherency. When modeled separately, forecast functions tend to diverge in the long run. The product-ratio model forecasts the population functions by modeling and forecasting the ratio and product of the populations. Coherence is imposed by constraining the forecast ratio function to stationary time series models. Suppose $\mathcal{X}^{(1)}(u)$ and $\mathcal{X}^{(2)}(u)$ are the smoothed functions from the two populations to be modeled together, we compute the products and ratios by:

$$p_t(u) = \sqrt{\mathcal{X}_t^{(1)}(u)\mathcal{X}_t^{(2)}(u)},$$

$$r_t(u) = \sqrt{\mathcal{X}_t^{(1)}(u)/\mathcal{X}_t^{(2)}(u)}.$$

The product $\{p_t(u)\}$ and ratio $\{r_t(u)\}$ functions are then decomposed using FPCA and the scores can be modeled separately with a stationary autoregressive moving average (ARMA)(p, q) (Box et al. 2015) in the product functions or an autoregressive fractionally integrated moving average (ARFIMA)(p, d, q) process (Granger & Joyeux 1980, Hosking 1981) in the ratio functions respectively. With the h -step-ahead forecast values for $\hat{p}_{n+h|n}(u)$ and $\hat{r}_{n+h|n}(u)$, the h -step-ahead forecast values for $\hat{\mathcal{X}}_{n+h|n}^{(1)}(u)$ and $\hat{\mathcal{X}}_{n+h|n}^{(2)}(u)$ can be derived by

$$\hat{\mathcal{X}}_{n+h|n}^{(1)}(u) = \hat{p}_{n+h|n}(u)\hat{r}_{n+h|n}(u),$$

$$\hat{\mathcal{X}}_{n+h|n}^{(2)}(u) = \hat{p}_{n+h|n}(u)/\hat{r}_{n+h|n}(u).$$

2.2.5 Bootstrap prediction interval

The point forecast itself does not provide information about the uncertainty of prediction. Constructing a prediction interval is an important part of evaluating forecast uncertainty, when the full predictive distribution is hard to specify.

The univariate model proposed by Hyndman & Ullah (2007), discussed in Section 2.2.1, computes the variance of the predicted function by adding up the variance of each component as well as the estimated error variance. The $(1 - \alpha) \times 100\%$ prediction interval is then constructed under the assumption of normality, where α denotes the level of significance. The same approach is used in the product-ratio model; however, when the normality assumption is violated, alternative approaches may be used.

Bootstrapping is used to construct prediction interval in the functional VECM that we propose. There are three sources of uncertainties in the prediction. The first is from the smoothing process. The second is from the remaining terms after the cut-off at K in the principal component regression: $\sum_{k=K+1}^n \xi_{t,k} \phi_k(u)$. If the correct number of dimensions K is picked, the residuals can be regarded as independent. The last source of uncertainty is from the prediction of scores. The smoothing errors are calculated under the assumption of normality and the other two kinds of errors are bootstrapped. All three uncertainties are added up to construct bootstrapped prediction functions. The steps are summarized in the following algorithm:

- 1) Smooth the functions with $Y_{t,j}^{(\omega)} = \mathcal{X}_t^{(\omega)}(u_j) + \eta_{t,j}^{(\omega)}$, $\omega = 1, 2$, where $\eta_{t,j}^{(\omega)}$ is the smoothing error with mean zero and estimated variance $\hat{\sigma}_{t,j}^2$, $j = 1, \dots, p$.
- 2) Perform FPCA on the smoothed functions $\mathcal{X}_t^{(1)}$ and $\mathcal{X}_t^{(2)}$ separately, and get K pairs of principal component scores $\xi_{t,k} = \left(\tilde{\xi}_{t,k}^{(1)}, \tilde{\xi}_{t,k}^{(2)} \right)^\top$.
- 3) Fit K VECM models to the principal component scores. From the fitted scores $\hat{\xi}_{t,k}$,

for $t = 1, \dots, n$ and $k = 1, \dots, K$, obtain the fitted functions $\widehat{\boldsymbol{\mathcal{X}}}_t = \left(\widehat{\boldsymbol{\mathcal{X}}}_t^{(1)}, \widehat{\boldsymbol{\mathcal{X}}}_t^{(2)} \right)^\top$.

- 4) Get residuals \boldsymbol{e}_t from $\boldsymbol{e}_t = \boldsymbol{\mathcal{X}}_t - \widehat{\boldsymbol{\mathcal{X}}}_t$.
- 5) Express the estimated VECM from step 3) in its VAR form: $\boldsymbol{\zeta}_{t,k} = \widehat{\boldsymbol{A}}_1 \boldsymbol{\zeta}_{t-1,k} + \widehat{\boldsymbol{A}}_2 \boldsymbol{\zeta}_{t-2,k} + \boldsymbol{\epsilon}_{t,k}$, $t = 1, \dots, n$ and $k = 1, \dots, K$. Construct K sets of bootstrap principal component scores time series $\boldsymbol{\zeta}_{t,k}^* = \widehat{\boldsymbol{A}}_1 \boldsymbol{\zeta}_{t-1,k}^* + \widehat{\boldsymbol{A}}_2 \boldsymbol{\zeta}_{t-2,k}^* + \boldsymbol{\epsilon}_{t,k}^*$, where the error term $\boldsymbol{\epsilon}_{t,k}^*$ is re-sampled with replacement from $\boldsymbol{\epsilon}_{t,k}$.
- 6) Refit a VECM with $\boldsymbol{\zeta}_{t,k}^*$ and make h -step-ahead predictions $\widehat{\boldsymbol{\zeta}}_{n+h|n}^*$ and hence a predicted function $\widehat{\boldsymbol{\mathcal{X}}}_{n+h|n}^*$.
- 7) Construct a bootstrapped h -step-ahead prediction for the function by

$$\widehat{\boldsymbol{\mathcal{X}}}_{n+h|n}^{**}(u_j) = \widehat{\boldsymbol{\mathcal{X}}}_{n+h|n}^*(u_j) + \boldsymbol{e}_t^* + \boldsymbol{\eta}_{t,j}^*$$

where \boldsymbol{e}_t^* is a re-sampled version of \boldsymbol{e}_t from step 4), and $\boldsymbol{\eta}_{t,j}^*$ are generated from a normal distribution with mean 0 and variance $\sigma_{t,j}^2$, where $\sigma_{t,j}^2$ is re-sampled from $\{\widehat{\sigma}_{1,j}^2, \dots, \widehat{\sigma}_{n,j}^2\}$ from step 1).

- 8) Repeat step 5) to 7) many times.
- 9) The $(1 - \alpha) \times 100\%$ point-wise prediction intervals can be constructed by taking the $\frac{\alpha}{2} \times 100\%$ and $(1 - \frac{\alpha}{2}) \times 100\%$ quantiles of the bootstrapped samples.

2.3 Forecast evaluation

We split the data set into a training set and a testing set. The four models are fitted to the data in the training set and predictions are made. The data in the testing set is then used for forecast evaluation. Following the early work by Faraway (2016), we allocate the first

two-thirds of the observations into the training set and the last third into the testing set.

We are using an expanding window approach. Suppose the size of the full data set is 60. The first 40 functions are modeled and one to twenty-step-ahead forecasts are produced. Then the first 41 functions are used to make one to nineteen-step-ahead forecasts. The process is iterated by increasing the sample size by one until reaching the end of the data. This produces 20 one-step-ahead forecasts, 19 two-step-ahead forecasts, ... and, finally, 1 twenty-step-ahead forecast. The forecast values are compared with the true values of the last 20 functions. Mean absolute prediction errors (MAPE) and mean squared prediction errors (MSPE) are used as measures of point forecast accuracy (Danesi et al. 2015). For each population, MAPE and MSPE can be calculated as:

$$\begin{aligned} \text{MAPE}(h) &= \frac{1}{(21-h) \times p} \sum_{\eta=h}^{20} \sum_{j=1}^p \left| Y_{n+\eta,j} - \hat{\mathcal{X}}_{n+\eta|n+\eta-h}(u_j) \right|, \\ \text{MSPE}(h) &= \frac{1}{(21-h) \times p} \sum_{\eta=h}^{20} \sum_{j=1}^p \left[Y_{n+\eta,j} - \hat{\mathcal{X}}_{n+\eta|n+\eta-h}(u_j) \right]^2, \end{aligned} \quad (2.4)$$

where $\hat{\mathcal{X}}_{n+\eta|n+\eta-h}$ represents the h -step-ahead prediction using the first $n + \eta - h$ years fitted in the model, and $Y_{n+\eta,j}$ denotes the holdout function.

For the interval forecast, coverage rate is a popular evaluation standard. However, coverage rate alone does not take into account the width of the prediction interval. Instead, the interval score is an appealing way that combines both a measure of coverage rate and the width of the prediction interval (Gneiting & Raftery 2007). If $\hat{\mathcal{X}}_{n+h|n}^u$ and $\hat{\mathcal{X}}_{n+h|n}^l$ are the upper and lower $(1 - \alpha) \times 100\%$ prediction bounds, and Y_{n+h} is the realized value,

the interval score at point u_j is:

$$\begin{aligned} S_\alpha(u_j) &= \left[\widehat{\mathcal{X}}_{n+h|n}^u(u_j) - \widehat{\mathcal{X}}_{n+h|n}^l(u_j) \right] \\ &+ \frac{2}{\alpha} \left[\widehat{\mathcal{X}}_{n+h|n}^l(u_j) - Y_{n+h,j} \right] \mathbb{1} \left\{ Y_{n+h,j} < \widehat{\mathcal{X}}_{n+h|n}^l(u_j) \right\} \\ &+ \frac{2}{\alpha} \left[Y_{n+h,j} - \widehat{\mathcal{X}}_{n+h|n}^u(u_j) \right] \mathbb{1} \left\{ Y_{n+h,j} > \widehat{\mathcal{X}}_{n+h|n}^u(u_j) \right\}, \end{aligned} \quad (2.5)$$

where α is the level of significance, and $\mathbb{1}\{\cdot\}$ is an indicator function. According to this standard, the best predicted interval is the one that gives the smallest interval score. In the functional case here, the point-wise interval scores are computed and the mean over the discretized ages is taken as a score for the whole curve. Then the score values are averaged across the forecast horizon to get a mean interval score at horizon h :

$$\bar{S}_\alpha(h) = \frac{1}{(21-h) \times p} \sum_{\eta=h}^{20} \sum_{j=1}^p S_\alpha \left[\widehat{\mathcal{X}}_{n+\eta|n+\eta-h}^u(u_j), \widehat{\mathcal{X}}_{n+\eta|n+\eta-h}^l(u_j); Y_{n+\eta,j} \right], \quad (2.6)$$

where p denotes the number of age groups and h denotes the forecast horizons.

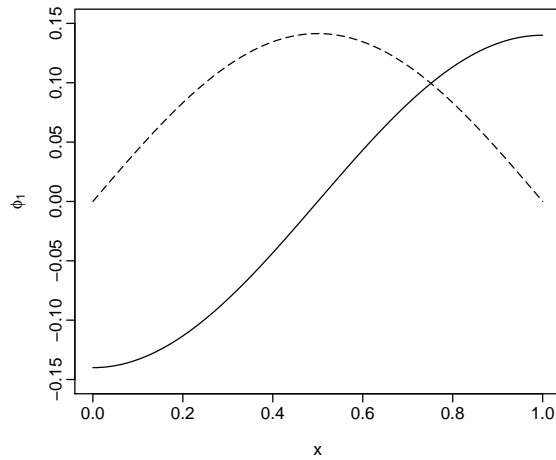
2.4 Simulation studies

In this section, we report the results from the prediction of simulated non-stationary functional time series using the models discussed in Section 2.2. We generated two series of correlated populations, each with two orthogonal basis functions. The simulated functions are constructed by

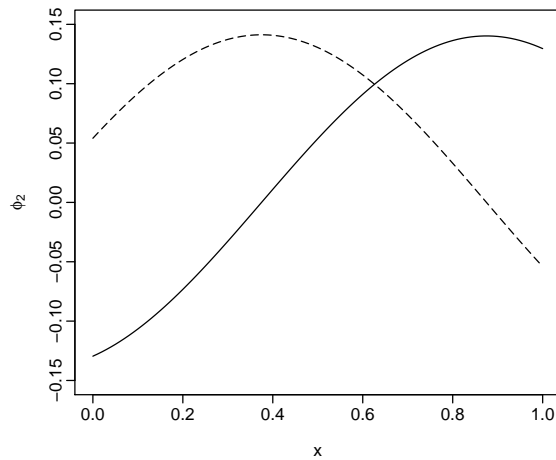
$$\mathcal{X}_t^{(\omega)}(u) = \zeta_{t,1}^{(\omega)} \phi_1^{(\omega)}(u) + \zeta_{t,2}^{(\omega)} \phi_2^{(\omega)}(u), \quad \omega = 1, 2. \quad (2.7)$$

The construction of basis functions is quite arbitrary with the only restriction being that of orthogonality. The two basis functions for the first population we used are

$\phi_1^{(1)}(u) = -\cos(\pi u)$ and $\phi_2^{(1)} = \sin(\pi u)$, and for the second population, these are $\phi_1^{(2)}(u) = -\cos(\pi u + \pi/8)$ and $\phi_2^{(2)}(u) = \sin(\pi u + \pi/8)$, where $u \in [0, 1]$. Here we are using $n = 100$ discrete data points for each function. As shown in Figure 2.1, the basis functions are scaled so that they had the L_2 norm of 1.



(a) Basis functions for population 1



(b) Basis functions for population 2

Figure 2.1: Simulated basis functions for the first and second populations

The principal component scores or coefficients $\zeta_{t,k}$ are generated with non-stationary

time series models and centered to have mean zero. In Section 2.4.1, we consider the case with co-integration and in Section 2.4.2, we consider the case without co-integration.

2.4.1 With co-integration

We first considered the case where there is a co-integration relationship between the scores of the two populations. Assuming that the principal component scores are first integrated, the two pairs of scores are generated with the following two models:

$$\begin{bmatrix} \Delta \zeta_{t,1}^{(1)} \\ \Delta \zeta_{t,1}^{(2)} \end{bmatrix} = \begin{bmatrix} -0.2 & 0.4 \\ 0.2 & -0.4 \end{bmatrix} \begin{bmatrix} \zeta_{t,1}^{(1)} \\ \zeta_{t,1}^{(2)} \end{bmatrix} + \begin{bmatrix} 0.4 & 0.3 \\ -0.3 & -0.4 \end{bmatrix} \begin{bmatrix} \Delta \zeta_{t-1,1}^{(1)} \\ \Delta \zeta_{t-1,1}^{(2)} \end{bmatrix} + \begin{bmatrix} \epsilon_{t,1}^{(1)} \\ \epsilon_{t,1}^{(2)} \end{bmatrix},$$

$$\begin{bmatrix} \Delta \zeta_{t,2}^{(1)} \\ \Delta \zeta_{t,2}^{(2)} \end{bmatrix} = \begin{bmatrix} -0.4 & 0.4 \\ 0.4 & -0.4 \end{bmatrix} \begin{bmatrix} \zeta_{t,2}^{(1)} \\ \zeta_{t,2}^{(2)} \end{bmatrix} + \begin{bmatrix} 0.3 & -0.2 \\ -0.2 & 0.3 \end{bmatrix} \begin{bmatrix} \Delta \zeta_{t-1,2}^{(1)} \\ \Delta \zeta_{t-1,2}^{(2)} \end{bmatrix} + \begin{bmatrix} \epsilon_{t,2}^{(1)} \\ \epsilon_{t,2}^{(2)} \end{bmatrix},$$

where $\epsilon_{t,k}$ are innovations that follow Gaussian distribution with mean zero and variance σ_k^2 . To satisfy the condition of decreasing eigenvalues: $\lambda_1 > \lambda_2$, we used $\sigma_1^2 = 0.1$ and $\sigma_2^2 = 0.01$.

It could easily be seen that the long term equilibrium for the first pair of scores is $-\zeta_{t,1}^{(1)} + 2\zeta_{t,1}^{(2)}$ and, for the second pair of scores, it is $-\zeta_{t,2}^{(1)} + \zeta_{t,2}^{(2)}$.

2.4.2 Without co-integration

When co-integration does not exist, there is no long term equilibrium between the two sets of scores, but they are still correlated through the coefficient matrix. We assumed that the first integrated scores follow a stable VAR(1) model:

$$\begin{bmatrix} \Delta \tilde{\zeta}_{t,1}^{(1)} \\ \Delta \tilde{\zeta}_{t,1}^{(2)} \end{bmatrix} = \begin{bmatrix} 0.4 & -0.3 \\ -0.2 & 0.4 \end{bmatrix} \begin{bmatrix} \Delta \tilde{\zeta}_{t-1,1}^{(1)} \\ \Delta \tilde{\zeta}_{t-1,1}^{(2)} \end{bmatrix} + \begin{bmatrix} \epsilon_{t,1}^{(1)} \\ \epsilon_{t,1}^{(2)} \end{bmatrix},$$

$$\begin{bmatrix} \Delta \tilde{\zeta}_{t,2}^{(1)} \\ \Delta \tilde{\zeta}_{t,2}^{(2)} \end{bmatrix} = \begin{bmatrix} 0.3 & 0.1 \\ 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} \Delta \tilde{\zeta}_{t-1,2}^{(1)} \\ \Delta \tilde{\zeta}_{t-1,2}^{(2)} \end{bmatrix} + \begin{bmatrix} \epsilon_{t,2}^{(1)} \\ \epsilon_{t,2}^{(2)} \end{bmatrix}.$$

For a VAR(1) model to be stable, it is required that $\det(\mathbf{I}_p - \mathbf{A}_1 z) = 0$ should have all roots outside the unit circle.

2.4.3 Results

The principal component scores are generated using the aforementioned two models for observations $t = 1, \dots, 60$. Two sets of simulated functions are generated using (2.7). We performed an FPCA on the two populations separately. Then the estimated principal component scores are modeled using the univariate model, the VAR model and the VECM.

We repeated the simulation procedures 150 times. In each simulation, 500 bootstrap samples are generated to calculate the prediction intervals. We show the MSPE and the mean interval scores at each forecast horizon in Figure 2.2. The three models performed almost equally well in the short term forecasts. In the long run, however, the functional VECM produced apparently better predictions than the other two models. This advantage grew bigger as the forecast horizons increased.

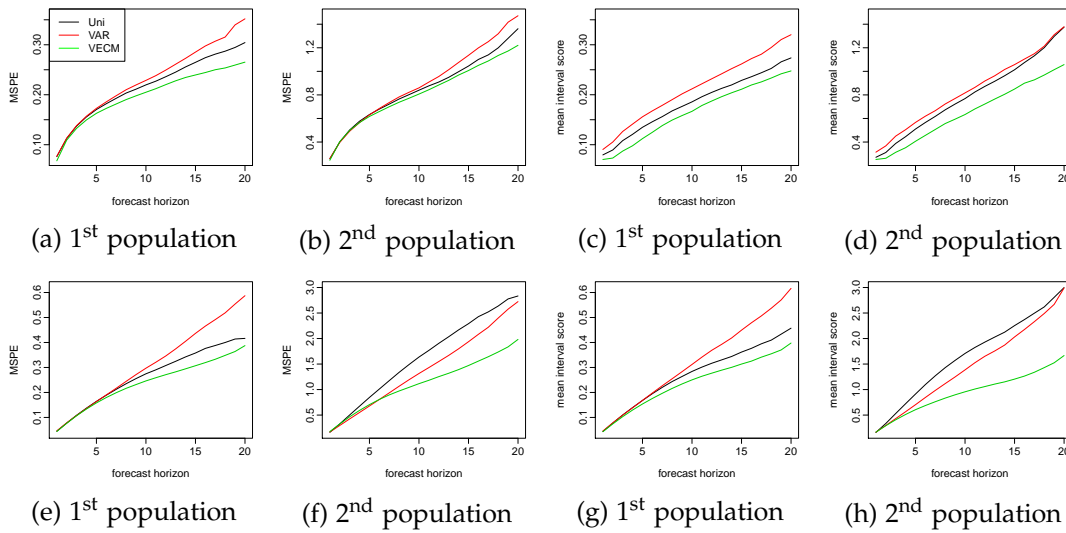


Figure 2.2: The first row presents MSPE and the mean interval scores for the two populations in the co-integration setting. The second row presents MSPE and the mean interval scores for the two populations without the co-integration setting

2.5 Empirical studies

To show that the proposed model outperformed the existing ones using real data, we applied the four models illustrated in Section 2.2 to the age- and sex-specific mortality rates in Switzerland and the Czech Republic. The observations are yearly mortality curves from ages 0 to 110 years, where the age is treated as the continuum in the rate function. Female and male curves are available from 1908 to 2014 in Human Mortality Database (2016). We only used data from 1950 to 2014 for analysis to avoid the possibly abnormal rates before 1950 due to war deaths. With the aim of forecasting, we considered the data before 1950 to be too distant to provide useful information. The data at ages 95 and older are grouped together, to avoid problems associated with erratic rates at these ages.

2.5.1 Swiss age-specific mortality rates

Figure 2.3 exhibits smoothed log mortality rates for female and male from 1950 to 2014. We use a rainbow plot (Hyndman & Shang 2010), where the red color represents the curves for more distant years and the purple color represents the curves for more recent years. The curves are smoothed using penalized regression splines with a monotonically increasing constraint after the age of 65 (Wood 1994*b*, Hyndman & Ullah 2007). Over a span of 65 years, the mortality rates in general have decreased over all ages, with exceptions in the male population at around age 20. Female rates have been slightly lower than male rates over the years.

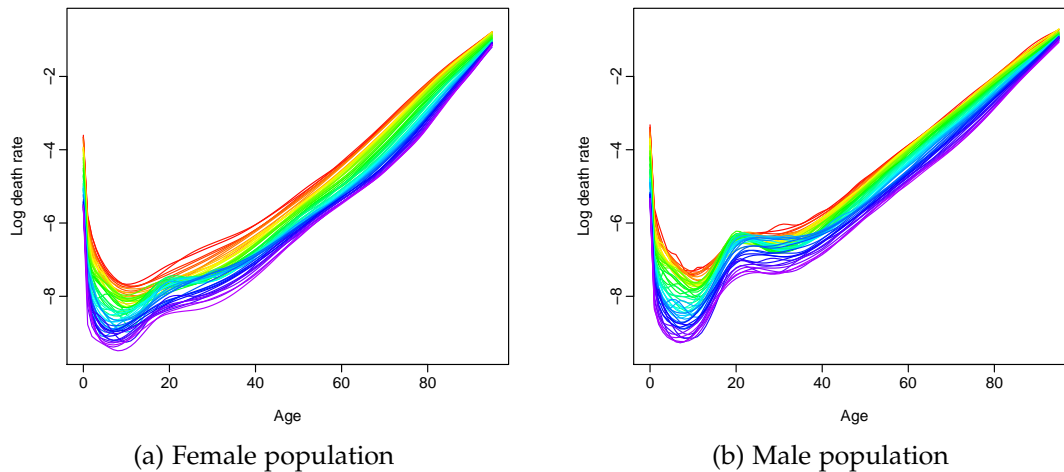


Figure 2.3: Smoothed log mortality rates in Switzerland from 1950 to 2014

First, we tested the stationarity of our data set. The Monte Carlo test, in which the null hypothesis is stationarity, is applied to both the male and female populations. We used data from all 65 of the years in our range and performed 5,000 Monte Carlo replications (Horvath et al. 2014). The p -values for the male and female populations are 0.0256 and 0.0276 respectively. These small p -values indicated a strong deviation from stationary functional time series.

The first 45 years of data (from 1950 to 1994) are allocated to the training set, and the last 20 years of data from (1995 to 2014) are allocated to the testing set. To choose the order K , we further divided the training set into two groups of 30 and 15 years. The model is fitted to the first 30 years from (1950 to 1979) and made forecasts for the next 15 years from (1980 to 1994). In both the VAR model and the functional VECM, K is chosen using:

$$K = \underset{m}{\operatorname{argmin}} \left\{ \frac{1}{15} \sum_{h=1}^{15} \sum_{j=0}^{95} \left[\hat{\mathcal{X}}_{n'+h|n'}(u_j; m) - Y_{n'+h,j} \right]^2 \right\},$$

where $\hat{\mathcal{X}}_{n'+h|n'}(u_j; m)$ denotes the h -step-ahead forecast based on the first $n' = 30$ years of data, with m dimensions retained. $Y_{n'+h}$ denotes the true rate at year $n' + h$. This selection scheme led to both the VAR and VECM models with $K = 3$ basis functions in this case, which explained 91.20%, 4.37% and 1.56% of the variation in the training set respectively. These add up to 97.13% of the total variances in the training data being explained. In the univariate and the product-ratio models, order $K = 6$ is used as in Hyndman & Booth (2008) and Hyndman et al. (2013), where they found that six components would suffice and that having more than six made no difference to the forecasts. Now that with chosen K values, the four models are fitted using an expanding window approach (as explained in Section 2.3). This produced 20 one-step-ahead forecasts, 19 two-step-ahead forecasts, . . . and, finally, 1 twenty-step-ahead forecast. These forecasts are compared with the holdout data from the years 1995 to 2014. We calculated MAPE and MSPE as point forecast errors using (2.4).

Table 2.1 presents the MSPE of log mortality rates. The smallest errors at each forecast horizon are highlighted in bold face. For the prediction of female rates, the proposed functional VECM has proved to make more accurate point forecasts for all forecast

horizons except for the twenty-step-ahead prediction. It should be noted that there is only one error estimate for the twenty-step-ahead forecast, so the error estimate may be quite volatile. The other three approaches are somewhat competitive for the eleven-step-ahead forecasts, or less. For the longer forecast horizons, the errors of the product-ratio method increase quickly. For the forecasting of male mortality rates, although the VAR model produces slightly smaller values of the forecast errors, there is hardly any difference between the four models in the short term. For long-term predictions, the product-ratio approach performs much better than the univariate and the VAR models, but the VECM still dominates. In fact, the product-ratio model usually outperforms the existing models for the male mortality forecasts, while for the female mortality forecasts, it is not as accurate.

To examine how the models perform in interval forecasts, (2.5) and (2.6) are used to calculate the mean interval scores. We generate 1000 bootstrap samples in the functional VECM and VAR. Table 2.2 shows the mean interval scores. The 80% prediction intervals are produced using the four different approaches. As explained earlier, smaller mean interval score values indicate better interval predictions. For the female forecasts, the functional VECM makes superior interval predictions at all forecast steps, while for the male forecasts, the product-ratio model and the VECM are very competitive, with the latter having a minor advantage in the mean value.

2.5.2 Czech Republic age-specific mortality rates

We have also applied the four models to other countries, such as Czech Republic, to show that the proposed functional VECM does not only work in the one case of Swiss mortality rates. The raw data are grouped and smoothed as that is done for the Swiss data. $K = 5$ is chosen in the VAR and the VECM, and the proportions of explained variance are 93.04%,

Table 2.1: MSPE for Swiss female and male rates (the smallest values are highlighted in bold)

h	Female				Male			
	UNI	VAR	PR	VECM	UNI	VAR	PR	VECM
1	0.081	0.082	0.076	0.074	0.050	0.048	0.049	0.049
2	0.085	0.088	0.079	0.075	0.056	0.052	0.053	0.053
3	0.090	0.094	0.084	0.078	0.065	0.059	0.060	0.060
4	0.096	0.104	0.091	0.082	0.077	0.067	0.070	0.069
5	0.103	0.112	0.098	0.086	0.090	0.078	0.080	0.078
6	0.109	0.119	0.107	0.090	0.107	0.093	0.093	0.089
7	0.117	0.130	0.119	0.096	0.129	0.115	0.109	0.104
8	0.125	0.140	0.130	0.102	0.149	0.136	0.124	0.119
9	0.136	0.151	0.145	0.111	0.171	0.160	0.139	0.129
10	0.145	0.163	0.157	0.116	0.198	0.191	0.160	0.149
11	0.156	0.171	0.173	0.125	0.224	0.223	0.178	0.162
12	0.167	0.186	0.195	0.133	0.261	0.269	0.206	0.184
13	0.174	0.192	0.210	0.137	0.299	0.317	0.232	0.201
14	0.188	0.203	0.238	0.145	0.344	0.361	0.260	0.213
15	0.183	0.209	0.254	0.141	0.396	0.414	0.293	0.228
16	0.197	0.219	0.281	0.152	0.460	0.444	0.332	0.239
17	0.209	0.223	0.327	0.164	0.538	0.556	0.373	0.251
18	0.209	0.233	0.354	0.165	0.649	0.652	0.416	0.263
19	0.197	0.232	0.457	0.162	0.792	0.733	0.502	0.253
20	0.144	0.249	0.493	0.175	0.904	0.753	0.525	0.270
Mean	0.145	0.165	0.203	0.120	0.298	0.286	0.213	0.158
Median	0.145	0.265	0.173	0.120	0.224	0.223	0.178	0.158

Table 2.2: Mean interval score (80%) for Swiss female and male rates (the smallest values are highlighted in bold)

h	Female				Male			
	UNI	VAR	PR	VECM	UNI	VAR	PR	VECM
1	1.089	1.042	0.865	0.852	0.871	0.767	0.657	0.715
2	1.114	1.042	0.878	0.864	0.964	0.786	0.699	0.748
3	1.153	1.059	0.909	0.880	1.088	0.852	0.759	0.791
4	1.204	1.102	0.954	0.902	1.243	0.911	0.838	0.839
5	1.254	1.136	0.997	0.926	1.407	1.011	0.909	0.887
6	1.306	1.169	1.046	0.964	1.594	1.134	1.005	0.954
7	1.358	1.234	1.113	0.996	1.789	1.289	1.113	1.059
8	1.413	1.276	1.166	1.026	1.969	1.430	1.190	1.133
9	1.483	1.349	1.241	1.088	2.134	1.587	1.282	1.204
10	1.532	1.426	1.287	1.113	2.326	1.798	1.388	1.338
11	1.608	1.479	1.358	1.170	2.476	2.012	1.475	1.458
12	1.661	1.591	1.437	1.209	2.655	2.303	1.609	1.628
13	1.716	1.647	1.463	1.237	2.819	2.618	1.706	1.767
14	1.766	1.723	1.540	1.281	3.001	2.892	1.793	1.891
15	1.705	1.775	1.571	1.262	3.145	3.082	1.892	1.963
16	1.774	1.790	1.638	1.304	3.309	3.180	1.957	1.986
17	1.852	1.860	1.760	1.352	3.521	3.692	2.041	2.011
18	1.819	1.884	1.767	1.368	3.632	4.148	2.036	2.051
19	1.795	1.986	1.941	1.360	3.683	4.254	2.175	1.974
20	1.679	2.347	2.176	1.398	3.873	3.595	2.375	1.978
Mean	1.514	1.496	1.355	1.128	2.375	2.167	1.445	1.419
Median	1.532	1.479	1.355	1.128	2.375	2.012	1.445	1.419

1.99%, 1.55%, 1.18% and 0.79% respectively, which add up to 98.55% of the total variance being explained. Figure 2.4 shows the MSPE and mean interval scores for the point and interval forecast evaluations. In order to compare with the VECM model in the literature, we also try fitting only the first set of principal component scores, shown in the figure by VECM*. Among all five models, the functional VECM produces better predictions in both the point and interval forecast. Compared to our model that uses 5 principal component scores, VECM* produces larger error especially in male forecasts. We consider an important fraction of information is lost if only the first set of principal component scores is used.

To examine whether or not the differences in the forecast errors are significant, we conduct the Diebold-Mariano test (Diebold & Mariano 1995). We used a null hypothesis where the two prediction methods had the same forecast accuracy at each forecast horizon, while the three alternative hypotheses used are that the functional VECM method produces more accurate forecasts than the three other methods. Thus, a small p -value is expected in favor of the alternatives. A squared error loss function is used and the p -values for one-sided tests are calculated at each forecast horizon, as shown in Figure 2.5. The p -values are hardly greater than zero at most forecast horizons. Almost all are below $\alpha = 0.05$, denoted by the horizontal line, with the exception of the nineteen- and twenty-step-ahead forecasts. We conclude that there is strong evidence that the functional VECM method produces more accurate forecasts than the other three for most of the forecast horizons.

In summary, we apply the proposed functional VECM to modeling female and male mortality rates in Switzerland and Czech Republic, and prove its advantage in forecasting.

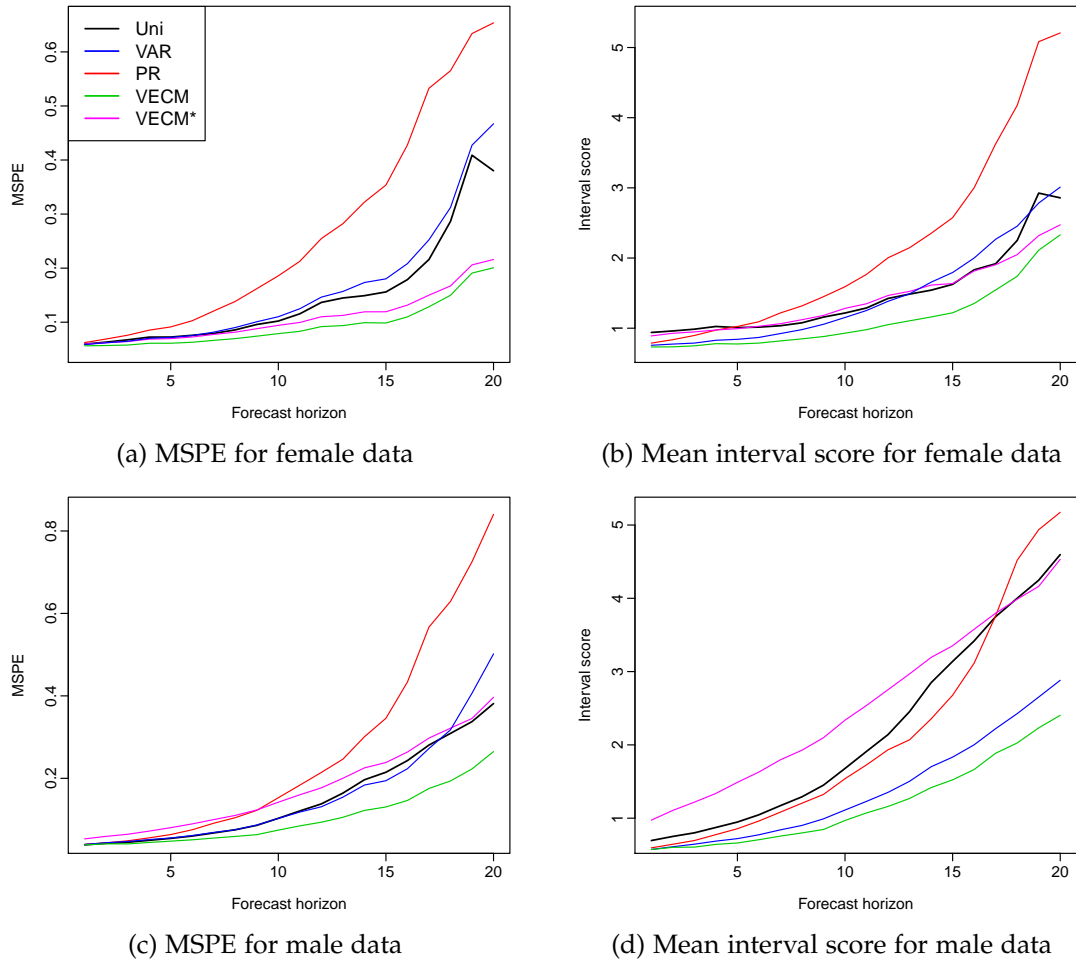
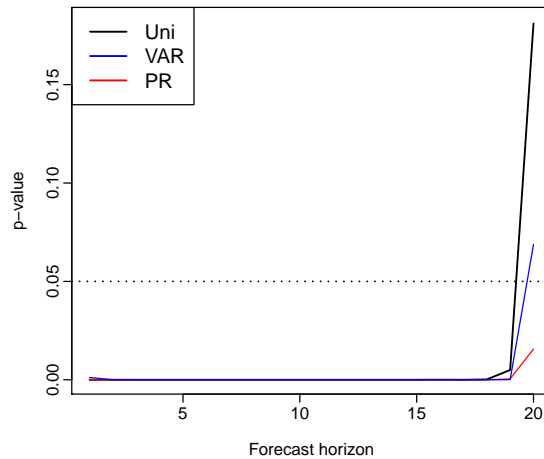
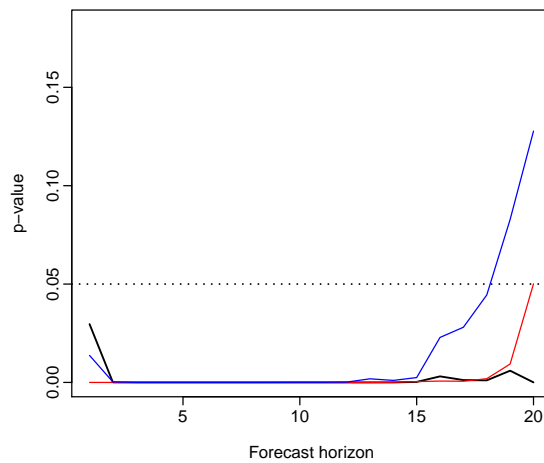


Figure 2.4: Czech Republic: forecast errors for female and male mortality rates (MSPE and Interval scores are presented)



(a) Female population



(b) Male population

Figure 2.5: Czech Republic: P -values for the three tests comparing a functional VECM to univariate, VAR and the product-ratio models respectively (the horizontal line is the default level of significance $\alpha = 0.05$)

2.6 Conclusions

We have proposed a functional VECM for the prediction of multivariate functional time series. Compared to the existing forecasting approaches, the proposed method has proved to perform well in both simulations and in empirical analyses. An algorithm to generate bootstrap prediction intervals is proposed and the results show a superiority in interval forecasts. The advantage of our method is the result of several factors: (1) the functional VECM model considered the covariance between different groups, rather than modeling the populations separately; (2) it could cope with data where the assumption of stationarity does not hold; (3) the forecast intervals using the proposed algorithm combine three sources of uncertainties. Bootstrapping is used to avoid the assumption of the distribution of the data.

We apply the proposed method as well as the existing methods to male and female mortality rates in Switzerland and the Czech Republic. The empirical studies provide evidence of the superiority of the functional VECM approach in both the point and interval forecasts, which are evaluated in MAPE, MSPE and interval scores respectively. Diebold-Mariano test results also show significantly improved forecast accuracy of our model. In most cases, when there is a long-run coherent structure in male and female mortality rates, the functional VECM is preferable. The long-term equilibrium constraint in the functional VECM ensures that divergence does not emerge.

While we use two populations for the illustration of the model and in the empirical analysis, the functional VECM can easily be applied to populations with more than two groups. A higher rank of co-integration order may need to be considered and the Johansen test can then be used to determine the rank (Johansen 1991).

Appendix: Functional principal component analysis

Let $\{\mathcal{X}_t(u), t \in Z\}$ be a set of functional time series in $L_2(\mathcal{I})$ from a separable Hilbert space \mathcal{H} . \mathcal{H} is characterized by the inner product $\langle \cdot, \cdot \rangle$, where $\langle \mathcal{X}_1, \mathcal{X}_2 \rangle = \int_{\mathcal{I}} \mathcal{X}_1(u) \mathcal{X}_2(u) dx$. We assume that $\mathcal{X}(u)$ has a continuous mean function $\mu(u)$ and covariance function $G(u, v)$:

$$\begin{aligned}\mu(u) &= \mathbb{E}[\mathcal{X}(u)], \\ G(u, v) &= \text{Cov}[\mathcal{X}(u), \mathcal{X}(v)] = \mathbb{E}\{[\mathcal{X}(u) - \mu(u)][\mathcal{X}(v) - \mu(v)]\},\end{aligned}$$

and thus the covariance operator for any $\mathcal{X}(u) \in \mathcal{H}$ is given by

$$C(v)(\mathcal{X}) = \int_{\mathcal{I}} G(v, u) \mathcal{X}(u) du.$$

The eigenequation $C(v)(\mathcal{X}) = \rho \mathcal{X}$ has solutions with orthonormal eigenfunctions $\phi_k(u)$, and associated eigenvalues λ_k for $k = 1, 2, \dots$ such that $\lambda_1 \geq \lambda_2 \geq \dots$ and $\sum_k \lambda_k < \infty$.

According to the Karhunen-Loève theorem, the function $\mathcal{X}(u)$ can be expanded by:

$$\mathcal{X}(u) = \mu(u) + \sum_{k=1}^{\infty} \tilde{\xi}_k \phi_k(u), \quad (2.8)$$

where $\{\phi_k(u)\}$ are orthogonal basis functions also on $L^2(\mathcal{I})$, and the principal component scores $\{\tilde{\xi}_k\}$ are uncorrelated random variables given by the projection of the centered function in the direction of the k^{th} eigenfunction:

$$\tilde{\xi}_k = \int_{\mathcal{I}} [\mathcal{X}(u) - \mu(u)] \phi_k(u) dx.$$

The principal component scores also satisfy:

$$E(\xi_k) = 0, \quad \text{Var}(\xi_k) = \lambda_k.$$

Functional principal component regression

According to (2.8), for a sequence of functional time series $\{\mathcal{X}_t(u)\}$, each element can be decomposed as:

$$\begin{aligned} \mathcal{X}_t(u) &= \mu(u) + \sum_{k=1}^{\infty} \xi_{t,k} \phi_k(u) \\ &= \mu(u) + \sum_{k=1}^K \xi_{t,k} \phi_k(u) + e_t(u), \end{aligned}$$

where $e_t(u)$ denotes the model truncation error function that captures the remaining terms. It is assumed that the scores follow $\xi_k \sim N(0, \lambda_k)$. Thus the functions can be characterized by the K -dimension vector $(\xi_1, \dots, \xi_K)^\top$.

Assorted approaches for selecting the number of principal components K include: a) ensuring that a certain fraction of the data variation is explained (Chiou 2012); b) cross-validation (Rice & Silverman 1991); c) bootstrapping (Hall & Vial 2006); d) information criteria (Yao et al. 2005a).

With the smoothed functions $\{\mathcal{X}_1(u), \dots, \mathcal{X}_n(u)\}$, the mean function $\mu(u)$ is estimated by

$$\hat{\mu}(u) = \frac{1}{n} \sum_{t=1}^n \mathcal{X}_t(u).$$

The covariance operator for a function g is estimated by

$$\widehat{C}(g) = \frac{1}{n} \sum_{t=1}^n \langle \mathcal{X}_t - \widehat{\mu}, g \rangle (\mathcal{X}_t - \widehat{\mu}),$$

where n is the number of observed curves. Sample eigenvalue and eigenfunction pairs $\widehat{\lambda}_k$ and $\widehat{\phi}_k(u)$ can be calculated from the estimated covariance operator using singular value decomposition. Empirical principal component scores $\zeta_{t,k}$ are obtained by $\zeta_{t,k} = \langle \mathcal{X}_t, \widehat{\phi}_k \rangle$ with numerical integration $\int_{\mathcal{I}} [\mathcal{X}_t(u) - \widehat{\mu}(x)] \widehat{\phi}_k(u) dx$. These simple estimators are proved to be consistent under weak dependence when the functions collected are dense and regularly spaced (Yao & Lee 2006, Hörmann & Kokoszka 2010). In sparse data settings, other methods should be applied. For instance, Yao et al. (2005a) propose principal component conditional expectation using pooled information between the functions to do estimation.

High-dimensional Functional Time Series Forecasting

3.1 Introduction

As the popularity of functional time series grows, there has been a rapid growing body of research on functional time series modeling and forecasting. From a parametric aspect, Bosq (2012) and Bosq & Blanke (2007) proposed the functional autoregressive of order 1 (FAR(1)) and derived one-step-ahead forecasts that are based on a regularized form of the Yule-Walker equations. Later, FAR(1) was extended to FAR(p), under which the order p can be determined via a sequential hypothesis testing procedure of Kokoszka & Reimherr (2013). Klepsch & Klüppelberg (2017) proposed the functional moving average (FMA) process and introduced an innovations algorithm to obtain the best linear predictor. Klepsch et al. (2017) extended the VAR model to the vector autoregressive moving average model. Recently, Li et al. (2020) considered long-range dependent curve time series and proposed a functional autoregressive fractionally integrated moving average model. From a nonparametric perspective, Besse et al. (2000) and Ferraty & Vieu (2006) proposed functional kernel regression to model the temporal dependence via a similarity measure defined by semi-metric, bandwidth, and kernel function. From a

semi-parametric viewpoint, Aneiros-Pérez & Vieu (2008) put forward a semi-functional partial linear model that combines both parametric and nonparametric models, and this model allows us to consider additive covariates and to use a continuous path in the past to predict future values of a stochastic process. Apart from the estimation of a conditional mean, Hörmann et al. (2013) considered a functional autoregressive conditional heteroskedasticity model for modeling conditional variance, while Aue et al. (2017) considered a functional generalized autoregressive conditional heteroskedasticity model. Kokoszka et al. (2017) considered a portmanteau test for testing autocorrelation under a functional analog of generalized autoregressive conditional heteroskedasticity model.

To deal with infinite dimensional functions, there is a demand for efficient data reduction techniques. Functional principal component analysis (FPCA) is the most commonly used approach that serves this purpose. FPCA performs eigendecomposition on the underlying variance functions. As in multivariate principal component analysis case, most of the variance structures are captured in a vector called the principal component scores. Some papers on FPCA include Hall et al. (2006) and Hall & Hosseini-Nasab (2006) on theoretical properties, Yao et al. (2005a) for sparse longitudinal data, and Locantore et al. (1999) and Viviani et al. (2005) for some interesting applications.

The existing FPCA method has been developed for independent observations, which is a serious weakness when we are dealing with functional time series. Thus we adopt a dynamic FPCA approach (Panaretos & Tavakoli 2013, Hörmann et al. 2015, Rice & Shang 2017), where serial dependence between the curves is taken into account. With dynamic FPCA, functional time series is reduced to a vector time series, where the individual component processes are mutually uncorrelated functional principal component (FPC) scores.

It is often the case that we collect a vector of N functions at a single time point t . If these N functions are assumed to be correlated, multivariate functional data models should be considered. Classical multivariate FPCA concatenates the multiple functions into one to perform univariate FPCA (see, e.g., Ramsay & Silverman 2005). Chiou et al. (2014) suggested normalizing each random function as a preliminary step before concatenation. Berrendero et al. (2011) studied a functional version of principal component analysis, where multivariate functional data are reduced to one or two functions rather than vectors. However, existing models dealing with multivariate functional data either fail to handle data with a large N or are difficult to implement practically.

The main contribution of this chapter is to propose a possible solution to modeling high-dimensional functional time series. By high dimension, we allow the dimension of the functional time series N to grow with the length of the observed functional time series T . We propose a twofold dimension reduction technique to represent the original multivariate functional time series with a low dimension scalar time series. The proposed model has three major advantages:

- 1) it models N functional time series simultaneously, taking the cross-covariance between the populations into account;
- 2) the model avoids the problem of curse of dimensionality, which is a major problem for traditional multivariate functional models;
- 3) the proposed model is conceptually simple and easy to implement.

Our model consists of three steps:

- 1) Dynamic FPCA is performed separately on each set of functional time series, resulting in N sets of principal component scores of low dimension p_0 (typically less than 5);

- 2) The first functional principal component scores from each of N sets of functional time series are combined into an $N \times 1$ vector. We fit factor models to the FPC scores to further reduce the dimension into an $r \times 1$ vector ($r \ll N$). The same is done for the second, third, and so on until the p_0^{th} FPC scores. The vector of N functional time series is reduced to $r \times p_0$ what we call factors.
- 3) A scalar time series model can be fitted to each factor and forecasts are produced. The forecast factors can be used to construct forecast functions.

The proposed dimension reduction model is essentially using a matrix of small dimension ($r \times p_0$) to represent the covariation of the original N functional time series. Elements of the reduced matrix are uncorrelated and it is adequate to model each element with scalar time series models.

In the second step mentioned above, we adopt factor models that are frequently used for dimension reduction for time series data. Some early application of factor analysis to multivariate time series include Anderson (1963), Priestley et al. (1974) and Brillinger (1981). Time series in high-dimensional settings where $N \rightarrow \infty$ together with T are studied in Chamberlain (1983), Bai (2003), and Lam et al. (2011). Among these, we adopt a similar approach to that considered in Lam et al. (2011), where the model is conceptually simple and the asymptotic properties are established. The reason why we use the technique is that the dimension reduction is based on the lag covariance of the data, which is suitable for time series data.

The remainder of the chapter is organized as follows. In Section 3.2, more detailed background on dynamic FPCA is introduced and the two-fold dimension reduction model is proposed. In Section 3.3, asymptotic results for the proposed model are given. We present simulation studies in Section 3.4. In Section 3.5, we apply our proposed model to Japanese age- and sex-specific mortality rate data. The conclusion is presented in

Section 3.6, and proofs are provided in the Appendix.

3.2 Twofold dimension reduction

We consider the stationary N -dimensional functional time series $\{\mathcal{X}_t : t = 1, \dots, T\}$, where T is the sample size. At time t , $\mathcal{X}_t = [\mathcal{X}_t^1(u), \dots, \mathcal{X}_t^N(u)]^\top$, and each $\mathcal{X}_t^{(i)}(u)$ takes values in the space $H := L^2(\mathcal{I})$ of real-valued square integrable functions on \mathcal{I} . The space H is a Hilbert space, equipped with the inner product $\langle x, y \rangle := \int_{\mathcal{I}} x(u)y(u)du$. The function norm is defined as $\|x\| := \langle x, x \rangle^{1/2}$. We could also look at the data in another direction, where we call $\{\mathcal{X}_t^{(i)}(u) : t = 1, \dots, T\}$ the i^{th} population of the functional time series, and there are N populations. Under our setting, both the sample size and the number of populations go to infinity, that is $N \rightarrow \infty, T \rightarrow \infty$.

The purpose is to reduce the dimension of the vector functional time series data \mathcal{X}_t . Our technique consists of performing FPCA on $\mathcal{X}_t^{(i)}$ for each population in the first step, resulting in $N \times p_0$ FPC scores, and then fitting factor models in the second step, getting $r \times p_0$ factors.

In Section 3.2.1 and 3.2.2, we will introduce the two models we use in our twofold dimension reduction. In Section 3.2.3, estimation process combining the two steps is explained. In Section 3.2.4, we illustrate how functional time series forecast can be performed.

3.2.1 Dynamic functional principal component analysis

For each $i \in \{1, \dots, N\}$, we assume that $\mathcal{X}_t^{(i)}$ has a continuous mean function $\mu^{(i)}(u)$ and an auto-covariance function at lag h , $c_h^{(i)}(u, v)$, where

$$\begin{aligned}\mu^{(i)}(u) &= \mathbb{E}[\mathcal{X}_t^{(i)}(u)], \\ c_h^{(i)}(u, v) &= \text{cov}[\mathcal{X}_t^{(i)}(u), \mathcal{X}_{t+h}^{(i)}(v)].\end{aligned}$$

The long-run covariance function is defined as

$$c^{(i)}(u, v) = \sum_{h=-\infty}^{\infty} c_h^{(i)}(u, v).$$

Using $c^{(i)}(u, v)$ as a kernel, we define the operator $C^{(i)}$ by:

$$C^{(i)}(x)(u) = \int_{\mathcal{I}} c^{(i)}(u, v)x(v)dv, \quad u, v \in \mathcal{I}.$$

For simplicity, we can also write

$$C^{(i)} = \sum_{h=-\infty}^{\infty} C_h^{(i)}, \tag{3.1}$$

where $C_h^{(i)}$ is the covariance operator at lag h . The operator is symmetric and non-negative definite. By Mercer's theorem, the operator $C^{(i)}$ admits an eigendecomposition

$$C^{(i)}(x) = \sum_{p=1}^{\infty} \lambda_p^{(i)} \langle x, \gamma_p^{(i)} \rangle \gamma_p^{(i)}, \tag{3.2}$$

where $(\lambda_p^{(i)} : p \geq 1)$ are the eigenvalues of $C^{(i)}$ in descending order and $(\gamma_p^{(i)} : p \geq 1)$ the corresponding normalized eigenfunctions. By Karhunen-Loève theorem, $\mathcal{X}_t^{(i)}(u)$ can be

represented with

$$\mathcal{X}_t^{(i)}(u) = \sum_{p=1}^{\infty} \beta_{p,t}^{(i)} \gamma_p^{(i)}(u),$$

where $\beta_{p,t}^{(i)} = \int_{\mathcal{I}} \mathcal{X}_t^{(i)}(u) \gamma_p^{(i)}(u) du$ is the p^{th} FPC score at time t . The infinite-dimensional functions can be approximated by the first p_0 FPC scores:

$$\mathcal{X}_t^{(i)}(u) = \sum_{p=1}^{p_0} \beta_{p,t}^{(i)} \gamma_p^{(i)}(u) + \theta_t^{(i)}(u), \quad (3.3)$$

where $\theta_t^{(i)}(u) = \sum_{p=p_0+1}^{\infty} \beta_{p,t}^{(i)} \gamma_p^{(i)}(u)$ captures the remaining terms cutting from $p = p_0 + 1$ to ∞ .

3.2.2 Factor model

With the first step dimension reduction, we now have FPC scores $\beta_{p,t}^{(i)}$ where $i = 1, \dots, N$. Define the vector FPC score as

$$\boldsymbol{\beta}_{p,t} = \left(\beta_{p,t}^1, \dots, \beta_{p,t}^N \right)^\top. \quad (3.4)$$

So $\boldsymbol{\beta}_{p,t}$ is the vector that contains the p^{th} FPC score of all N functional time series. We consider the following factor model for each $p = 1, \dots, p_0$. Let

$$\boldsymbol{\beta}_{p,t} = \mathbf{A}_p \mathbf{f}_{p,t} + \mathbf{e}_{p,t}, \quad t = 1, \dots, T, \quad (3.5)$$

where $\mathbf{f}_{p,t}$ is an $r \times 1$ unobserved factor time series; \mathbf{A}_p is an $N \times r$ unknown constant factor loading matrix. We need to fit p_0 factor models to the FPC scores. The factor model is similar to the model in Lam et al. (2011). The difference is that in their paper, the e_t 's

are assumed to be white noise with mean zero and a constant covariance matrix. In our settings, there is no model on the error term, and $\mathbf{e}_{p,t}$ is what is left after taking out the main explaining factors $\mathbf{f}_{p,t}$. To write out $\mathbf{e}_{p,t}$:

$$\mathbf{e}_{p,t} = \mathbf{A}'_{p,t} \mathbf{f}'_{p,t},$$

where \mathbf{A}'_p is an $N \times (N - r)$ matrix, the columns of which are orthogonal to the columns of \mathbf{A}_p ; and $\mathbf{f}'_{p,t}$ is $(N - r) \times 1$ vector.

Combing (3.3) and (3.5), the original functional time series can be modeled as

$$\mathcal{X}_t^{(i)}(u) = \sum_{p=1}^{p_0} [\mathbf{A}_p \mathbf{f}_{p,t}]^{(i)} \gamma_p^{(i)}(u) + \epsilon_t^{(i)}(u), \quad (3.6)$$

where $[\cdot]^{(i)}$ denotes the i^{th} element in the vector. Note that the i^{th} element in the vector $\mathbf{A}_p \mathbf{f}_{p,t}$ is in fact $\boldsymbol{\alpha}^{(i)} \mathbf{f}_{p,t}$, where $\boldsymbol{\alpha}^{(i)}$ is the i^{th} row in the matrix \mathbf{A}_p . The resulting dimension reduced factor $\mathbf{f}_{p,t}$ does not rely on i . The error term $\epsilon_t^{(i)}(u)$ contains the accumulated error from both steps:

$$\epsilon_t^{(i)}(u) = \theta_t^{(i)}(u) + \sum_{p=1}^{p_0} [\mathbf{e}_{p,t}]^{(i)} \gamma_p(u).$$

Before the estimation process is introduced, we make a few notations and definitions.

Define

$$\boldsymbol{\Sigma}_{\beta}^{(p)}(h) = \text{cov}(\boldsymbol{\beta}_{p,t+h}, \boldsymbol{\beta}_{p,t}), \quad \boldsymbol{\Sigma}_f^{(p)}(h) = \text{cov}(\mathbf{f}_{p,t+h}, \mathbf{f}_{p,t}), \quad \boldsymbol{\Sigma}_e^{(p)}(h) = \text{cov}(\mathbf{e}_{p,t+h}, \mathbf{e}_{p,t}),$$

and also the cross-covariance between the factor and the error term

$$\boldsymbol{\Sigma}_{f,e}^{(p)}(h) = \text{cov}(\mathbf{f}_{p,t+h}, \mathbf{e}_{p,t}), \quad \boldsymbol{\Sigma}_{f,e}^{(p)}(-h) = \text{cov}(\mathbf{e}_{p,t+h}, \mathbf{f}_{p,t}).$$

Using (3.5), we can write the relation as:

$$\boldsymbol{\Sigma}_{\beta}^{(p)}(h) = \mathbf{A}_p \boldsymbol{\Sigma}_f^{(p)}(h) \mathbf{A}_p^{\top} + \mathbf{A}_p \boldsymbol{\Sigma}_{f,e}^{(p)}(h) + \mathbf{A}_p \boldsymbol{\Sigma}_{f,e}^{(p)}(-h) + \boldsymbol{\Sigma}_e^{(p)}(h). \quad (3.7)$$

Let

$$\mathbf{L}^{(p)} = \sum_{h=1}^{h_0} \boldsymbol{\Sigma}_{\beta}^{(p)}(h) \boldsymbol{\Sigma}_{\beta}^{(p)}(h)^{\top}, \quad (3.8)$$

where h_0 is a constant.

Plugging (3.7) into (3.8), we have

$$\mathbf{L}^{(p)} = \mathbf{L}^{(p)*} + \mathbf{E}^{(p)}, \quad (3.9)$$

where

$$\mathbf{L}^{(p)*} = \mathbf{A}_p \left[\sum_{h=1}^{h_0} \{ \boldsymbol{\Sigma}_f^{(p)}(h) \mathbf{A}_p^{\top} + \boldsymbol{\Sigma}_{f,e}^{(p)}(h) + \boldsymbol{\Sigma}_{f,e}^{(p)}(-h) \} \{ \boldsymbol{\Sigma}_f^{(p)}(h) \mathbf{A}_p^{\top} + \boldsymbol{\Sigma}_{f,e}^{(p)}(h) + \boldsymbol{\Sigma}_{f,e}^{(p)}(-h) \}^{\top} \right] \mathbf{A}_p^{\top},$$

and

$$\begin{aligned} \mathbf{E}^{(p)} = & \mathbf{A}_p \left[\boldsymbol{\Sigma}_f^{(p)}(h) \mathbf{A}_p^{\top} + \boldsymbol{\Sigma}_{f,e}^{(p)}(h) + \boldsymbol{\Sigma}_{f,e}^{(p)}(-h) \right] \boldsymbol{\Sigma}_e^{(p)\top}(h) \\ & + \boldsymbol{\Sigma}_e^{(p)}(h) \left[\boldsymbol{\Sigma}_f^{(p)}(h) \mathbf{A}_p^{\top} + \boldsymbol{\Sigma}_{f,e}^{(p)}(h) + \boldsymbol{\Sigma}_{f,e}^{(p)}(-h) \right]^{\top} \mathbf{A}_p^{\top} + \boldsymbol{\Sigma}_e^{(p)}(h) \boldsymbol{\Sigma}_e^{(p)\top}(h). \end{aligned} \quad (3.10)$$

If we perform eigendecomposition on the middle part within the square brackets of $\mathbf{L}^{(p)*}$,

then $L^{(p)*} = A_p \mathbf{U}_p D_p \mathbf{U}_p^\top A_p^\top$, where D_p is the diagonal matrix with the first r largest eigenvalues. \mathbf{U}_p is an orthogonal matrix, so that $A_p \mathbf{U}_p$ is a rotation on the matrix A_p . We use $A_p \mathbf{U}_p$ as the matrix A_p . Thus, $L^{(p)*} = A_p D_p A_p^\top$. Let the columns of A_p be the eigenvectors of the matrix $L^{(p)*}$ corresponding to the first r largest eigenvalues in descending order. The matrix D_p is then a diagonal matrix with the first r eigenvalues on its diagonal.

3.2.3 Estimation

We need to estimate A_p , $f_{p,t}$ and $\gamma_p^{(i)}(u)$ in (3.6). In the dynamic FPCA step, the long-run covariance function $c^{(i)}(u, v)$ can be estimated by:

$$\hat{c}^{(i)}(u, v) = \sum_{|h| \leq q} W\left(\frac{h}{q}\right) \hat{c}_h^{(i)}(u, v),$$

and the covariance operator by:

$$\hat{C}^{(i)}(x)(u) = \int_{\mathcal{I}} \hat{c}^{(i)}(u, v) x(v) dv,$$

or

$$\hat{C}^{(i)} = \sum_{|h| \leq q} W\left(\frac{h}{q}\right) \hat{C}_h^{(i)}, \quad (3.11)$$

where

$$\hat{c}_h^{(i)}(u, v) = \begin{cases} \frac{1}{T-h} \sum_{j=1}^{T-h} [\mathcal{X}_j^{(i)}(u) - \bar{\mathcal{X}}_j(u)] [\mathcal{X}_{j+h}(v) - \bar{\mathcal{X}}(v)], & h \geq 0 \\ \frac{1}{T-h} \sum_{j=1-h}^T [\mathcal{X}_j^{(i)}(u) - \bar{\mathcal{X}}_j(u)] [\mathcal{X}_{j+h}(v) - \bar{\mathcal{X}}(v)], & h < 0 \end{cases}.$$

Here, $W(\cdot)$ is a weight function with $W(0) = 1, W(u) = W(-u), W(u) = 0$ if $|u| > m$ for some $m > 0$, and W is continuous on $[-m, m]$. Some possible choices include Bartlett, Parzen, Tukey-Hanning, quadratic spectral and flat-top functions (Andrews 1991, Andrews & Monahan 1992). In this chapter, we use $W(h/q) = 1 - |h|/q$, where q is a bandwidth parameter. Conditions will be imposed on q in Section 3.3.

By performing eigendecomposition on $\widehat{C}^{(i)}$, we can estimate empirical eigenfunctions $\widehat{\gamma}_p^{(i)}(u)$ and the empirical FPC scores $\widetilde{\beta}_{p,t}^{(i)} = \int_{\mathcal{I}} \mathcal{X}_t^{(i)}(u) \widehat{\gamma}_p^{(i)}(u) du$, calculated by numerical integration.

The estimates $\widetilde{\beta}_{p,t}^{(i)}$ are combined into a vector

$$\widetilde{\beta}_{p,t} = \left(\widetilde{\beta}_{p,t}^1, \dots, \widetilde{\beta}_{p,t}^N \right), \quad (3.12)$$

and fitted to a factor model. The estimation of latent factors for high-dimensional time series can be found in Lam et al. (2011). The idea of estimation is that in the previously defined matrix $\mathbf{L}^{(p)} = \mathbf{L}^{(p)*} + \mathbf{E}^{(p)}$ in (3.9), when the term $\mathbf{E}^{(p)}$ related to error covariance is small, such that $\mathbf{L}^{(p)*}$ is close to $\mathbf{L}^{(p)}$, we can use the eigendecomposition of $\mathbf{L}^{(p)}$ to estimate the eigendecomposition of $\mathbf{L}^{(p)*}$. Details can be found in Appendix.

Then, a natural estimator for \mathbf{A}_p can be found by performing eigendecomposition on an estimated version of $\mathbf{L}^{(p)}$. It is defined as $\widehat{\mathbf{A}}_p = (\widehat{\mathbf{a}}_{p,1}, \dots, \widehat{\mathbf{a}}_{p,r})$, where $\widehat{\mathbf{a}}_{p,j}$ is the j^{th} eigenvector of $\widehat{\mathbf{L}}^{(p)}$, and

$$\widehat{\mathbf{L}}^{(p)} = \sum_{h=1}^{h_0} \widehat{\Sigma}_{\beta}^{(p)}(h) \widehat{\Sigma}_{\beta}^{(p)}(h)^{\top}, \quad \widehat{\Sigma}_{\beta}^{(p)}(h) = \frac{1}{T-h} \sum_{t=1}^{T-h} \left(\widetilde{\beta}_{p,t+h} - \widetilde{\beta}_p \right) \left(\widetilde{\beta}_{p,t} - \widetilde{\beta}_p \right)^{\top}, \quad (3.13)$$

where $\widetilde{\beta}_p = 1/T \sum_{t=1}^T \widetilde{\beta}_{p,t}$. Thus we estimate the p^{th} factor by:

$$\widehat{\mathbf{f}}_{p,t} = \widehat{\mathbf{A}}_p^{\top} \widetilde{\beta}_{p,t}. \quad (3.14)$$

The estimated dimension reduced FPC scores are

$$\widehat{\boldsymbol{\beta}}_{p,t} = \widehat{\mathbf{A}}_p \widehat{\mathbf{f}}_{p,t}.$$

The estimator for the original function $\mathcal{X}_t^{(i)}(u)$ is:

$$\widehat{\mathcal{X}}_t^{(i)}(u) = \sum_{p=1}^{p_0} [\widehat{\boldsymbol{\beta}}_{p,t}]^{(i)} \widehat{\gamma}_p^{(i)}(u) = \sum_{p=1}^{p_0} [\widehat{\mathbf{A}}_p \widehat{\mathbf{f}}_{p,t}]^{(i)} \widehat{\gamma}_p^{(i)}(u), \quad i = 1, \dots, N, \quad t = 1, \dots, T, \quad (3.15)$$

where again $[\cdot]^{(i)}$ denotes the i^{th} element of the vector.

3.2.4 Forecasting

With twofold dimension reduction, information of serial correlation is contained in the factors $\mathbf{f}_{p,t}$. To forecast N -dimensional functional time series, we could instead make forecast on the estimated factors. Scalar or vector time series models could be applied. We suggest univariate time series models: autoregressive moving average (ARMA) models, for instance, since the factors are mutually uncorrelated. Recall that we have retained p_0 FPC scores and r factors in the estimation process. Consequently, we need to fit $r \times p_0$ ARMA models on the estimated factors $\{\widehat{\mathbf{f}}_{p,1}, \dots, \widehat{\mathbf{f}}_{p,t}\}$. The prediction of the functions could be calculated as:

$$\widehat{\mathcal{X}}_{T+h|T}^{(i)}(u) = \sum_{p=1}^{p_0} [\widehat{\mathbf{A}}_p \widehat{\mathbf{f}}_{p,T+h|T}]^{(i)} \widehat{\gamma}_p^{(i)}(u), \quad i = 1, \dots, N, \quad (3.16)$$

where $\widehat{\mathcal{X}}_{t+h|t}^{(i)}(u)$ is the h -step-ahead forecast at time t , and h denotes a forecast horizon.

Prediction intervals for functions could also be constructed. In this chapter, we use a

bootstrapping approach. The bootstrapped function forecast is

$$\widehat{\mathcal{X}}_{T+h|T}^{(i),b}(u) = \sum_{p=1}^{p_0} \left[\widehat{\mathbf{A}}_p \widehat{\mathbf{f}}_{p,T+h|T}^b \right]^{(i)} \widehat{\gamma}_p^{(i)}(u) + \widehat{\varepsilon}_{T+h|T}^{(i),b}(u), \quad b = 1, \dots, B,$$

where B is the number of bootstraps, and $\widehat{\mathbf{f}}_{p,T+h|T}^b$ is the bootstrapped prediction of the factor. A bootstrapped residual $\widehat{\varepsilon}_{T+h|T}^{(i),b}(u)$ is added, which is resampled from $\{\mathcal{X}_t^{(i)}(u) - \widehat{\mathcal{X}}_t^{(i)}(u)\}$. Lower and Upper prediction bounds are calculated as the $100 \times (\alpha/2)^{\text{th}}$ and $100 \times (1 - \alpha/2)^{\text{th}}$ percentile of the bootstrapped forecasts, where α is the level of significance.

The bootstrapped prediction of the factors $\widehat{\mathbf{f}}_{p,T+h|T}^b$ can be constructed in several ways. Here in this chapter, we use an intuitive approach consisting of four steps:

1. Fit an ARMA(p, q) model to the first half of the estimated factors $\widehat{f}_{p,1}, \dots, \widehat{f}_{p,T_1}$, $T_1 = T/2$, and make h -step ahead prediction \tilde{f}_{p,T_1+h} . The prediction error is $\zeta_{p,1} = \widehat{f}_{p,T_1+h} - \tilde{f}_{p,T_1+h}$.
2. Fit another ARMA(p, q) model to $\widehat{f}_{p,1}, \dots, \widehat{f}_{p,T_1+1}$, and make h -step ahead prediction \tilde{f}_{p,T_1+1+h} . The prediction error is $\zeta_{p,2} = \widehat{f}_{p,T_1+1+h} - \tilde{f}_{p,T_1+1+h}$.
3. Following the first two steps we acquire j prediction errors such that $T_1 + j + h = T$. Resample ζ_p^* from $\{\zeta_{p,1}, \dots, \zeta_{p,j}\}$.
4. Construct the bootstrap prediction $\widehat{f}_{p,T+h|T}^b = \widehat{f}_{p,T+h|T} + \zeta_p^*$, where $\widehat{f}_{p,T+h|T}$ is the point forecast.

The process above needs to be repeated r times to generate each element in $\widehat{\mathbf{f}}_{p,T+h|T}^b$.

Alternatively, the bootstrapped functions can be used to construct prediction region (Zhu & Politis 2017). Denote $q(\alpha)$ as the α -quantile of $\|\widehat{\mathcal{X}}_{T+h|T}^{i,b} - \widehat{\mathcal{X}}_{T+h|T}^i\|_2$. Then the

$(1 - \alpha) \times 100\%$ bootstrap predictive region consists of all \mathcal{X} such that

$$\left\| \mathcal{X} - \widehat{\mathcal{X}}_{T+h|T}^{(i)} \right\|_2 \leq q(\alpha).$$

3.3 Asymptotic properties

We derive consistency results for dimension reduced estimates of the functions. Lemmas and detailed proofs are provided in the Appendix. First, some assumptions are made on the functional time series and the FPCA estimation process.

Assumption 1. For each $i = 1, \dots, N$, the functions $\{\mathcal{X}_t^{(i)}(u), t \in \mathbb{Z}\}$ are stationary and L^4 - m -approximable, which also satisfies

$$\sup_t \|\mathcal{X}_t^{(i)}(u)\| < \infty$$

Assumption 2. For each $i = 1, \dots, N$, the h lag covariance operator satisfies

$$\sum_{h=-\infty}^{\infty} \|C_h^{(i)}\|_{\mathcal{S}} < \infty,$$

where $\|\cdot\|_{\mathcal{S}}$ denotes the Hilbert-Schmidt norm.

In Assumption 1, it is assumed the dependence structure of the functional time series for each population. The definition of L^p - m -approximable in Assumption 1 and the Hilbert-Schmidt norm in Assumption 2 can be found in the Appendix. We also provide a simple sufficient condition for Assumption 2 to hold, as in the following proposition.

Proposition 1. If \mathcal{X}_t is L^2 - m -approximate, then Assumption 2 holds.

An example of functional time series that satisfies this assumption is the simplest

functional AR(1) model. The proof of the proposition and that functional AR(1) is L^2 -m-approximate is included in the Appendix.

Assumption 3. For each $i = 1, \dots, N$, the eigenvalues $\lambda_p^{(i)}, p = 1, \dots$ are distinct.

Assumption 4. In the estimation of the functional principal components, for each population i , the empirical eigenfunctions are in the same direction of the true function, i.e., $\langle \gamma_p^{(i)}, \widehat{\gamma}_p^{(i)} \rangle > 0$

Assumption 5. In the estimation of the functional principal components, for each population i , the bandwidth parameter $q^3 = o(T/N)$, and $q \rightarrow \infty$.

Assumption 3 is a very common assumption in FPCA. Further, Assumption 4 ensures that we choose the correct sign for each eigenfunction. This assumption is used to serve for theoretical proof. In practice, the sign of the estimated eigenfunction does not make a difference because the problem vanishes once we take the product of the estimated eigenfunction and the corresponding estimated FPC score. Assumption 5 imposes a condition on the rate of the bandwidth parameter q , which has been previously defined in (3.11). The two conditions in Assumption 5 also imply $N = o(T)$, that is the number of populations grows not as rapidly as the sample size.

The following assumptions are made on the second dimension reduction step, the factor model as stated in (3.5). As in Section 3.2.2, in the following, we again omit the subscript p for conciseness. First, let's define some notations. We use $\|M\|_2$ to denote the L_2 norm of the matrix or vector. When M is a matrix, it is the greatest singular value. We use $\|M\|_{\min}$ to denote the smallest singular value. We use $a \asymp b$ to denote $\{a = O(b)\} \cap \{b = O(a)\}$, that is a and b are of the same order.

Assumption 6. For p in 1 to p_0 , $\|\Sigma_f^{(p)}(h)\|_2 \asymp N^{1-\delta} \asymp \|\Sigma_f^{(p)}(h)\|_{\min}$, where $0 \leq \delta < 1$.

Assumption 7. For p in 1 to p_0 , $\|\Sigma_{e,f}^{(p)}(h)\|_2 = O(\|\Sigma_f^{(p)}(h)\|_2)$,
and $\|\Sigma_e^{(p)}(h)\|_2 = O(\min\{NT^{-1/2}, 1\})$.

In Assumption 6, it is assumed that the order of the lag covariance of factor $f_{p,t}$ is related to the dimension of $\beta_{p,t}$ by a factor $\delta \in [0, 1)$. In Assumption 7, it is assumed that the strength of the lag cross-covariance between factors and errors is not bigger than that of the lag covariance of the factors, and that the lag covariance of the error term is at least bounded or of constant rate. Since what the model does essentially is principal component analysis, we want to ensure that most of the covariation of $\beta_{p,t}$ is contained in the lower dimension factors $f_{p,t}$.

Assumption 8.

$$\|\theta_t^{(i)}(u)\|_2 = o_P(1), \quad N \rightarrow \infty$$

where $\theta_t^{(i)}(u)$ is defined in (3.3).

Assumption 9. For each i ,

$$[e_{p,t}]^{(i)} = O_P\left(\frac{1}{\sqrt{N}}\right), \quad N, T \rightarrow \infty,$$

where $e_{p,t}$ is defined in (3.5), and $[e_{p,t}]^{(i)}$ denotes the i th element in the vector $e_{p,t}$.

In Assumption 8 and 9, it is assumed the error terms in both dimension reduction steps to be small, which is a natural assumption in principal component analysis.

Theorem 1. Under Assumptions 1 to 5, $\|\tilde{\beta}_{p,t} - \beta_{p,t}\|_2$ converges to zero in probability as $N, T \rightarrow \infty$, where the vectors $\beta_{p,t}$ and $\tilde{\beta}_{p,t}$ are defined in (3.4) and (3.12).

Theorem 2. Under Assumptions 1 to 7, assuming $N^\delta T^{-1/2} = o(1)$, we have

$$\|\hat{\mathbf{A}}_p - \mathbf{A}_p\|_2 = O_P\left(N^\delta T^{-1/2} + N^{\delta-1}\right) = o_P(1)$$

In Theorem 2, we have proved the convergence rate of the estimated factor loadings. When δ is 0, the convergence rate becomes $(T^{-1/2} + N^{-1})$, which is quite fast. However when δ is close to 1, the rate of convergence is very slow.

Theorem 3. *Under Assumptions 1 to 7,*

$$\frac{1}{N} \left\| \widehat{\boldsymbol{\beta}}_{p,t} - \boldsymbol{\beta}_{p,t} \right\|_2 = O_P \left(N^{(\delta-1)/2} T^{-1/2} \right) + O_P \left(\frac{1}{N} \right).$$

Theorem 4. *Under Assumptions 1 to 9, assuming $N^{\delta/2} T^{-1/2} = o(1)$,*

$$\frac{1}{N} \sum_{i=1}^N \left\| \widehat{\mathcal{X}}_t^{(i)}(u) - \mathcal{X}_t^{(i)}(u) \right\| = o_P(1), \quad N, T \rightarrow \infty$$

Theorem 3 states the convergence rate for the estimated FPC scores $\boldsymbol{\beta}_{p,t}$. Theorem 4 proves that the approximated functions are good estimates of the true functions. The rate of convergence is calculated in the Appendix.

We also investigate the prediction error of the model. After dimension reduction, classic time series models are fitted to the estimated factors $\widehat{\boldsymbol{f}}_{p,t}$. In this chapter, we use the AR(1) model as an example in the proof for asymptotic property. Let $f_{p,t}^{(i)}$ denote the i th element in the vector $\boldsymbol{f}_{p,t}$. The AR(1) model is

$$f_{p,t}^{(i)} = \phi_{p,i} f_{p,t-1}^{(i)} + \omega_{p,t}^{(i)}, \quad t = 2, \dots,$$

where $\phi_{p,i}$ is the AR coefficient which satisfy $|\phi_{p,i}| < 1$, and $\omega_{p,t}^{(i)}$ is the white noise. We define $\Gamma = \max_{p,i}(\Gamma_{p,i})$, where $\Gamma_{p,i} = \left| \sum_{j=1}^{h-1} \phi_{p,i}^j \omega_{p,T+h-j}^{(i)} \right|$. We have the following theorem.

Theorem 5. *Under Assumptions 1 to 9, assuming $N^{\delta/2} T^{-1/2} = o(1)$,*

$$\frac{1}{N} \sum_{i=1}^N \left\| \widehat{\mathcal{X}}_{T+h|T}^{(i)}(u) - \mathcal{X}_{T+h}^{(i)}(u) \right\| = o_P(1) + O_P(N^{-1/2}\Gamma), \quad N, T \rightarrow \infty,$$

We see that the forecast error includes a component that converges to zero which comes from the estimation error, and a component that is $O_p(N^{-1/2}\Gamma)$ which measures the error from the forecast model. In the ordinary setting of univariate AR(1) model, $\Gamma = O(1)$. So the second part also converges, that is $N^{-1/2}\Gamma = o_p(1)$.

For the simplicity of theoretical proofs, we use AR(1) model as an example. However, with easy but tedious work, this theorem can be extended to all stationary ARMA models. We could find an appropriate Γ that is $O(1)$ so that the error from the forecast model is $o_p(1)$

3.4 Simulation Studies

We illustrate our method using simulated data. We compare results using the proposed high-dimensional functional time series (HDFTS) model and a univariate functional time series (FTS) model where each population is modeled and predicted as a independent functional time series.

3.4.1 Data generation

We generate N populations of functional time series data. The i^{th} function at time t is constructed by

$$\mathcal{X}_t^{(i)}(u) = \sum_{p=1}^2 \beta_{p,t}^{(i)} \gamma_p^{(i)}(u) + \theta_t^{(i)}(u), \quad t = 1, \dots, T, \quad i = 1, \dots, N$$

where $\theta_t^{(i)}(u) = \sum_{p=3}^{\infty} \beta_{p,t}^{(i)} \gamma_p^{(i)}(u)$.

The coefficients $\beta_{p,t}^{(i)}$ for all N populations are combined and generated by

$$\beta_{p,t} = A_p f_{p,t}, \quad p = 1, \dots, \infty,$$

where $\boldsymbol{\beta}_{p,t} = \{\beta_{p,t}^1, \dots, \beta_{p,t}^N\}$. \mathbf{A}_p is a $N \times N$ matrix, and $\mathbf{f}_{p,t}$ is a $N \times 1$ vector.

We assume that $\beta_{p,t}^{(i)}$ have mean 0 and variance 0 when $p > 3$, so we only construct the coefficients $\boldsymbol{\beta}_{p,t}$ for $p = 1, 2, 3$.

The first set of coefficients $\boldsymbol{\beta}_{1,t}$ for N populations are generated with $\boldsymbol{\beta}_{1,t} = \mathbf{A}_1 \mathbf{f}_{1,t}$. Each element in matrix \mathbf{A}_1 is generated by $a_{ij} = N^{-1/4} \times b_{ij}$, where $b_{ij} \sim \mathcal{N}(2, 4)$.

The factors $\mathbf{f}_{1,t}$ are generated using autoregressive model of order 1 (AR(1)). Define the i^{th} element in vector $\mathbf{f}_{1,t}$ as $f_{1,t}^{(i)}$. Then, $f_{1,t}^1$ is generated by $f_{1,t}^1 = 0.5f_{1,t-1}^1 + \omega_t$, where ω_t are independent $\mathcal{N}(0, 1)$ random variables. We generate $f_{1,t}^{(i)}$, $i = 2, \dots, N$ by $f_{1,t}^{(i)} = (1/N)g_t^{(i)}$, where $g_t^{(i)}$, $i = 2, \dots, N$ are also AR(1) and follow $g_t^{(i)} = 0.2g_{t-1}^{(i)} + \omega_t$. It is ensured that most of the variance of $\boldsymbol{\beta}_{1,t}$ can be explained by one factor. The second coefficient $\boldsymbol{\beta}_{2,t}$ are constructed the same way as $\boldsymbol{\beta}_{1,t}$.

We also generate the third FPC scores $\boldsymbol{\beta}_{3,t}$ but with small values. \mathbf{A}_3 is generated by $a_{ij} = N^{-1/4} \times b_{ij}$, where $b_{ij} \sim \mathcal{N}(0, 0.04)$. The factors $\mathbf{f}_{3,t}$ are generated as $\mathbf{f}_{1,t}$.

The three basis functions are constructed by $\gamma_1^{(i)}(u) = \sin(2\pi u + \pi i/2)$, $\gamma_2^{(i)}(u) = \cos(2\pi u + \pi i/2)$ and $\gamma_3^{(i)}(u) = \sin(4\pi u + \pi i/2)$, where $u \in [0, 1]$. The functional time series for the i^{th} population is constructed by

$$\mathcal{X}_t^{(i)}(u) = [\boldsymbol{\beta}_{1,t}]_i \gamma_1^{(i)}(u) + [\boldsymbol{\beta}_{2,t}]_i \gamma_2^{(i)}(u) + [\boldsymbol{\beta}_{3,t}]_i \gamma_3^{(i)}(u),$$

where $[\cdot]_i$ denotes the i^{th} element of the vector.

3.4.2 Model fitting

Simulated data are generated under different settings of N and T values. The proposed model is fitted to the data. The bandwidth parameter is simply chosen as \sqrt{T} in each case. We use fraction of variation explained (FVE) to choose both the number of FPC score \hat{p}_0 and the number of retained factors \hat{r} . We require the first \hat{p}_0 FPC scores to explain 99%

of each population of functional time series, and the first \hat{r} factors to explain also 99% of each FPC scores. For different populations, the chosen number of FPC scores can be different according to FVE. Our model requires to select the same number of FPC scores for each population. Therefore, we choose \hat{p}_0 to be the largest number of FPC scores needed. The number of retained factors \hat{r} in the factor models can also be different for each $p = 1, \dots, \hat{p}_0$. Therefore, we use $\hat{r}_1, \dots, \hat{r}_{\hat{p}_0}$ to denote the number of factors chosen for each FPC score.

We first look at the estimation error under different settings. The estimation error is calculated using mean norm of residuals (MNR):

$$MNR = \frac{1}{NT} \sum_{i=1}^N \sum_{t=1}^T \sqrt{\sum_{j=1}^w [\mathcal{X}_t^{(i)}(u_j) - \hat{\mathcal{X}}_t^{(i)}(u_j)]^2},$$

where $\mathcal{X}_t^{(i)}(u_j)$ denotes the function value at discrete time point u_j , for $j = 1 \dots, w$, and w is the total number of discrete points in $[0, 1]$. We use $w = 51$ throughout simulation study.

Next we compare the forecast performance of our model with the independent forecast model. The independent forecast model follows similar idea of the approach in Hyndman & Ullah (2007) for each population. When fitting models to the data, we use expanding window prediction. The simulated data is divided into a training set with size T_1 and a test set with size T_2 , and $T_1 + T_2 = T$. In this study, we use $T_2 = (1/4) \times T$. The proposed models are fitted to the training set and forecasts are made based on fitted models. Then the test set is used for forecast evaluation. Each time we increase the training size by one and refit the model. New forecasts are made each time. Finally all the prediction errors for one-, two-, and three-step-ahead forecasts are collected and means are taken.

We use discretized mean absolute forecast error (MAFE) and mean squared forecast

error (MSFE).

$$\text{MAFE}(h) = \frac{1}{(T_2 + 1 - h) \times w} \sum_{\eta=h}^{T_2} \sum_{j=1}^w \left| \mathcal{X}_{T_1+\eta}(u_j) - \hat{\mathcal{X}}_{T_1+\eta|T_1+\eta-h}(u_j) \right|,$$

$$\text{MSFE}(h) = \frac{1}{(T_2 + 1 - h) \times w} \sum_{\eta=h}^{T_2} \sum_{j=1}^w \left[\mathcal{X}_{T_1+\eta}(u_j) - \hat{\mathcal{X}}_{T_1+\eta|T_1+\eta-h}(u_j) \right]^2,$$

where $\hat{\mathcal{X}}_{T_1+\eta|T_1+\eta-h}$ represents the h -step-ahead prediction using data $t = 1, \dots, T_1 + \eta - h$ fitted in the model, and $\mathcal{X}_{T_1+\eta}(u_j)$ denotes the holdout function.

For each combination of N and T values, we replicate the simulation 100 times, and calculate the mean of the errors MNR, MAFE and MSFE.

3.4.3 Results

The estimation errors of the proposed model are presented in Table 3.1. With the increase of N and T , the MNR becomes smaller, which can be seen as a concordant with Theorem 4. Under all four settings, we select the number of FPC scores to be two. The number of factors selected for the first and second FPC scores \hat{r}_1 and \hat{r}_2 are different in each simulation, but mostly equal to two or three.

Table 3.1: The MNR under different settings

(N, T)	MNR	\hat{p}_0
(20, 20)	1.756	2
(40, 50)	1.226	2
(60, 80)	0.804	2
(100, 150)	0.622	2

Table 3.2 shows the sample mean of the MAFE and MSFE in different settings. Each number in the table is the mean of the h -step-ahead errors taken over all N populations. The smaller value is in bold face. It can be seen that the proposed model produces smaller

forecast errors in almost all settings and all forecast horizons.

Table 3.2: The mean MAFE and MSFE values when fitting the independent functional time series model and the proposed high-dimensional functional time series model two models for one-, two-, and three-step-ahead forecasts

(N, T)	h	MAFE		MSFE	
		FTS	HDFTS	FTS	HDFTS
(20, 20)	1	1.136	1.134	2.595	2.597
	2	1.310	1.266	3.539	3.256
	3	1.380	1.324	3.983	3.630
(40, 50)	1	0.917	0.872	1.688	1.532
	2	1.038	0.970	2.187	1.908
	3	1.096	0.998	2.438	2.005
(60, 80)	1	0.817	0.763	1.327	1.163
	2	0.935	0.858	1.752	1.473
	3	0.980	0.877	1.942	1.540
(100, 150)	1	0.688	0.644	0.992	0.863
	2	0.799	0.737	1.331	1.128
	3	0.856	0.764	1.520	1.231

3.5 Mortality rate forecast

We also illustrate our method using an empirical data set. The Japanese sub-national mortality rates in 47 prefectures are used to demonstrate the effectiveness of our proposed method. Available from the Japanese Mortality Database (2017), the data set contains yearly age-specific mortality rates over a span of 41 years from 1975 to 2015. The observations are the yearly mortality curves from ages 0 to 110 years, where age is treated as the continuum in the rate function. In this study, the data at ages 95 and older are grouped together, to avoid problems associated with erratic rates at these ages.

A graphical display of the functional time series is presented in Figure 3.1. The figure

presents the log smoothed female age-specific mortality rates in the Tokyo prefecture, where the red lines represent more distant data and the purple lines represent more recent years. The curves are smoothed using penalized regression splines with a monotonically increasing constraint after the age of 65 (see Wood 1994a, Hyndman & Ullah 2007).

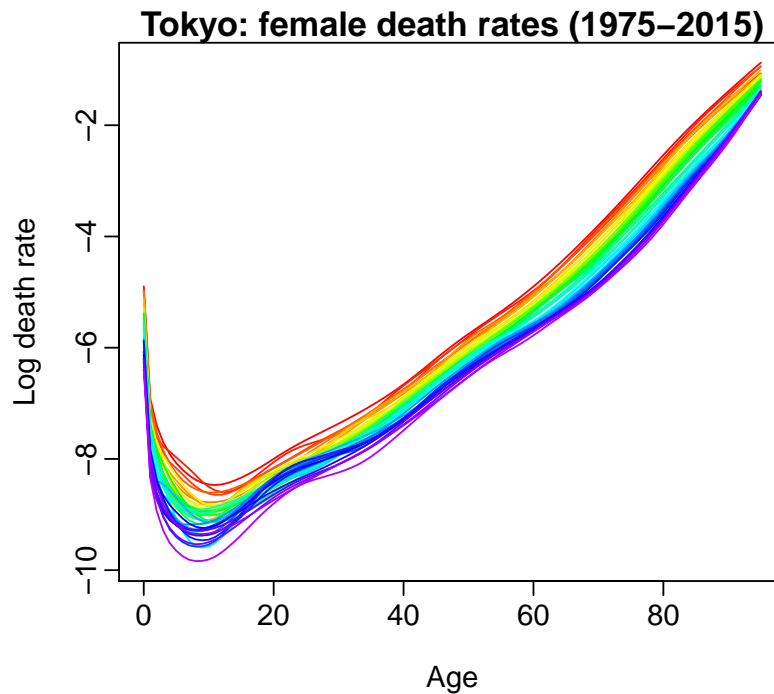


Figure 3.1: Log smoothed female mortality in the Tokyo prefecture from 1975 to 2015

The dimension of the functional time series is $N = 47$, which is greater than the sample size $T = 41$. With the twofold dimension reduction model, we use the first three FPC scores for each population, and the first three factors.

The expanding window approach is again used as described in the simulation study. The first 26 years of data (from 1975 to 2000) are allocated to the training set, and the last 15 years of data (from 2001 to 2015) are allocated to the testing set. In each fitting process, the bandwidth parameter q is chosen as $\sqrt{T^*}$, where T^* is the number of years

fitted to the model. We compare the forecast accuracy of our proposed method with the independent functional time series model, where each sub-national population is forecast individually using the approach in Hyndman & Ullah (2007). With Hyndman and Ullah's mortality model, the FPCA is performed on the functional time series of each prefecture and FPC scores are estimated and fitted to classical time series model to make predictions. Then functional forecast are produced using predicted scores. In this method, the prediction intervals are constructed by calculating the total variance of the principal components. This independent forecast model does not take into account the dependence between the sub-national populations.

For female mortality rates, prediction errors are calculated and show that the proposed method outperforms the independent model in general. Specifically, in 24 out of 47 prefectures, the proposed model produces smaller MAFE in all one to ten forecast horizons. In 45 out of 47 prefectures, the proposed model produces smaller mean MAFE taken across all forecast horizons.

We have also fitted the male mortality data of the 47 prefectures. The female and male forecast errors are summarized in Table 3.3. Each number in the table is the mean error taken across 47 prefectures. FTS stands for the alternative independent Hyndman and Ullah method and HDFTS stands for the proposed high-dimensional functional time series model. For female data, our model outperforms independent FTS model in both MAFE and MSFE values. For male data, however, our model produces smaller MAFE but does not have an advantage in MSFE values.

Prediction intervals are calculated based on the bootstrap approach. We use interval score as an evaluation for interval forecast. Let $\hat{\mathcal{X}}_{n+h|n}^u$ and $\hat{\mathcal{X}}_{n+h|n}^l$ denote the upper and lower $(1 - \alpha) \times 100\%$ prediction bounds, and \mathcal{X}_{n+h} is the realized value. The discretized

Table 3.3: MAFE and MSFE for the Japanese female and male rates

h	Female				Male			
	MAFE		MSFE		MAFE		MSFE	
	FTS	HDFTS	FTS	HDFTS	FTS	HDFTS	FTS	HDFTS
1	0.174	0.164	0.293	0.286	0.266	0.261	0.609	0.619
2	0.179	0.165	0.315	0.293	0.274	0.261	0.634	0.607
3	0.182	0.168	0.323	0.285	0.280	0.274	0.646	0.673
4	0.186	0.170	0.336	0.302	0.291	0.281	0.673	0.684
5	0.187	0.169	0.334	0.310	0.294	0.280	0.662	0.691
6	0.197	0.175	0.373	0.337	0.311	0.293	0.714	0.758
7	0.207	0.174	0.406	0.343	0.325	0.300	0.749	0.801
8	0.217	0.178	0.441	0.365	0.342	0.314	0.809	0.860
9	0.229	0.181	0.478	0.384	0.357	0.322	0.841	0.913
10	0.232	0.188	0.479	0.419	0.365	0.323	0.838	0.906
Mean	0.199	0.173	0.378	0.332	0.311	0.291	0.717	0.751
Median	0.192	0.172	0.354	0.323	0.302	0.287	0.693	0.724

interval score at point u_j is defined as

$$\begin{aligned}
S_\alpha(u_j) = & \left[\hat{\mathcal{X}}_{n+h|n}^u(u_j) - \hat{\mathcal{X}}_{n+h|n}^l(u_j) \right] \\
& + \frac{2}{\alpha} \left[\hat{\mathcal{X}}_{n+h|n}^l(u_j) - \mathcal{X}_{n+h}(u_j) \right] \mathbb{1} \left\{ \mathcal{X}_{n+h}(u_j) < \hat{\mathcal{X}}_{n+h|n}^l(u_j) \right\} \\
& + \frac{2}{\alpha} \left[\mathcal{X}_{n+h}(u_j) - \hat{\mathcal{X}}_{n+h|n}^u(u_j) \right] \mathbb{1} \left\{ \mathcal{X}_{n+h}(u_j) > \hat{\mathcal{X}}_{n+h|n}^u(u_j) \right\},
\end{aligned}$$

where α is the level of significance, and $\mathbb{1}\{\cdot\}$ is a binary indicator function. According to this standard, the best predicted interval is the one that gives the smallest interval score. In the functional case here, the point-wise interval scores are computed and the mean over the discretized ages is taken as a score for the whole curve. Then the average scores over all populations are calculated. Mean interval scores are shown in Figure 3.2. Though the values are not different by large scales, the proposed high-dimensional FTS model has an apparent advantage in interval predictions especially in long-run forecast.

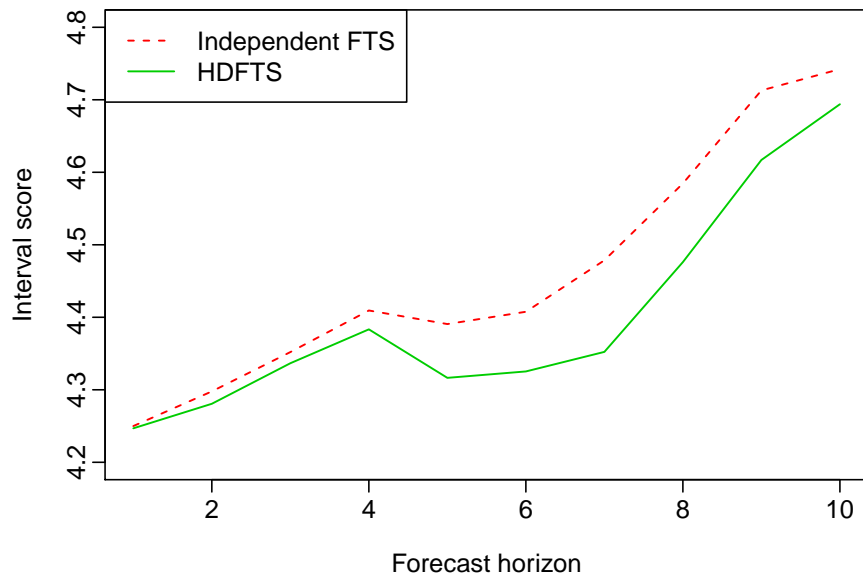


Figure 3.2: Mean interval score for one- to ten-step-ahead forecast. The green solid line represents the mean interval score for the high-dimensional functional time series model. The red colored dotted line represents the mean interval score for independent functional time series forecast

3.6 Conclusion

In this chapter, we have proposed a two-fold dimension reduction model for modeling and forecasting high-dimensional functional time series. Our approach utilizes dynamic FPCA and factor model to represent original data with low-dimensional time series. This offers a solution to the issue of the curse of dimensionality in high-dimensional data settings. We have also provided the asymptotic properties of the model when the dimension of the functional data N grows with the sample size T . When the tail terms in both dimension reduction steps converge to zero, the estimation error can be proved to converge to zero. Compared to the existing forecasting approaches, the proposed method has been proven to perform well both in simulations and in an empirical data analysis.

3.7 Appendix

Lemmas

We introduce some lemmas needed for the proofs of theorems. Lemma 1 is from the Lemma 4 in the supplement of Hörmann et al. (2015). Lemma 2 bounds the difference of eigenfunctions by the difference in covariance operators. Lemma 3 is known as Kronecker's Lemma, which helps with the proof of Lemma 4. The consistency of the estimated functional covariance operator is proved in Lemma 4. Lemma 5 is from Theorem 8.1.10 in Golub & Loan (2012). The idea of Lemma 5 is that the distance between two matrices determines the distance between the eigenvectors of the two matrices, which enables the proof for Theorem 2.

Let L_H^p be the space of H valued random variables \mathcal{X} and define $v_p(X) = (E\|\mathcal{X}\|^p)^{1/p}$.

Definition 1. A sequence $\{\mathcal{X}_t(u)\} \in L_H^p$ is called L^p - m -approximable if each \mathcal{X}_t admits the

representation,

$$\mathcal{X}_t = f(\epsilon_t, \epsilon_{t-1}, \dots),$$

where the ϵ_t are i.i.d. elements taking values in a measurable space S , and f is a measurable function $f : S^\infty \rightarrow H$. Moreover we assume that if $\{\epsilon'_t\}$ is an independent copy of $\{\epsilon_t\}$ defined on the same probability space, then letting

$$\mathcal{X}_t^{(m)} = f(\epsilon_t, \epsilon_{t-1}, \dots, \epsilon_{t-m+1}, \epsilon'_{t-m}, \epsilon'_{t-m-1}, \dots),$$

we have

$$\sum_{m=1}^{\infty} v_p(\mathcal{X}_m - \mathcal{X}_m^{(m)}) < \infty$$

To serve for the following lemmas, let $\mathcal{L} = \mathcal{L}(H, H)$ be the set of bounded linear operators from H to H . For $\Psi \in \mathcal{L}$, we define the operator norm $\|\Psi\|_{\mathcal{L}} = \sup_{\|x\| \leq 1} \|\Psi(x)\|$. A linear operator $\Psi \in \mathcal{L}(H, H)$ is Hilbert–Schmidt if for some orthogonal basis $(v_k, k \geq 1)$, we have $\|\Psi\|_{\mathcal{S}}^2 = \sum_{k \geq 1} \|\Psi(v_k)\|^2 < \infty$. Then $\|\Psi\|_{\mathcal{S}}$ defines the Hilbert–Schmidt norm. Recall that for any Hilbert–Schmidt operator $\Psi \in \mathcal{L}$, $\|\Psi\|_{\mathcal{L}} \leq \|\Psi\|_{\mathcal{S}}$.

Lemma 1. Assume that (\mathcal{X}_t) is an L^4 - m -approximable series. Let C_h and \widehat{C}_h be defined as in (3.1) and (3.11). Then, for all $|h| < T$, $E(\|C_h - \widehat{C}_h\|_{\mathcal{S}}) \leq U\sqrt{(|h| \vee 1)/T}$, where U is a constant, and T is the sample size.

Lemma 2. Suppose $C, K \in \mathcal{L}$ are two compact operators with singular value decompositions

$$C(x) = \sum_{j=1}^{\infty} \lambda_j \langle x, v_j \rangle f_j, \quad K(x) = \sum_{j=1}^{\infty} \gamma_j \langle x, u_j \rangle g_j.$$

Let $v'_j = \widehat{c}_j v_j$, where $\widehat{c}_j = \text{sgn}(\langle u_j, v_j \rangle)$. If C is Hilbert-Schmidt, symmetric and positive definite, and its eigenvalues satisfy

$$\lambda_1 > \dots > \lambda_d > \lambda_{d+1},$$

then

$$\|u_j - v'_j\| \leq \frac{2\sqrt{2}}{\alpha_j} \|K - C\|_{\mathcal{L}}, \quad 1 \leq j \leq d,$$

where $\alpha_1 = \lambda_1 - \lambda_2$ and $\alpha_j = \min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})$, $2 \leq j \leq d$

Lemma 3. If $(x_n)_{n=1}^{\infty}$ is an infinite sequence of real numbers such that

$$\sum_{m=1}^{\infty} x_m = s$$

exists and is finite, then we have for all $0 < b_1 \leq b_2 \leq b_3 \leq \dots$, and $b_n \rightarrow \infty$ that

$$\lim_{n \rightarrow \infty} \frac{1}{b_n} \sum_{k=1}^n b_k x_k = 0.$$

Lemma 4. Consider the estimator

$$\widehat{C} = \sum_{h=-q}^q \left(1 - \frac{|h|}{q}\right) \widehat{C}_h,$$

where \widehat{C}_h is the autocovariance operator at lag h . \widehat{C} is a consistent estimator for C .

Proof.

$$\begin{aligned}
\|C - \widehat{C}\|_S &= \left\| \sum_{h \in \mathbb{Z}} C_h - \sum_{h=-q}^q \left(1 - \frac{|h|}{q}\right) \widehat{C}_h \right\|_S \\
&\leq \left\| \sum_{h=-q}^q \left(1 - \frac{|h|}{q}\right) (\widehat{C}_h - C_h) \right\|_S + \left\| \frac{1}{q} \sum_{h=-q}^q |h| C_h \right\|_S + \left\| \sum_{|h|>q} C_h \right\|_S \\
&\leq \sum_{h=-q}^q \left(1 - \frac{|h|}{q}\right) \|C_h - \widehat{C}_h\|_S + \frac{1}{q} \sum_{h=-q}^q |h| \|C_h\|_S + \sum_{|h|>q} \|C_h\|_S.
\end{aligned}$$

On the right hand side of the above inequality, the second term tends to zero by Lemma 3. The last term tends to zero by Assumption 2.

For the first term,

$$\Pr \left(\sum_{h=-q}^q \left(1 - \frac{|h|}{q}\right) \|C_h - \widehat{C}_h\|_S > \epsilon \right) \leq \frac{1}{\epsilon} \sum_{h=-q}^q \left(1 - \frac{|h|}{q}\right) E[\|C_h - \widehat{C}_h\|_S] \leq U_1 \frac{q^{3/2}}{\sqrt{T}},$$

for some U_1 .

Thus $\|C - \widehat{C}\|_S = O_P(q^{3/2}/\sqrt{T}) = o_P(1)$, under Assumption 5. \square

Lemma 5 (Theorem 8.1.10 in Golub & Loan (2012)). *Suppose A and $A + E$ are $n \times n$ symmetric matrices and that $Q = [Q_1 \quad Q_2]$, where Q_1 has size $n \times r$ and Q_2 has size $n \times (n - r)$, is an orthogonal matrix such that $\text{span}(Q_1)$ is an invariant subspace for A ; that is $A \times \text{span}(Q_1) \subset \text{span}(A)$. Partition the matrices $Q^\top A Q$ and $Q^\top E Q$ as follows:*

$$Q^\top A Q = \begin{pmatrix} D_1 & \mathbf{0} \\ \mathbf{0} & D_2 \end{pmatrix}, \quad Q^\top E Q = \begin{pmatrix} E_{11} & E_{21}^\top \\ E_{21} & E_{22} \end{pmatrix}.$$

If $\text{sep}(D_1, D_2) := \min_{\lambda \in \lambda(D_1), \mu \in \lambda(D_2)} |\lambda - \mu| > 0$, where $\lambda(M)$ denotes the set of eigenvalues of

the matrix \mathbf{M} , and $\|E\|_2 \leq \text{sep}(\mathbf{D}_1, \mathbf{D}_2)/5$, then there exists a matrix $\mathbf{P} \in \mathbb{R}^{(n-r) \times r}$ with

$$\|\mathbf{P}\|_2 \leq \frac{4}{\text{sep}(\mathbf{D}_1, \mathbf{D}_2)} \|E_{21}\|_2,$$

such that the columns of $\widehat{\mathbf{Q}}_1 = (\mathbf{Q}_1 + \mathbf{Q}_2\mathbf{P})(\mathbf{I} + \mathbf{P}^\top\mathbf{P})^{-1/2}$ define an orthonormal basis for a subspace that is invariant for $\mathbf{A} + \mathbf{E}$.

Proof of Proposition 1

Without loss of generality, we assume that $E(\mathcal{X}_0^{(i)}) = 0$. Since $\mathcal{X}_0^{(i)}$ and $\mathcal{X}_h^{i,(h)}$, $h \geq 1$, are independent,

$$\|C_h^{(i)}\|_S = \left\| E\mathcal{X}_0^{(i)} \otimes (\mathcal{X}_h^{(i)} - \mathcal{X}_h^{i,(h)}) \right\|_S \leq (E\|\mathcal{X}_0^{(i)}\|^2)^{1/2} (E\|\mathcal{X}_h^{(i)} - \mathcal{X}_h^{i,(h)}\|^2)^{1/2},$$

which completes the proof.

It is easy to prove that a functional AR(1) model is L^2 - m -approximate (Hörmann & Kokoszka 2010). A functional AR(1) takes the form $\mathcal{X}_t(u) = \Psi(\mathcal{X}_{t-1})(u) + \epsilon_t(u)$, where $\Psi \in \mathcal{L}$ satisfies $\|\Psi\|_{\mathcal{L}} < 1$, and $\epsilon_t(u) \in L^2_H$ is i.i.d. with mean zero. The model can be written as the expansion $\mathcal{X}_t(u) = \sum_{j=0}^{\infty} \Psi^j(\epsilon_{t-j})(u)$, where Ψ^j is the j th iterate of the operator Ψ . Let $\mathcal{X}_t^{(m)} = \sum_{j=0}^m \Psi^j(\epsilon_{t-j}) + \sum_{j=m+1}^{\infty} \Psi^j(\epsilon_{t-j}^{(t)})$, where $\epsilon_k^{(t)}$ is an independent copy of ϵ_k for each t . It is easy to verify that for any $A \in \mathcal{L}$, $v_p(A(Y)) \leq \|A\|_{\mathcal{L}} v_p(Y)$. Since $\mathcal{X}_m - \mathcal{X}_m^{(m)} = \sum_{j=1}^{\infty} [\Psi^j(\epsilon_{m-j}) - \Psi^j(\epsilon_{m-j}^{(m)})]$, it follows that $v_p(\mathcal{X}_m - \mathcal{X}_m^{(m)}) \leq 2 \sum_{j=m}^{\infty} \|\Psi\|_{\mathcal{L}}^j v_p(\epsilon_0) = O(1) \times v(\epsilon_0) \|\Psi\|_{\mathcal{L}}^m$. By assumption $v_2(\epsilon_0) < \infty$, therefore $\sum_{m=1}^{\infty} v_2(\mathcal{X}_m - \mathcal{X}_m^{(m)}) < \infty$.

Proof of theorems

Proof of Theorem 1

Proof. We first prove the consistency for $\tilde{\beta}_{p,t}^{(i)}$, which is defined in (3.12):

$$\begin{aligned} |\beta_{p,t}^{(i)} - \tilde{\beta}_{p,t}^{(i)}| &= \left| \langle \mathcal{X}_t^{(i)}, \gamma_p^{(i)} \rangle - \langle \mathcal{X}_t^{(i)}, \hat{\gamma}_p^{(i)} \rangle \right| \\ &\leq \left| \langle \mathcal{X}_t^{(i)}, \gamma_p^{(i)} - \hat{\gamma}_p^{(i)} \rangle \right|_2 \\ &\leq \|\mathcal{X}_t^{(i)}\| \|\gamma_p^{(i)} - \hat{\gamma}_p^{(i)}\|_2. \end{aligned}$$

Under Assumption 4, $\hat{\gamma}_p^{(i)}$ is in the same direction as $\gamma_p^{(i)}$. Using Lemma 2

$$\|\gamma_p^{(i)} - \hat{\gamma}_p^{(i)}\|_2 \leq \frac{2\sqrt{2}}{\alpha_p^{(i)}} \|C^{(i)} - \hat{C}^{(i)}\|_{\mathcal{S}}, \quad \text{for } p = 1, \dots, p_0,$$

where

$$\alpha_1^{(i)} = \lambda_1^{(i)} - \lambda_2^{(i)}, \quad \alpha_p^{(i)} = \min(\lambda_{p-1}^{(i)} - \lambda_p^{(i)}, \lambda_p^{(i)} - \lambda_{p+1}^{(i)}),$$

with $\lambda_p^{(i)}$ defined in (3.2). Under Assumption 3, $\alpha_p^{(i)}$ is not equal to 0.

By Assumption 2 and Lemma 4, we have

$$|\tilde{\beta}_{p,t}^{(i)} - \beta_{p,t}^{(i)}| = O_P\left(\frac{q^{3/2}}{\sqrt{T}}\right).$$

Now for $i = 1, \dots, N$, $\beta_{p,t}^{(i)}$ and $\tilde{\beta}_{p,t}^{(i)}$ are combined into vectors $\tilde{\boldsymbol{\beta}}_{p,t}$ and $\boldsymbol{\beta}_{p,t}$. Then

$$\|\tilde{\boldsymbol{\beta}}_{p,t} - \boldsymbol{\beta}_{p,t}\|_2 = O_P\left(\frac{q^{3/2}\sqrt{N}}{\sqrt{T}}\right).$$

Under Assumption 5, $\sup_t \|\tilde{\boldsymbol{\beta}}_{p,t} - \boldsymbol{\beta}_{p,t}\|_2 = o_P(1)$. The proof is complete.

Since we are following a similar process for each p^{th} FPC score, we will, without confusion, omit the subscript p in the following proof, so that β_t denotes $\beta_{p,t}$, A denotes A_p and f_t denotes $f_{p,t}$. The matrices L , L^* and covariance matrices also denote the ones corresponding to a specific p , $p = 1, \dots, p_0$. \square

Proof of Theorem 2

Proof. Recall that we have modeled the FPC scores as:

$$\beta_t = A f_t + e_t.$$

And we have defined in Section 2.2 that

$$\begin{aligned} L &= \Sigma_\beta(h) \Sigma_\beta^\top(h) \\ &= A \left[\sum_{h=1}^{h=h_0} \{ \Sigma_f(h) A^\top + \Sigma_{f,e}(h) + \Sigma_{f,e}(-h) \} \{ \Sigma_f(h) A^\top + \Sigma_{f,e}(h) + \Sigma_{f,e}(-h) \}^\top \right] A^\top + E \\ &= ADA^\top + E \\ &= L^* + E, \end{aligned} \tag{3.17}$$

where ADA^\top is the eigendecomposition of L^* , as elucidated in Section 2.2.

To apply Lemma 5 on L^* and \widehat{L} , let B be an orthogonal complement of A , then $L^*B = 0$, and

$$\begin{pmatrix} A^\top \\ B^\top \end{pmatrix} L^* \begin{pmatrix} A & B \end{pmatrix} = \begin{pmatrix} D & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix},$$

with $\text{sep}(D, \mathbf{0}) = \lambda_{\min}(D)$, as defined in Lemma 5.

Next, we will find the order of $\lambda_{\min}(\mathbf{D})$. Define

$$\begin{aligned}\mathbf{W}_f(h_0) &= (\boldsymbol{\Sigma}_f(1), \dots, \boldsymbol{\Sigma}_f(h_0)), \\ \mathbf{W}_{f,e}(h_0) &= (\boldsymbol{\Sigma}_{f,e}(1), \dots, \boldsymbol{\Sigma}_{f,e}(h_0)) \quad \mathbf{W}_{f,e}(-h_0) = (\boldsymbol{\Sigma}_{f,e}(-1), \dots, \boldsymbol{\Sigma}_{f,e}(-h_0)),\end{aligned}$$

so the summation in (3.17) can be written out as $\mathbf{D} = \{\mathbf{W}_f(h_0)(\mathbf{I}_{h_0} \otimes \mathbf{A}^\top) + \mathbf{W}_{f,e}(h_0) + \mathbf{W}_{f,e}(-h_0)\} \{\mathbf{W}_f(h_0)(\mathbf{I}_{h_0} \otimes \mathbf{A}^\top) + \mathbf{W}_{f,e}(h_0) + \mathbf{W}_{f,e}(-h_0)\}^\top$.

Let $\sigma_k(\mathbf{M})$ be the k^{th} singular value in descending order of matrix \mathbf{M} . We have

$$\begin{aligned}\lambda_{\min} &= \left[\sigma_r \{ \mathbf{W}_f(h_0)(\mathbf{I}_{h_0} \otimes \mathbf{A}^\top) + \mathbf{W}_{f,e}(h_0) + \mathbf{W}_{f,e}(-h_0) \} \right]^2 \\ &\geq \left[\sigma_r \{ \mathbf{W}_f(h_0)(\mathbf{I}_{h_0} \otimes \mathbf{A}^\top) \} - \sigma_1 \{ \mathbf{W}_{f,e}(h_0) \} - \sigma_1 \{ \mathbf{W}_{f,e}(-h_0) \} \right]^2 \\ &= \left[\sigma_r \{ \mathbf{W}_f(h_0) \} - \sigma_1 \{ \mathbf{W}_{f,e}(h_0) \} - \sigma_1 \{ \mathbf{W}_{f,e}(-h_0) \} \right]^2 \\ &\asymp \sigma_r^2 \{ \mathbf{W}_f(h_0) \} \asymp N^{2-2\delta},\end{aligned}$$

where the last step used Assumption 6 and 7. Hence,

$$N^{2-2\delta} = O\{\lambda_{\min}(\mathbf{D})\}. \quad (3.18)$$

Next, we will find the distance between \mathbf{L}^* and $\widehat{\mathbf{L}}$. Let $\mathbf{E}_L = \widehat{\mathbf{L}} - \mathbf{L}^*$,

$$\begin{aligned}\|\mathbf{E}_L\|_2 &= \|\widehat{\mathbf{L}} - \mathbf{L} + \mathbf{E}\|_2 \\ &\leq \|\widehat{\mathbf{L}} - \widetilde{\mathbf{L}}\|_2 + \|\widetilde{\mathbf{L}} - \mathbf{L}\|_2 + \|\mathbf{E}\|_2,\end{aligned} \quad (3.19)$$

where $\widehat{\mathbf{L}}$ is defined in (13), and

$$\widetilde{\mathbf{L}} := \sum_{h=1}^{h_0} \widetilde{\boldsymbol{\Sigma}}_\beta(h) \widetilde{\boldsymbol{\Sigma}}_\beta(h)^\top,$$

with

$$\tilde{\Sigma}_\beta(h) = \frac{1}{T-k} \sum_{t=1}^{T-k} (\beta_{t+k} - \bar{\beta})(\beta_t - \bar{\beta})^\top,$$

where $\bar{\beta} = (1/T) \sum_{t=1}^T \beta_t$.

We will find the order of each of the three terms on the right hand side in (3.19). Let's first consider the last term. By (10) and Assumption 7, we get

$$\begin{aligned} \|E\|_2 &\leq (\|\Sigma_f(h)\|_2 + \|\Sigma_{f,e}(h)\|_2) \|\Sigma_e(h)\|_2 + \|\Sigma_e(h)\|_2^2 \\ &= \begin{cases} O(N^{1-\delta}), & \text{when } NT^{-1/2} \rightarrow \infty \\ O(N^{2-\delta}T^{-1/2}), & \text{when } NT^{-1/2} = o(1) \end{cases} \end{aligned} \quad (3.20)$$

Next consider

$$\begin{aligned} \|E_{L_1}\|_2 &:= \|\hat{L} - \tilde{L}\|_2 \\ &\leq \sum_{h=1}^{h_0} \left(\|\hat{\Sigma}_\beta(h) - \tilde{\Sigma}_\beta(h)\|_2^2 + 2\|\tilde{\Sigma}_\beta(h)\|_2 \times \|\hat{\Sigma}_\beta(h) - \tilde{\Sigma}_\beta(h)\|_2 \right). \end{aligned}$$

We can find the rate for each term on the right hand side.

$$\begin{aligned} \|\hat{\Sigma}_\beta(h) - \tilde{\Sigma}_\beta(h)\|_2 &\leq \frac{1}{T-h} \sum_{t=1}^{T-h} \left\| (\tilde{\beta}_{t+h} - \bar{\beta})(\tilde{\beta}_t - \bar{\beta})^\top - (\beta_{t+h} - \bar{\beta})(\beta_t - \bar{\beta})^\top \right\|_2 \\ &\leq \frac{1}{T-h} \sum_{t=1}^{T-h} \left[\|(\tilde{\beta}_{t+h} - \beta_{t+h}) + (\bar{\beta} - \tilde{\beta})\|_2 \|(\beta_t - \bar{\beta})^\top\|_2 + \right. \\ &\quad \left. \|\tilde{\beta}_{t+h} - \bar{\beta}\|_2 \|(\tilde{\beta}_t - \beta_t)^\top + (\bar{\beta} - \tilde{\beta})^\top\|_2 \right] = o_p(1), \end{aligned} \quad (3.21)$$

where Theorem 1 is used.

Under Assumption 2, the functional time series $\mathcal{X}_t^{(i)}(u)$ is L^4 -m-approximable. Then

each element of $\tilde{\Sigma}_\beta(h) - \Sigma_\beta(h)$ is $O_P(T^{-1/2})$, so

$$\|\tilde{\Sigma}_\beta(h) - \Sigma_\beta(h)\|_2 = O_P(NT^{-1/2}) \quad (3.22)$$

$$\begin{aligned} \|\Sigma_\beta(h)\|_2 &= \|A\Sigma_f(h)A^\top + A\Sigma_{f,e}(h) + A\Sigma_{f,e}(-h) + \Sigma_e(h)\|_2 \\ &\leq \|\Sigma_f(h)\|_2 + 2\|\Sigma_{f,e}(h)\|_2 + \|\Sigma_e(h)\|_2 \\ &= O(N^{1-\delta}), \end{aligned} \quad (3.23)$$

where Assumptions 6 and 7 are used.

$$\begin{aligned} \|\tilde{\Sigma}_\beta(h)\|_2 &\leq \|\tilde{\Sigma}_\beta(h) - \Sigma_\beta(h)\|_2 + \|\Sigma_\beta(h)\|_2 \\ &= O_P(NT^{-1/2}) + O_P(N^{1-\delta}). \end{aligned}$$

Using (3.21) and (10),

$$\|\mathbf{E}_{L_1}\|_2 = o_P(NT^{-1/2} + N^{1-\delta}). \quad (3.24)$$

Last we consider

$$\begin{aligned} \|\mathbf{E}_{L_2}\|_2 &:= \|\tilde{\mathbf{L}} - \mathbf{L}\|_2 \\ &\leq \sum_{h=1}^{h_0} \left(\|\tilde{\Sigma}_\beta(h) - \Sigma_\beta(h)\|_2^2 + 2\|\Sigma_\beta(h)\| \times \|\tilde{\Sigma}_\beta(h) - \Sigma_\beta(h)\|_2 \right) \\ &= O_P(N^2T^{-1}) + O(N^{1-\delta}) \times O_P(NT^{-1/2}) \\ &= O_P(N^{2-\delta}T^{-1/2}), \end{aligned} \quad (3.25)$$

where (3.22) and (3.23) are used.

Thus combing (3.20), (3.24) and (3.25),

$$\begin{aligned}\|\mathbf{E}_L\|_2 &\leq \|\mathbf{E}_{L_1}\|_2 + \|\mathbf{E}_{L_2}\|_2 + \|\mathbf{E}\|_2 \\ &= O_P(N^{2-\delta}T^{-1/2} + N^{1-\delta})\end{aligned}$$

Compared with the order for $\text{sep}(\mathbf{D}, \mathbf{0})$ in (3.18),

$$\begin{aligned}\|\mathbf{E}_L\|_2 &= O_P(N^{2-\delta}T^{-1/2} + N^{1-\delta}) = o_P(N^{2-2\delta}) \\ &= O_P\{\text{sep}(\mathbf{D}, \mathbf{0})\} = O_P\{\lambda(\mathbf{D})\}\end{aligned}$$

According to Lemma 5 there exists a matrix \mathbf{P} such that

$$\|\mathbf{P}\|_2 \leq \frac{4}{\text{sep}(\mathbf{D}, \mathbf{0})} \|\mathbf{E}_L\|_2,$$

and $\widehat{\mathbf{A}} = (\mathbf{A} + \mathbf{B}\mathbf{P})(\mathbf{I} + \mathbf{P}^\top \mathbf{P})^{-1/2}$.

Thus

$$\begin{aligned}\|\widehat{\mathbf{A}} - \mathbf{A}\|_2 &= \|[\mathbf{A}(\mathbf{I} - (\mathbf{I} + \mathbf{P}^\top \mathbf{P})^{1/2}) + \mathbf{B}\mathbf{P}] [(\mathbf{I} + \mathbf{P}^\top \mathbf{P})^{-1/2}]\|_2 \\ &\leq \|\mathbf{I} - (\mathbf{I} + \mathbf{P}^\top \mathbf{P})^{1/2}\|_2 + \|\mathbf{P}\|_2 \leq 2\|\mathbf{P}\|_2.\end{aligned}$$

Since $N^{2-2\delta} = O(\lambda_{\min}(\mathbf{D}))$,

$$\begin{aligned}\|\widehat{\mathbf{A}} - \mathbf{A}\|_2 &= O_P\left(\frac{N^{2-\delta}T^{-1/2} + N^{1-\delta}}{N^{2-2\delta}}\right) \\ &= O_P(N^\delta T^{-1/2} + N^{\delta-1}).\end{aligned}$$

This completes the proof. \square

Proof of Theorem 3

Proof. Recall our model is $\beta_t = \mathbf{A}f_t + e_t$. Now we consider

$$\begin{aligned} \widehat{\mathbf{A}}\widehat{f}_t - \mathbf{A}f_t &= \widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top \widetilde{\beta}_t - \mathbf{A}f_t \\ &= \widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top [\beta_t + (\widetilde{\beta}_t - \beta_t)] - \mathbf{A}f_t \\ &= \widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top \beta_t - \mathbf{A}f_t + \widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top (\widetilde{\beta}_t - \beta_t) \\ &= \mathbf{K}_1 + \mathbf{K}_2, \end{aligned}$$

where

$$\mathbf{K}_1 = \widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top \beta_t - \mathbf{A}f_t = \widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top \mathbf{A}f_t - \mathbf{A}f_t + \widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top e_t = \mathbf{K}_{1,1} + \mathbf{K}_{1,2} + \mathbf{K}_{1,3},$$

with $\mathbf{K}_{1,1} = (\widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top - \mathbf{A}\mathbf{A}^\top)\mathbf{A}f_t$, $\mathbf{K}_{1,2} = \widehat{\mathbf{A}}(\widehat{\mathbf{A}} - \mathbf{A})^\top e_t$ and $\mathbf{K}_{1,3} = \widehat{\mathbf{A}}\mathbf{A}^\top e_t$. We have

$$\|\mathbf{K}_{1,1}\|_2 = O_P(\|\widehat{\mathbf{A}} - \mathbf{A}\|_2 \|f_t\|_2) = O_P(N^{(1-\delta)/2} \|\widehat{\mathbf{A}} - \mathbf{A}\|_2).$$

As in Theorem 2, $\|\widehat{\mathbf{A}} - \mathbf{A}\|_2 = o_P(1)$ and $\|\mathbf{A}\|_2 = 1$, we have $\|\mathbf{K}_{1,2}\|_2$ dominated by $\|\mathbf{K}_{1,3}\|_2$ in probability. Hence we only need to consider $\mathbf{K}_{1,3}$. Now consider for $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_r)$, the random variable $\mathbf{a}_j^\top e_t$, with

$$\mathbb{E}(\mathbf{a}_j^\top e_t) = 0, \quad \text{var}(\mathbf{a}_j^\top e_t) = \mathbf{a}_j^\top \boldsymbol{\Sigma}_e \mathbf{a}_j \leq \lambda_{\max}(\boldsymbol{\Sigma}_e) < c < \infty,$$

for $j = 1, \dots, r$ by Assumption 7, where c is a constant independent of T and r . Hence $\mathbf{a}_j^\top \mathbf{e}_t = O_P(1)$. We then have

$$\|\mathbf{K}_{1,3}\|_2 = \|\widehat{\mathbf{A}}\mathbf{A}^\top \mathbf{e}_t\|_2 \leq \|\mathbf{A}^\top \mathbf{e}_t\|_2 = \sum_{j=1}^r (\mathbf{a}_j^\top \mathbf{e}_t)^2 = O_P(1).$$

Thus,

$$\|\mathbf{K}_1\|_2 = O_P(N^{(1-\delta)/2} \|\widehat{\mathbf{A}} - \mathbf{A}\|_2 + 1).$$

$$\begin{aligned} \|\mathbf{K}_2\|_2 &= \|\widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top (\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t)\|_2 \\ &\leq \|\widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top\|_2 \|\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t\|_2 \\ &\leq \left(\|\widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top - \mathbf{A}\mathbf{A}^\top\|_2 + \|\mathbf{A}\mathbf{A}^\top\|_2 \right) \|\widetilde{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t\|_2 \\ &= O_P \left(\left[\|\widehat{\mathbf{A}} - \mathbf{A}\|_2 + 1 \right] \frac{q^{3/2}\sqrt{N}}{\sqrt{T}} \right) \\ &= o_P(1). \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{1}{N} \|\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}_t\|_2 &= \frac{1}{N} \left\| \widehat{\mathbf{A}}\widehat{\mathbf{f}}_t - \mathbf{A}\mathbf{f}_t - \mathbf{e}_t \right\|_2 \\ &= \frac{1}{N} (\|\mathbf{K}_1\|_2 + \|\mathbf{K}_2\|_2 + \|\mathbf{e}_t\|_2) \\ &= O_P(N^{(\delta-1)/2} \|\widehat{\mathbf{A}} - \mathbf{A}\|_2 + \frac{1}{N}) \\ &= O_P(N^{(\delta-1)/2} T^{-1/2}) + O_P\left(\frac{1}{N}\right), \end{aligned}$$

where $\|\mathbf{e}_t\|_2 = O_P(1)$ by Assumption 9.

The proof is complete. □

Proof of Theorem 4

Proof. We are comparing the true functions $\mathcal{X}_t^{(i)}(u)$ and estimated functions $\widehat{\mathcal{X}}_t^{(i)}(u)$:

$$\begin{aligned}\mathcal{X}_t^{(i)}(u) &= \sum_{p=1}^{p_0} [\boldsymbol{\beta}_{p,t}]^{(i)} \gamma_p^{(i)} + \boldsymbol{\epsilon}_t^{(i)}(u) \\ \widehat{\mathcal{X}}_t^{(i)}(u) &= \sum_{p=1}^{p_0} [\widehat{\boldsymbol{\beta}}_{p,t}]^{(i)} \widehat{\gamma}_p^{(i)}(u).\end{aligned}$$

$$\begin{aligned}& \frac{1}{N} \sum_{i=1}^N \|\widehat{\mathcal{X}}_t^{(i)}(u) - \mathcal{X}_t^{(i)}(u)\|_2 \\ & \leq \frac{1}{N} \sum_{i=1}^N \sum_{p=1}^{p_0} \left\| [\widehat{\boldsymbol{\beta}}_{p,t}]^{(i)} \widehat{\gamma}_p^{(i)}(u) - [\boldsymbol{\beta}_{p,t}]^{(i)} \gamma_p^{(i)}(u) \right\|_2 + \frac{1}{N} \sum_{i=1}^N \|\boldsymbol{\epsilon}_t^{(i)}(u)\|_2 \\ & = \frac{1}{N} \sum_{p=1}^{p_0} \sum_{i=1}^N \left\| \{[\widehat{\boldsymbol{\beta}}_{p,t}]^{(i)} - [\boldsymbol{\beta}_{p,t}]^{(i)}\} \widehat{\gamma}_p^{(i)}(u) + [\boldsymbol{\beta}_{p,t}]^{(i)} (\widehat{\gamma}_p^{(i)} - \gamma_p^{(i)}) \right\|_2 + \frac{1}{N} \sum_{i=1}^N \|\boldsymbol{\epsilon}_t^{(i)}(u)\|_2 \\ & \leq \frac{1}{N} \sum_{p=1}^{p_0} \sum_{i=1}^N \left\{ \left\| [\widehat{\boldsymbol{\beta}}_{p,t}]^{(i)} - [\boldsymbol{\beta}_{p,t}]^{(i)} \right\| \|\widehat{\gamma}_p^{(i)}(u)\|_2 + \left\| [\boldsymbol{\beta}_{p,t}]^{(i)} \right\| \|\widehat{\gamma}_p^{(i)} - \gamma_p^{(i)}\|_2 \right\} + \frac{1}{N} \sum_{i=1}^N \|\boldsymbol{\epsilon}_t^{(i)}(u)\|_2 \\ & \leq \frac{1}{N} \sum_{p=1}^{p_0} \left\{ \sqrt{\sum_{i=1}^N [\widehat{\boldsymbol{\beta}}_{p,t} - \boldsymbol{\beta}_{p,t}]^2 \sum_{i=1}^N \|\widehat{\gamma}_p^{(i)}\|_2^2} + \sqrt{\sum_{i=1}^N [\boldsymbol{\beta}_{p,t}]^{(i)2} \sum_{i=1}^N \|\widehat{\gamma}_p^{(i)} - \gamma_p^{(i)}\|_2^2} \right\} \\ & + \frac{1}{N} \sum_{i=1}^N \|\boldsymbol{\epsilon}_t^{(i)}(u)\|_2 \quad (\text{Cauchy inequality used}) \\ & \leq \frac{1}{N} \sum_p^{p_0} \left\{ \sqrt{\|\widehat{\boldsymbol{\beta}}_{p,t} - \boldsymbol{\beta}_{p,t}\|_2^2 \sum_{i=1}^N \left[\|\widehat{\gamma}_p^{(i)} - \gamma_p^{(i)}\|_2^2 + \|\gamma_p^{(i)}\|_2^2 + 2\|\widehat{\gamma}_p^{(i)} - \gamma_p^{(i)}\|_2 \|\gamma_p^{(i)}\|_2 \right]} \right. \\ & \left. + \sqrt{\|\boldsymbol{\beta}_{p,t}\|_2^2 \sum_{i=1}^N \|\widehat{\gamma}_p^{(i)} - \gamma_p^{(i)}\|_2^2} \right\} + \frac{1}{N} \sum_{i=1}^N \|\boldsymbol{\epsilon}_t^{(i)}(u)\|_2.\end{aligned}$$

Under Assumption 8 and 9, $\|\boldsymbol{\epsilon}_t^{(i)}(u)\|_2 = o_p(1)$. With the results from Theorem 1, 2

and 3,

$$\begin{aligned}
\frac{1}{N} \sum_{i=1}^N \|\widehat{\mathcal{X}}_t^{(i)}(u) - \mathcal{X}_t^{(i)}(u)\|_2 &= O_P\left(\frac{1}{N} \sqrt{(N^{(1+\delta)/2} T^{-1/2} + 1) \times N + N^{1-\delta} \times \frac{Nq^3}{T}}\right) \\
&= O_P(N^{\delta/2} T^{-1/2} + N^{-1/2} + N^{-\delta/2} q^{3/2} T^{-1/2}) \\
&= o_P(1).
\end{aligned}$$

The proof is complete. \square

Proof of Theorem 5

Proof. First we need to make notations clear that the true factors $\{f_{p,1}, \dots, f_{p,T+h}\}$ are defined in (6), the estimated factors $\{\widehat{f}_{p,1}, \dots, \widehat{f}_{p,T}\}$ are defined in (15), and the forecast factors $\{\widehat{f}_{p,T+1|T}, \dots, \widehat{f}_{p,T+h|T}\}$ are defined in (16), where each is an r -dimension vector.

Using the model $\boldsymbol{\beta}_{p,t} = \mathbf{A} f_{p,t} + \mathbf{e}_{p,t}$, and that $\mathbf{A}_p^\top \mathbf{A}_p = \mathbf{I}$, we have $f_{p,t} = \mathbf{A}_p^\top (\boldsymbol{\beta}_{p,t} - \mathbf{e}_{p,t})$. $\widehat{f}_{p,t}$ is estimated using (14). Then

$$\begin{aligned}
\|f_{p,t} - \widehat{f}_{p,t}\|_2 &= \|\mathbf{A}_p^\top \boldsymbol{\beta}_{p,t} - \widehat{\mathbf{A}}_p^\top \widetilde{\boldsymbol{\beta}}_{p,t} - \mathbf{A}_p^\top \mathbf{e}_{p,t}\|_2 \\
&\leq \|\mathbf{A}_p^\top - \widehat{\mathbf{A}}_p^\top\|_2 \|\boldsymbol{\beta}_{p,t}\|_2 + \|\widehat{\mathbf{A}}_p\|_2 \|\boldsymbol{\beta}_{p,t} - \widetilde{\boldsymbol{\beta}}_{p,t}\|_2 + \|\mathbf{A}_p^\top \mathbf{e}_{p,t}\|_2 \\
&= O_P(\|\mathbf{A}_p^\top - \widehat{\mathbf{A}}_p^\top\|_2 \|\boldsymbol{\beta}_{p,t}\|_2 + \|\boldsymbol{\beta}_{p,t} - \widetilde{\boldsymbol{\beta}}_{p,t}\|_2) \\
&= O_P\{(N^\delta T^{-1/2} + N^{\delta-1}) N^{(1-\delta)/2} + 1\} \\
&= O_P(N^{(1+\delta)/2} T^{-1/2} + 1)
\end{aligned} \tag{3.26}$$

where the results of Theorem 1 and Theorem 2 are used.

To make forecast, we assume r scalar time series models to each element $f_{p,t}^{(i)}, i =$

$1, \dots, r$ in $f_{p,t}$. In this case, we use AR(1) model, that is:

$$f_{p,t}^{(i)} = \phi_{p,i} f_{p,t-1}^{(i)} + \omega_{p,t}^{(i)} \quad t = 2, \dots, T+h.$$

The forecast factor follows

$$\widehat{f}_{p,t|T}^{(i)} = \widehat{\phi}_{p,i} \widehat{f}_{p,t-1|T}^{(i)} \quad t = T+1, \dots, T+h,$$

where $\widehat{\phi}_{p,i}$ is estimated autoregressive parameter calculated using estimated factors. Then iteratively we can deduce:

$$f_{p,T+h}^{(i)} = \phi_{p,i}^h f_{p,T}^{(i)} + \sum_{j=1}^{h-1} \phi_{p,i}^j \omega_{p,T+h-j}^{(i)} \quad (3.27)$$

and

$$\widehat{f}_{p,T+h|T}^{(i)} = \widehat{\phi}_{p,i}^h \widehat{f}_{p,T}^{(i)}. \quad (3.28)$$

By definition, the parameter $\phi_{p,i} = \rho_{p,i}(1)$, where $\rho_{p,i}(1)$ is the lag 1 autocorrelation of $f_{p,t}^{(i)}$. The estimator for AR(1) parameter is $\widehat{\phi}_{p,i} = \widehat{\rho}_{p,i}(1)$, which is the sample lag 1 autocorrelation of $\widehat{f}_{p,t}^{(i)}$, and is calculated as

$$\widehat{\rho}_{p,i}(1) = \frac{\sum_{t=1}^{T-1} \widehat{f}_{p,t}^{(i)} \widehat{f}_{p,t+1}^{(i)}}{\sum_{t=1}^T \widehat{f}_{p,t}^{(i)2}}.$$

Define also $\widetilde{\phi}_{p,i}$ as the sample AR parameter calculated using the true factors $f_{p,1}^{(i)}, \dots, f_{p,T}^{(i)}$,

that is

$$\tilde{\phi}_{p,i} = \frac{\sum_{t=1}^{T-1} f_{p,t}^{(i)} f_{p,t+1}^{(i)}}{\sum_{t=1}^T f_{p,t}^{(i)^2}}.$$

And,

$$\begin{aligned} |\hat{\phi}_{p,i} - \tilde{\phi}_{p,i}| &= |\hat{\rho}_{p,i}(1) - \tilde{\rho}_{p,i}(1)| \\ &= \left| \frac{\sum_{t=1}^{T-1} \hat{f}_{p,t}^{(i)} \hat{f}_{p,t+1}^{(i)}}{\sum_{t=1}^T \hat{f}_{p,t}^{(i)^2}} - \frac{\sum_{t=1}^{T-1} f_{p,t}^{(i)} f_{p,t+1}^{(i)}}{\sum_{t=1}^T f_{p,t}^{(i)^2}} \right| \\ &= \left| \frac{\sum_{t=1}^T f_{p,t}^{(i)^2} \sum_{t=1}^{T-1} \hat{f}_{p,t}^{(i)} \hat{f}_{p,t+1}^{(i)} - \sum_{t=1}^T \hat{f}_{p,t}^{(i)^2} \sum_{t=1}^{T-1} f_{p,t}^{(i)} f_{p,t+1}^{(i)}}{\sum_{t=1}^T \hat{f}_{p,t}^{(i)^2} \sum_{t=1}^T f_{p,t}^{(i)^2}} \right| \\ &= \frac{1}{\Delta} \left| \left(\sum_{t=1}^T f_{p,t}^{(i)^2} - \sum_{t=1}^T \hat{f}_{p,t}^{(i)^2} \right) \sum_{t=1}^{T-1} \hat{f}_{p,t}^{(i)} \hat{f}_{p,t+1}^{(i)} - \left(\sum_{t=1}^{T-1} f_{p,t}^{(i)} f_{p,t+1}^{(i)} - \sum_{t=1}^{T-1} \hat{f}_{p,t}^{(i)} \hat{f}_{p,t+1}^{(i)} \right) \sum_{t=1}^T \hat{f}_{p,t}^{(i)^2} \right| \\ &= \frac{1}{\Delta} \left| \sum_{t=1}^T (f_{p,t}^{(i)} + \hat{f}_{p,t}^{(i)}) (f_{p,t}^{(i)} - \hat{f}_{p,t}^{(i)}) \sum_{t=1}^{T-1} \hat{f}_{p,t}^{(i)} \hat{f}_{p,t+1}^{(i)} \right. \\ &\quad \left. - \sum_{t=1}^{T-1} [(f_{p,t}^{(i)} - \hat{f}_{p,t}^{(i)}) f_{p,t+1}^{(i)} + (f_{p,t+1}^{(i)} - \hat{f}_{p,t+1}^{(i)}) \hat{f}_{p,t}^{(i)}] \sum_{t=1}^T \hat{f}_{p,t}^{(i)^2} \right| \\ &\leq \frac{1}{\Delta} \left\{ \sum_{t=1}^T |f_{p,t}^{(i)} + \hat{f}_{p,t}^{(i)}| |f_{p,t}^{(i)} - \hat{f}_{p,t}^{(i)}| \sum_{t=1}^{T-1} |\hat{f}_{p,t}^{(i)} \hat{f}_{p,t+1}^{(i)}| + \right. \\ &\quad \left. + \sum_{t=1}^{T-1} [|f_{p,t}^{(i)} - \hat{f}_{p,t}^{(i)}| |f_{p,t+1}^{(i)}| + |f_{p,t+1}^{(i)} - \hat{f}_{p,t+1}^{(i)}| |\hat{f}_{p,t}^{(i)}|] \sum_{t=1}^T |\hat{f}_{p,t}^{(i)^2}| \right\}, \end{aligned}$$

where $i = 1, \dots, r$. We denote the order for $|f_{p,t}^{(i)}|$ as $\eta_{p,i}$ and the order for $|\hat{f}_{p,t}^{(i)}|$ as $\eta'_{p,i}$, that is, $|f_{p,t}^{(i)}| \asymp \eta_{p,i}$ and $|\hat{f}_{p,t}^{(i)}| \asymp \eta'_{p,i}$. Then, $\Delta = \sum_{t=1}^T \hat{f}_{p,t}^{(i)^2} \sum_{t=1}^T f_{p,t}^{(i)^2} \asymp T^2 \eta_{p,i}^2 \eta'_{p,i}{}^2$. It can be calculated that the terms contained in the braces in the above inequality are of order

$T^2(\eta_{p,i} + \eta'_{p,i})\eta_{p,i}^2 |f_{p,t}^{(i)} - \widehat{f}_{p,t}^{(i)}|$. Thus,

$$\begin{aligned} |\widehat{\phi}_{p,i} - \tilde{\phi}_{p,i}| &\asymp \frac{|f_{p,t}^{(i)}| + |\widehat{f}_{p,t}^{(i)}|}{|f_{p,t}^{(i)2}|} |f_{p,t}^{(i)} - \widehat{f}_{p,t}^{(i)}| = \frac{|f_{p,t}^{(i)} - \widehat{f}_{p,t}^{(i)}|}{|f_{p,t}^{(i)}|} + \frac{|f_{p,t}^{(i)} - \widehat{f}_{p,t}^{(i)}| |\widehat{f}_{p,t}^{(i)}|}{|f_{p,t}^{(i)2}|} \\ &\leq \frac{|f_{p,t}^{(i)} - \widehat{f}_{p,t}^{(i)}|}{|f_{p,t}^{(i)}|} + \frac{|f_{p,t}^{(i)} - \widehat{f}_{p,t}^{(i)}| (|f_{p,t}^{(i)}| + |f_{p,t}^{(i)} - \widehat{f}_{p,t}^{(i)}|)}{|f_{p,t}^{(i)2}|} \\ &= \frac{2|f_{p,t}^{(i)} - \widehat{f}_{p,t}^{(i)}|}{|f_{p,t}^{(i)}|} + \frac{|f_{p,t}^{(i)} - \widehat{f}_{p,t}^{(i)}|^2}{|f_{p,t}^{(i)2}|}. \end{aligned}$$

Let

$$\Omega = \frac{|f_{p,t}^{(i)} - \widehat{f}_{p,t}^{(i)}|}{|f_{p,t}^{(i)}|},$$

then

$$\begin{aligned} \Omega &= O_P\left(\frac{N^{(1+\delta)/2} T^{-1/2} + 1}{N^{(1-\delta)/2}}\right) \\ &= O_P(N^\delta T^{-1/2} + N^{(\delta-1)/2}) \\ &= o_P(1), \end{aligned}$$

where (3.26) and Assumption 7 are used. And $|\widehat{\phi}_{p,i} - \tilde{\phi}_{p,i}| = O_P(\Omega)$.

According to basic time series textbooks, we know $|\tilde{\phi}_{p,i} \phi_{p,i}| = O_P(1/\sqrt{T})$. Then

$$\begin{aligned} |\widehat{\phi}_{p,i} - \phi_{p,i}| &= |\widehat{\phi}_{p,i} - \tilde{\phi}_{p,i} + \tilde{\phi}_{p,i} - \phi_{p,i}| \\ &\leq |\widehat{\phi}_{p,i} - \tilde{\phi}_{p,i}| + |\tilde{\phi}_{p,i} - \phi_{p,i}| \\ &= O_P(N^\delta T^{-1/2} + N^{(\delta-1)/2}). \end{aligned}$$

Using mean value theorem, there exists $\phi'_{p,i}$ between $\phi_{p,i}$ and $\widehat{\phi}_{p,i}$ such that

$$\begin{aligned} \left| \phi_{p,i}^h - \widehat{\phi}_{p,i}^h \right| &= \left| h \phi'^{h-1}_{p,i} (\phi_i - \widehat{\phi}_i) \right| \\ &= O_P(N^\delta T^{-1/2} + N^{(\delta-1)/2}), \end{aligned}$$

The above holds because $|\phi_{p,i}| < 1$, due to stationarity.

By (3.27) and (3.28), we can write

$$\begin{aligned} \left| \mathbf{f}_{p,T+h}^{(i)} - \widehat{\mathbf{f}}_{p,T+h|T}^{(i)} \right| &= \left| \phi_{p,i}^h \mathbf{f}_{p,T}^{(i)} - \widehat{\phi}_{p,i}^h \widehat{\mathbf{f}}_{p,T}^{(i)} + \sum_{j=1}^{h-1} \phi_{p,i}^j \omega_{p,T+h-j}^{(i)} \right| \\ &\leq \left| \phi_{p,i}^h - \widehat{\phi}_{p,i}^h \right| |\widehat{\mathbf{f}}_{p,T}^{(i)}| + \left| \mathbf{f}_{p,T}^{(i)} - \widehat{\mathbf{f}}_{p,T}^{(i)} \right| |\phi_{p,i}^h| + \left| \sum_{j=1}^{h-1} \phi_{p,i}^j \omega_{p,T+h-j}^{(i)} \right| \\ &= O_P \left\{ (N^\delta T^{-1/2} + N^{(\delta-1)/2}) (N^{(1+\delta)/2} T^{-1/2} + N^{(1-\delta)/2}) \right. \\ &\quad \left. + N^{(1+\delta)/2} T^{-1/2} + N^{(\delta-1)/2} + \Gamma_i \right\} \\ &= O_P(N^{(1+3\delta)/2} T^{-1} + N^{(1+\delta)/2} T^{-1/2} + \Gamma_{p,i}) \end{aligned}$$

where we define $\Gamma_{p,i}$ as the order of $\left| \sum_{j=1}^{h-1} \phi_{p,i}^j \omega_{p,T+h-j}^{(i)} \right|$. Then we combine $\mathbf{f}_{p,T+h}^{(i)}$ and $\widehat{\mathbf{f}}_{p,T+h|T}^{(i)}$ into vectors. The vectors $\mathbf{f}_{p,T+h}$ and $\widehat{\mathbf{f}}_{p,T+h|T}$ are of length r , which is a constant.

Thus, the vector norm

$$\|\mathbf{f}_{p,T+h} - \widehat{\mathbf{f}}_{p,T+h|T}\|_2 = O_P(N^{(1+3\delta)/2} T^{-1} + N^{(1+\delta)/2} T^{-1/2} + \Gamma_p),$$

where $\Gamma_p = \max_i(\Gamma_{p,i})$.

The prediction error for functional principal component scores

$$\begin{aligned}
\left\| \widehat{\boldsymbol{\beta}}_{p,T+h|T} - \boldsymbol{\beta}_{p,T+h} \right\|_2 &= \left\| \widehat{\mathbf{A}}_p \widehat{\mathbf{f}}_{p,T+h|T} - \mathbf{A}_p \mathbf{f}_{p,T+h} - \mathbf{e}_{p,T+h} \right\|_2 \\
&= \left\| \widehat{\mathbf{A}}_p - \mathbf{A}_p \right\|_2 \left\| \mathbf{f}_{p,T+h} \right\|_2 + \left\| \widehat{\mathbf{f}}_{p,T+h|T} - \mathbf{f}_{p,T+h} \right\|_2 \left\| \widehat{\mathbf{A}}_p \right\|_2 + \left\| \mathbf{e}_{p,T+h} \right\|_2 \\
&= O_p(N^{(1+\delta)/2} T^{-1/2} + N^{(1+3\delta)/2} T^{-1} + \Gamma_p + 1)
\end{aligned}$$

Recall that in Section 2, the model and predicted function are as follows

$$\begin{aligned}
\mathcal{X}_{T+h}^{(i)}(u) &= \sum_p^{p_0} [\boldsymbol{\beta}_{p,T+h}]_i \gamma_p^{(i)} + \boldsymbol{\epsilon}_{T+h}^{(i)}(u) \\
\widehat{\mathcal{X}}_{T+h|T}^{(i)}(u) &= \sum_p^{p_0} [\widehat{\boldsymbol{\beta}}_{p,T+h|T}]_i \widehat{\gamma}_p^{(i)}(u).
\end{aligned}$$

Using similar calculations as in the proof of Theorem 4,

$$\begin{aligned}
&\frac{1}{N} \sum_{i=1}^N \left\| \widehat{\mathcal{X}}_{T+h|T}^{(i)}(u) - \mathcal{X}_{T+h}^{(i)}(u) \right\|_2 \\
&\leq \frac{1}{N} \sum_p^{p_0} \left\{ \sqrt{\left\| \widehat{\boldsymbol{\beta}}_{p,T+h|T} - \boldsymbol{\beta}_{p,T+h} \right\|_2^2 \sum_{i=1}^N \left[\left\| \widehat{\gamma}_p^{(i)} - \gamma_p^{(i)} \right\|_2^2 + \left\| \gamma_p^{(i)} \right\|_2^2 + 2 \left\| \widehat{\gamma}_p^{(i)} - \gamma_p^{(i)} \right\|_2 \left\| \gamma_p^{(i)} \right\|_2 \right]} \right. \\
&\quad \left. + \sqrt{\left\| \boldsymbol{\beta}_{p,T+h} \right\|_2^2 \sum_{i=1}^N \left\| \widehat{\gamma}_p^{(i)} - \gamma_p^{(i)} \right\|_2^2} \right\} + \frac{1}{N} \sum_{i=1}^N \left\| \boldsymbol{\epsilon}_{T+h}^{(i)}(u) \right\|_2 \\
&= O_p \left(\frac{1}{N} \sqrt{(N^{(1+\delta)} T^{-1} + N^{(1+3\delta)} T^{-2} + \Gamma^2 + 1) N + N^{1-\delta} N q^3 T^{-1/2}} \right) \\
&= O_p(N^{\delta/2} T^{-1/2} + N^{-1/2} \Gamma + N^{-1/2} + N^{-\delta/2} q^{3/2} T^{-1/2}) \\
&= o_p(1) + O_p(N^{-1/2} \Gamma),
\end{aligned}$$

where $\Gamma = \max_p(\Gamma_p)$.

The proof is complete.

□

Factor-Augmented Smoothing Model for Functional data

4.1 Introduction

One main challenge in functional data analysis (FDA) lies in the fact that we are not able to observe functional curves directly, but only the discrete points, which are often contaminated by measurement errors. Moreover, when the number of discrete points is much larger than the number of curves, we encounter the so-called curse of dimensionality. We address these challenges by introducing a factor-augmented smoothing technique.

We denote a random sample of n functional data as $\mathcal{X}_i(u), i = 1, \dots, n$, and $u \in \mathcal{I} \subset \mathbb{R}$, where \mathcal{I} is a compact interval on the real line \mathbb{R} . In practice, the observed data are discrete points and are often contaminated by noise or measurement error. We use Y_{ij} to represent the j th observation on the i th subject; the observed data can then be expressed as a "signal plus noise" model:

$$Y_{ij} = \mathcal{X}_i(u_j) + \eta_{ij}, \quad j = 1, \dots, p; \quad i = 1, \dots, n.$$

We use $\mathcal{X}_i(u_j)$ to denote the realization of the j th discrete point on the curve $\mathcal{X}_i(\cdot)$, and η_{ij} is the noise or measurement error. We assume that measurement error only takes place

where the measurements are taken; thus, the error η_{ij} is a multivariate term of dimension p . Though in practice, the signal function component $X_i(u_j)$ is of the same p dimension, it differs from η_{ij} in nature. Although functions are potentially infinite-dimensional, we may impose smoothing assumptions on the functions, which usually implies functions possess one or more derivatives. This smoothness feature is used to separate the functions from the measurement errors - a procedure called functional smoothing.

When the variance of the noise level is a tiny fraction of the variance of the function, we say the signal-to-noise ratio is high. In this case, classic smoothing tools apply to functional data, including kernel methods, (Wand & Jones 1995); local polynomial smoothing (Fan & Gijbels 1996), and spline smoothing (Wahba 1990, Eubank 1999, Green & Silverman 1999). With pre-smoothed functions, estimates such as mean and covariance functions can be further obtained. More recent studies on functional smoothing approaches include Cai & Yuan (2011), Yao & Li (2013) and Zhang & Wang (2016). In this article, we apply basis smoothing to the functions $\mathcal{X}_i(u)$; that is, we represent $\mathcal{X}_i(u)$ as $\mathcal{X}_i(u) = \sum_{k=1}^K c_{ik}\phi_k(u)$, where $\phi_k(u)$ are the basis functions and the c_{ik} are the smoothing coefficients. The smoothing model then becomes

$$Y_{ij} = \sum_{k=1}^K c_{ik}\phi_k(u) + \eta_{ij}, \quad j = 1, \dots, p; \quad i = 1, \dots, n.$$

When the signal-to-noise level is low, smoothing tools may not be adequate in removing the measurement error and may cause inefficient estimation of the smoothing coefficients. Let us take a further look at the measurement error η_{ij} . In FDA, it is often the case that the number of discrete points p on each subject is large compared with the sample size n . Hence the term η_{ij} is a high-dimensional component. In this case, the observed data are in fact a mixture of functional data and high-dimensional data. The existence of the large measurement error η_{ij} raises the problem of the curse of dimensionality, which

naturally calls for the application of dimension reduction models to η_{ij} . Many studies have been conducted on various dimension reduction techniques for high-dimensional data; among these, factor models are widely used (Fan et al. 2008, Lam et al. 2011).

We propose using a factor model for the measurement error term. The high-dimensional measurement error is assumed to be driven by a small number of latent factors.

$$\eta_{ij} = \mathbf{a}_j^\top \mathbf{f}_i + \epsilon_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, p,$$

where $\mathbf{f}_i \in \mathbb{R}^r$ are the unobserved factors, $\mathbf{a}_j \in \mathbb{R}^r$ are the unobserved factor loadings, r is the number of latent factors, and ϵ_{ij} are idiosyncratic errors with mean zero. Thus, the observed data Y_{ij} can be written as the sum of two components:

$$Y_{ij} = \sum_{k=1}^K c_{ik} \phi_k(u) + \mathbf{a}_j^\top \mathbf{f}_i + \epsilon_{ij}, \quad i = 1, \dots, n, j = 1, \dots, p.$$

In the following, we refer to η_{ij} as the factor model term and ϵ_{ij} as the error term.

Since the latent factors are unobserved, we propose an iterative approach to simultaneously estimate the smooth function and the factors. Principal component analysis (PCA) is used as a tool to solve the factor model and penalized least squares is applied to construct the estimator for the smoothing coefficient c_{ik} . We also establish the asymptotic theories of the smoothing coefficient estimator. The consistency of the estimator is proved, as well as the asymptotic distribution of the estimator. The interplay between the smooth component and the factor model component is manifested; in particular, the smoothing estimator is projected onto the orthogonal complement of the space spanned by the factors \mathbf{f}_i .

The proposed estimator is asymptotically unbiased as well as efficient. When the expectation of η_{ij} is zero, ignoring the factor model component will not introduce bias to

the estimation of the smoothing coefficient, but will lead to a less efficient estimator. It is shown that the variance of the proposed estimator reaches the Cramér-Rao lower bound. When the expectation of η_{ij} is not zero, the smoothing coefficient estimator of a model without the factor model component will be both inefficient and biased. This is supported by the simulation results in Section 4.6.3.

The factor-augmented smoothing model is motivated by three considerations, as listed below. In these three cases, using the proposed model remedies the defects of the traditional smoothing model.

1. In traditional smoothing models, the measurement error η_{ij} is assumed to be small and the covariance between the measurement error in the functional dimension is ignored. This is an unrealistic assumption when the measurement error is large. With the factor model applied, we assume that the covariance in measurement error can be captured by a small number of factors. This is often reasonable in practice because the occurrence of systematic measurement error is usually driven by a few common factors.
2. When the number or shape of the smoothing basis are incorrectly identified, the smoothing model will lead to an erroneous coefficient estimate and large residuals. The proposed model deals with this problem since the unexplained variation from the mis-identification of the basis can be modeled with a small number of factors.
3. When there are step jumps in the mean level of the functions, neglecting the mean shift in smoothing models will result in large residuals at the point where the jumps take place. The changes in the mean levels of the functions come from a universal source and thus can be modeled by common factors.

In the remainder of this chapter, we elaborate on the three motivations in detail,

with examples given in Section 4.2. In Section 4.3, the model is formally stated and the iterative estimation approach is provided. We discuss the asymptotic properties of the smoothing coefficients under assumptions in Section 4.4. In Section 4.5, we consider from the statistical inference aspect of the model and propose a covariance matrix estimator for the raw data. In Section 4.6, we conduct Monte-Carlo simulations on the proposed model under various settings. A few real data examples are given in Section 4.7, and the conclusions are drawn in Section 4.8. Last, we provide the proofs of the relevant theorems and the lemmas in the Appendix.

4.2 Motivation

We introduce three examples to motivate the proposed model. In these cases, the smoothing model is not adequate to capture the signal information in the raw data. In the first example, when large measurement error exists, the residuals after smoothing are large with some extreme values. In the second example, when the basis functions are selected incorrectly, part of the variation in the functions cannot be captured by the smoothing model. In the third example, when there are step jumps in the functional data, the residuals after smoothing contain gaps. These examples demonstrate that further modeling of the residuals is needed.

4.2.1 Functional data with measurement error

Figure 4.1 shows the rainbow plots of the average daily temperature and log precipitation at 35 locations in Canada. Due to the nature of the two kinds of data, it is reasonable to assume that temperature and log precipitation are functions over time. The two graphs, however, display distinct features. In the temperature plot, though there are some perturbations, it is relatively easy to discern the shape of each curve; while in the

precipitation plot, there is a great amount of variability in the raw data, such that it is almost impossible to observe the underlying shape of the curves.

Smooth temperature data can be retrieved without much difficulty using basic smoothing techniques. The residuals are small, with constant variation. On the other hand, for the precipitation data, the residuals after smoothing exhibit a high level of variation, and even contain some extreme values. Our model endeavors to further explain the large residuals in similar cases to the precipitation data; we will show the fitting result in Section 4.7.

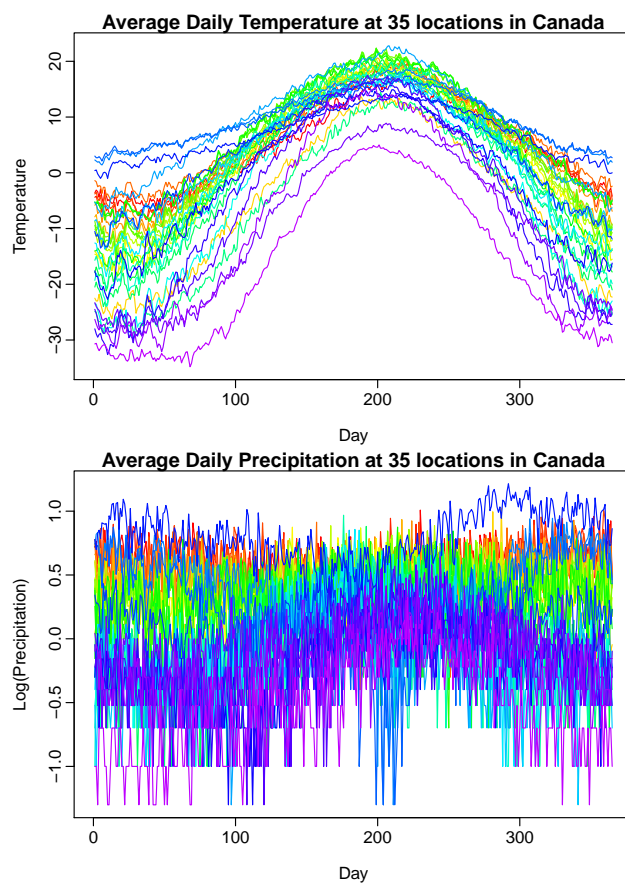


Figure 4.1: Average daily temperature and log precipitation in 35 Canadian weather stations averaged over 1960 to 1994

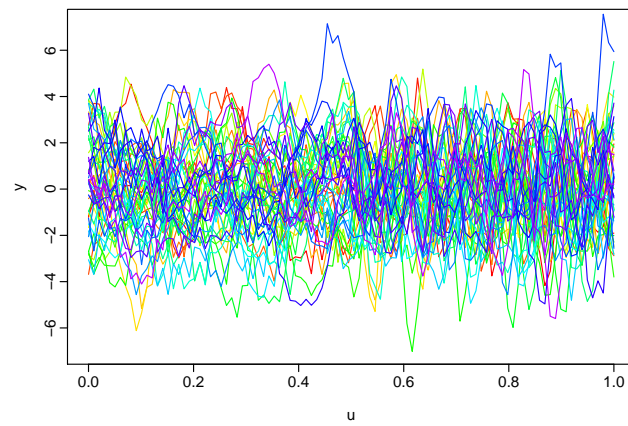
Remark 1. *Our model also serves as a pre-smooth process for further statistical inferences. In functional linear models where the smoothed curves are used as either functional response or functional covariates, it is essential that accurate estimates of pre-smoothed curves are used. Take the Canadian weather data as an example; the precipitation measured over time can be used as a functional dependent variable where the temperature measured over time can be used as the functional independent variable. In this case, we want to recover accurate smoothed weather curves before fitting the regression model.*

4.2.2 Mis-identification of the basis function

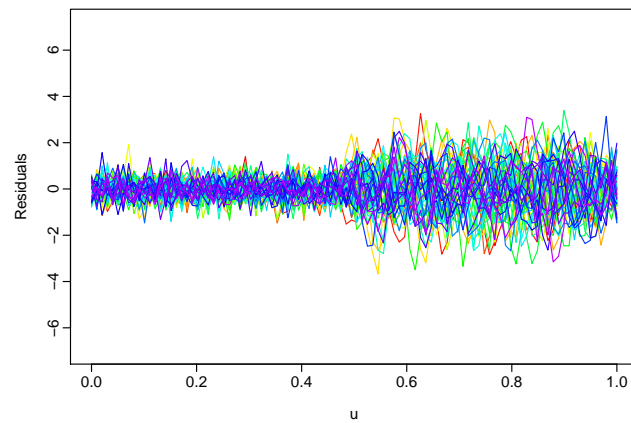
It is important to choose the appropriate basis functions in the smoothing method. In this example, we show the inadequacy of the smoothing model when the basis functions are mis-identified. We generate the functional data using basis functions with changing frequencies. The raw data are shown in Figure 4.2 (a). Fourier basis functions are used. In the second half of the data, the frequency of the Fourier basis functions increases, so the data exhibit more variation toward the right end. Suppose that we were not aware of the change in the frequencies in the basis functions, and still used the basis of the first half of the data for the whole curves. The consequences of mis-identification of the basis functions when a smoothing model is applied can be observed in Figure 4.2 (b). The residuals are large in the second half. The smoothing model fails to reduce the residuals; a factor model can be used to further model the signal hidden in the large residuals. The data generation process and further analysis can be found in Section 4.6.5.

4.2.3 Functional data with step jumps in the mean level

We provide another example of functional data with step jumps to motivate our proposed model. Suppose we observed a sample of the raw functional data as shown in Figure 4.3 (a).



(a) Raw data



(b) Residuals

Figure 4.2: Simulated sample of functional data with changing basis functions

It can be seen that there is a jump at around $u = 0.5$. The jump applies to all the data in the sample, so this sudden shift is in the mean level. We will explain how the data are generated in Section 4.6.6. The residuals after smoothing are presented in Figure 4.3 (b). The large residuals around the jump make it clear that, without measures to deal with the step jumps, smoothing itself is not enough to model these kind of data. We show in Section 4.6.6 that the proposed model applied to the same data generates smaller residuals and also has less flexibility. This is indeed one of the main goals in view of model selection.

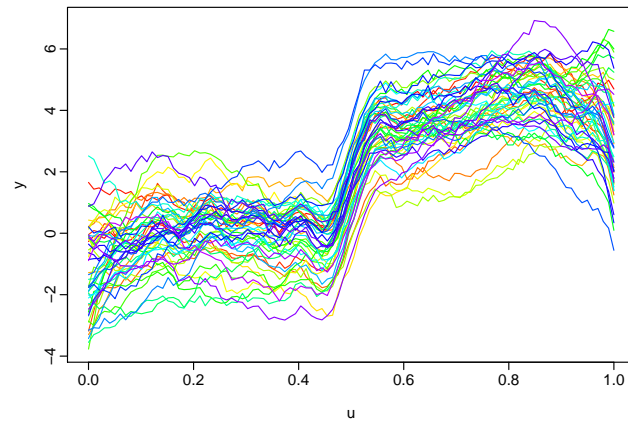
4.3 Model specification and estimation

In this section, we formally state the proposed model in Section 4.3.1 and provide the estimation method in Section 4.3.2. We first show how the smoothing coefficient c_i and the latent factors f_i are estimated separately, and then introduce an iterative approach to simultaneously find these estimates.

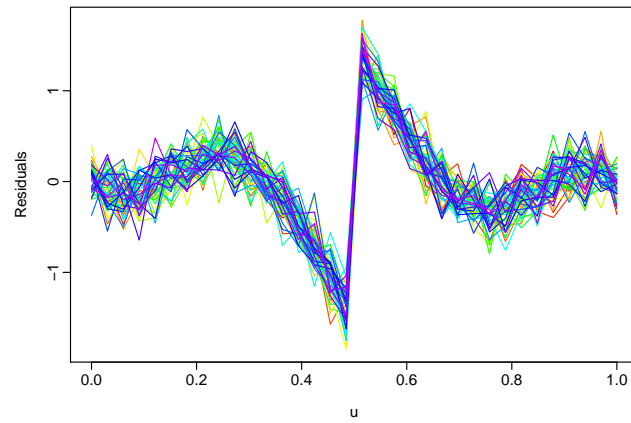
4.3.1 Model statement

We consider a sample of functional data $\mathcal{X}_i(u)$, which takes values in the space $H := L^2(\mathcal{I})$ of real-valued square integrable functions on \mathcal{I} . The space H is a Hilbert space, equipped with the inner product $\langle x, y \rangle := \int x(u)y(u)du$. The function norm is defined as $\|x\| := \langle x, x \rangle^{1/2}$. The functional nature of $\mathcal{X}_i(u)$ allows us to represent it as a linear expansion of a set of K smooth basis functions.

$$\mathcal{X}_i(u) = \sum_{k=1}^K c_{ik} \phi_k(u), \quad u \in \mathcal{I},$$



(a) Raw data



(b) Residuals

Figure 4.3: Simulated sample of functional data with step jump

where $\phi_k(u)$ is a set of common basis functions, and c_{ik} is the k th coefficient for the i th curve. Therefore, we can write the full model as

$$Y_{ij} = \sum_{k=1}^K c_{ik} \phi_k(u_j) + \eta_{ij},$$

$$\eta_{ij} = \mathbf{a}_j^\top \mathbf{f}_i + \epsilon_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, p,$$

where $\mathbf{f}_i \in \mathbb{R}^r$ are the unobserved common factors, $\mathbf{a}_j \in \mathbb{R}^r$ are the unobserved factor loadings and r is the number of factors. We call this proposed model the factor-augmented smoothing model (FASM). For the model to be identifiable, we require the following condition.

Identification Condition 1. *We require*

- (i) $\mathcal{X}_i(u)$ is independent of η_{ij} , and
- (ii) $\frac{1}{p} \sum_{j=1}^p \mathbf{a}_j \mathbf{a}_j^\top \xrightarrow{p} \Sigma_a > 0$ for some $r \times r$ matrix Σ_a , as $p \rightarrow \infty$;
 $\frac{1}{n} \sum_{i=1}^n \mathbf{f}_i \mathbf{f}_i^\top \xrightarrow{p} \Sigma_f > 0$ for some $r \times r$ matrix Σ_f , as $n \rightarrow \infty$.

The first part of the identification condition ensures the signal function component and the factor model component are independent. The second part ensures the existence of r factors, each of which makes a nontrivial contribution to the variance of η_{ij} , which in turn guarantees the identifiability between the factors and the error term ϵ_{ij} .

We treat the basis functions $\phi_k(u)$ as known, and the number K fixed. This is, of course, a simplification to accommodate for the theoretical proofs. In real data analysis, there exists a variety of choices for the basis functions and the decision can be quite subjective. For example, Fourier bases are preferred for periodic data, while spline basis systems are most commonly used for non-periodic data. Other bases include wavelet,

polynomial, and some ad-hoc basis functions. It can also be extended to allow the number of K to go to infinity.

4.3.2 Estimation

We can write the model for the i th object as

$$Y_i = \Phi c_i + A f_i + \epsilon_i, \quad (4.1)$$

where

$$Y_i = \begin{bmatrix} Y_{i1} \\ \vdots \\ Y_{ip} \end{bmatrix}, \quad c_i = \begin{bmatrix} c_{i1} \\ \vdots \\ c_{iK} \end{bmatrix}, \quad \Phi = \begin{bmatrix} \phi_1(u_1) & \dots & \phi_K(u_1) \\ \vdots & & \vdots \\ \phi_1(u_p) & \dots & \phi_K(u_p) \end{bmatrix}, \quad A = \begin{bmatrix} a_1^\top \\ \vdots \\ a_p^\top \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{ip} \end{bmatrix}.$$

Combining all the objects, we have in matrix form

$$Y = \Phi C + A F^\top + E, \quad (4.2)$$

where Y is $p \times n$ and C is a $K \times n$ matrix containing all the coefficients. The matrix $F = (f_1, \dots, f_n)^\top$ is $n \times r$ and $E = (\epsilon_1, \dots, \epsilon_n)$ is $n \times p$. Since Φ is assumed to be known, we illustrate how the parameters C, A and f are estimated in the following.

For the latent factor estimation, there is an identification problem such that $A F^\top = A U U^{-1} F^\top$ for any $r \times r$ invertible matrix U . Thus, we impose the normalization restriction on the factor loading matrix A

$$A^\top A / p = I_r. \quad (4.3)$$

It is also required for the factor matrix that $F^\top F$ is a diagonal matrix.

We propose to implement penalized least squares, where the objective function is defined as

$$\text{SSR}(c_i, \mathbf{A}, f) = \sum_{i=1}^n \left[(Y_i - \Phi c_i - \mathbf{A} f_i)^\top (Y_i - \Phi c_i - \mathbf{A} f_i) + \alpha \times \text{PEN}(\mathcal{X}_i) \right],$$

where $\text{PEN}(\mathcal{X}_i)$ is a penalty term used for regularization, and α is the tuning parameter controlling the degree of smoothness. The same α is used for all the functional observations i . This is a simplified case, where we assume a similar degree of smoothness for all the curves. The tuning parameter can be chosen by cross-validation or information criteria. We intend to penalize the "roughness" of the function term. To quantify the notion of "roughness" in a function, we use the square of the second derivative. Define the measure of roughness as

$$\text{PEN}_2(\mathcal{X}_i) = \int_{\mathcal{I}} [D^2 \mathcal{X}_i(s)]^2 ds,$$

where $D^2 \mathcal{X}_i$ denotes taking the second derivative of the function \mathcal{X}_i . The larger the tuning parameter α , the smoother the estimated functions we obtain. Further, denote

$$\Phi(u) = [\phi_1(u), \dots, \phi_K(u)]^\top. \quad (4.4)$$

Then

$$\mathcal{X}_i(u) = c_i^\top \Phi(u).$$

We can re-express the roughness penalty $\text{PEN}_2(\mathcal{X}_i)$ in matrix form as the following:

$$\begin{aligned}
 \text{PEN}_2(\mathcal{X}_i) &= \int_{\mathcal{I}} [D^2 \mathcal{X}_i(s)]^2 ds \\
 &= \int_{\mathcal{I}} [D^2 \mathbf{c}_i^\top \boldsymbol{\Phi}(s)]^2 ds \\
 &= \int_{\mathcal{I}} \mathbf{c}_i^\top D^2 \boldsymbol{\Phi}(s) D^2 \boldsymbol{\Phi}^\top(s) \mathbf{c}_i ds \\
 &= \mathbf{c}_i^\top \left[\int_{\mathcal{I}} D^2 \boldsymbol{\Phi}(s) D^2 \boldsymbol{\Phi}'(s) ds \right] \mathbf{c}_i \\
 &= \mathbf{c}_i^\top \mathbf{R} \mathbf{c}_i
 \end{aligned}$$

where $\mathbf{R} = \int_{\mathcal{I}} D^2 \boldsymbol{\Phi}(s) D^2 \boldsymbol{\Phi}'(s) ds$. The penalty term differs for each subject only by the coefficient \mathbf{c}_i .

Thus, the objective function can be written as

$$\text{SSR}(\mathbf{c}_i, \mathbf{A}, \mathbf{f}) = \sum_{i=1}^n \left[(\mathbf{Y}_i - \boldsymbol{\Phi} \mathbf{c}_i - \mathbf{A} \mathbf{f}_i)^\top (\mathbf{Y}_i - \boldsymbol{\Phi} \mathbf{c}_i - \mathbf{A} \mathbf{f}_i) + \alpha \mathbf{c}_i^\top \mathbf{R} \mathbf{c}_i \right],$$

subject to the constraint $\mathbf{A}^\top \mathbf{A} / p = \mathbf{I}_r$.

Our aim is to estimate the smoothing coefficient \mathbf{c}_i . We left multiply each term in (4.1) by a matrix to project the factor model term onto a zero matrix. Define the projection matrix

$$\mathbf{M}_A \equiv \mathbf{I}_p - \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top = \mathbf{I}_p - \mathbf{A} \mathbf{A}^\top / p. \quad (4.5)$$

Then

$$\mathbf{M}_A \mathbf{A} \mathbf{f}_i = \left(\mathbf{I}_p - \mathbf{A} \mathbf{A}^\top / p \right) \mathbf{A} \mathbf{f}_i = \left(\mathbf{A} - \mathbf{A} \mathbf{A}^\top \mathbf{A} / p \right) \mathbf{f}_i = \mathbf{0}.$$

So we estimate c_i from the projected equation

$$M_A Y_i = M_A \Phi c_i + M_A \epsilon_i.$$

The projected objective function becomes

$$\text{SSR}(c_i, A) = \sum_{i=1}^n \left[(M_A Y_i - M_A \Phi c_i)^\top (M_A Y_i - M_A \Phi c_i) + \alpha c_i^\top R c_i \right]. \quad (4.6)$$

By taking the derivative of $\text{SSR}(c_i, A)$ with respect to each c_i , we can solve for the estimator \hat{c}_i .

$$\frac{\partial \text{SSR}(c_i, A)}{\partial c_i} = (M_A Y_i - M_A \Phi c_i)^\top (M_A \Phi) + 2\alpha c_i^\top R.$$

Setting the derivative to zero and rearranging the terms, we have

$$\left(\Phi^\top M_A^\top M_A \Phi + \alpha R^\top \right) c_i = \Phi^\top M_A^\top M_A Y_i.$$

Using the fact that

$$M_A^\top M_A = \left(I_p - A A^\top / p \right)^\top \left(I_p - A A^\top / p \right) = M_A,$$

we obtain the least squares estimator for c_i given A

$$\hat{c}_i = \left(\Phi^\top M_A \Phi + \alpha R^\top \right)^{-1} \Phi^\top M_A Y_i.$$

Next, to estimate \mathbf{A} and f_i , we focus on the factor model

$$\boldsymbol{\eta}_i = \mathbf{A}f_i + \boldsymbol{\epsilon}_i,$$

and in matrix form

$$\mathbf{Z} = \mathbf{A}\mathbf{F}^\top + \mathbf{E},$$

where $\mathbf{Z} = (\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_n)$. In high dimensions, the unknown factors and loadings are typically estimated by least squares (i.e., the principal component analysis; see, e.g., Fan et al. (2008), Onatski (2012)). The least squares objective function is

$$\text{tr} \left[(\mathbf{Z} - \mathbf{A}\mathbf{F}^\top)(\mathbf{Z} - \mathbf{A}\mathbf{F}^\top)^\top \right]. \quad (4.7)$$

Minimizing the objective function with respect to \mathbf{F}^\top , we have $\mathbf{F}^\top = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{Z} = \mathbf{A}^\top \mathbf{Z} / p$, using (4.3). Substituting in (4.7), we obtain the objective function

$$\begin{aligned} & \text{tr} \left[(\mathbf{Z} - \mathbf{A}\mathbf{A}^\top \mathbf{Z} / p)(\mathbf{Z} - \mathbf{A}\mathbf{A}^\top \mathbf{Z} / p)^\top \right] \\ &= \text{tr} \left(\mathbf{Z}\mathbf{Z}^\top - \mathbf{Z}\mathbf{Z}^\top \mathbf{A}\mathbf{A}^\top / p - \mathbf{Z}\mathbf{Z}^\top \mathbf{A}\mathbf{A}^\top / p + \mathbf{A}\mathbf{A}^\top \mathbf{Z}\mathbf{Z}^\top \mathbf{A}\mathbf{A}^\top / p^2 \right) \\ &= \text{tr}(\mathbf{Z}\mathbf{Z}^\top) - \text{tr}(\mathbf{A}^\top \mathbf{Z}\mathbf{Z}^\top \mathbf{A}) / p, \end{aligned}$$

where the last equality uses (4.3) and that $\text{tr}(\mathbf{Z}\mathbf{Z}^\top \mathbf{A}\mathbf{A}^\top) = \text{tr}(\mathbf{A}^\top \mathbf{Z}\mathbf{Z}^\top \mathbf{A})$. Thus, minimizing the objective function is equivalent to maximizing $\text{tr}(\mathbf{A}^\top \mathbf{Z}\mathbf{Z}^\top \mathbf{A}) / p$. The estimator for \mathbf{A} is obtained by finding the first r eigenvectors corresponding to the r largest eigenvalues

of the matrix $\mathbf{Z}\mathbf{Z}^\top$ in descending order, where

$$\mathbf{Z}\mathbf{Z}^\top = \sum_{i=1}^n \boldsymbol{\eta}_i \boldsymbol{\eta}_i^\top = \sum_{i=1}^n (\mathbf{Y}_i - \boldsymbol{\Phi} \mathbf{c}_i)(\mathbf{Y}_i - \boldsymbol{\Phi} \mathbf{c}_i)^\top.$$

Therefore, knowing \mathbf{c}_i , we solve for $\hat{\mathbf{A}}$ using

$$\left[\frac{1}{np} \sum_{i=1}^n (\mathbf{Y}_i - \boldsymbol{\Phi} \mathbf{c}_i)(\mathbf{Y}_i - \boldsymbol{\Phi} \mathbf{c}_i)^\top \right] \hat{\mathbf{A}} = \hat{\mathbf{A}} \mathbf{V}_{np}, \quad (4.8)$$

where \mathbf{V}_{np} is a $r \times r$ diagonal matrix containing the r eigenvalues of the matrix in the square brackets in decreasing order. The coefficient $\frac{1}{np}$ is used for scaling.

Remark 2. *The number of factors r is assumed to be known in this chapter. In practice, r is selected based on some criteria regarding the eigenvalues. There has been many studies on this topic. Examples include Bai & Ng (2002), where two model selection criteria functions were proposed; Onatski (2010), where the number of factors was estimated using differenced eigenvalues; and Ahn & Horenstein (2013), where this number was selected based on the ratio of two adjacent eigenvalues.*

It can be seen that \mathbf{A} is needed to find $\hat{\mathbf{c}}_i$, and in turn \mathbf{c}_i is needed to find $\hat{\mathbf{A}}$. The final estimator $(\hat{\mathbf{c}}_i, \hat{\mathbf{A}})$ is the solution of the set of equations

$$\begin{cases} \hat{\mathbf{c}}_i = (\boldsymbol{\Phi}^\top \mathbf{M}_{\hat{\mathbf{A}}} \boldsymbol{\Phi} + \alpha \mathbf{R}^\top)^{-1} \boldsymbol{\Phi}^\top \mathbf{M}_{\hat{\mathbf{A}}} \mathbf{Y}_i, & i = 1, \dots, n \\ \left[\frac{1}{np} \sum_{i=1}^n (\mathbf{Y}_i - \boldsymbol{\Phi} \hat{\mathbf{c}}_i)(\mathbf{Y}_i - \boldsymbol{\Phi} \hat{\mathbf{c}}_i)^\top \right] \hat{\mathbf{A}} = \hat{\mathbf{A}} \mathbf{V}_{np}, \end{cases} \quad (4.9)$$

Since there is no closed-form expression of $\hat{\mathbf{A}}$ and $\hat{\mathbf{c}}_i$, we propose using numerical iterations to find the estimates. The details of these iterations are as follows:

1. Denote the initial value as $\widehat{\mathbf{A}}^{(0)}$. Using (4.9), we obtain

$$\widehat{\mathbf{c}}_i^{(0)} = \left(\mathbf{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}^{(0)}} \mathbf{\Phi} + \alpha \mathbf{R}^\top \right)^{-1} \mathbf{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}^{(0)}} \mathbf{Y}_i.$$

2. With $\widehat{\mathbf{c}}_i^{(t)}$, we substitute into the second equation of (4.9) to obtain

$$\widehat{\mathbf{A}}^{(t+1)} = (\widehat{\mathbf{a}}_1^{(t+1)}, \dots, \widehat{\mathbf{a}}_r^{(t+1)})^\top,$$

where $\widehat{\mathbf{a}}_j$ is the eigenvector of the matrix $\frac{1}{np} \sum_{i=1}^n (\mathbf{Y}_i - \mathbf{\Phi} \widehat{\mathbf{c}}_i^{(t+1)}) (\mathbf{Y}_i - \mathbf{\Phi} \widehat{\mathbf{c}}_i^{(t+1)})^\top$ corresponding to its j th largest eigenvalue.

3. With $\widehat{\mathbf{A}}^{(t+1)}$, we obtain $\widehat{\mathbf{c}}_i^{(t+1)} = \left(\mathbf{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}^{(t+1)}} \mathbf{\Phi} + \alpha \mathbf{R}^\top \right)^{-1} \mathbf{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}^{(t+1)}} \mathbf{Y}_i$ using (4.9)
4. We then repeat step 2 and 3 until $\|\widehat{\mathbf{c}}_i^{(t+1)} - \widehat{\mathbf{c}}_i^{(t)}\| < \delta$, where δ is a small positive constant.

Remark 3. In this chapter, we use $\widehat{\mathbf{A}}^{(0)} = \mathbf{0}$. This means we start by ignoring the factor model component so the initial value for the smoothing coefficient $\widehat{\mathbf{c}}_i^{(0)} = \left(\mathbf{\Phi}^\top \mathbf{\Phi} + \alpha \mathbf{R}^\top \right)^{-1} \mathbf{\Phi}^\top \mathbf{Y}_i$, which is simply the ridge estimator. The convergence of Newton's numeric iteration requires the convergence of this estimator, which in turn requires the factor model component η_{ij} to have an expectation zero.

Remark 4. Common methods for selecting the shrinkage parameter α include the Akaike's Information Criterion (AIC, Akaike (1974)) and the Bayesian Information Criterion (BIC, Schwarz (1978)) and cross-validation. In this chapter, we use the mean generalized cross-validation (mGCV) method (Golub et al. 1979). We define, at step t ,

$$mGCV^{(t)} = \frac{1}{n} \sum_{i=1}^n \frac{pSSE_i^{(t)}}{[p - df^{(t)}(\alpha)]^2}, \quad (4.10)$$

where $SSE_i^{(t)}$ is the sum of squares residual for the i th object at step t and $df^{(t)}(\alpha)$ is the equivalent degrees of freedom measure, which can be calculated as

$$df^{(t)}(\alpha) = \text{trace} \left[\Phi(\Phi^\top M_{\hat{A}^{(t)}} \Phi + \alpha \mathbf{R})^{-1} \Phi^\top M_{\hat{A}^{(t)}} \right]. \quad (4.11)$$

At each step of the iteration, the tuning parameter α is chosen by minimizing the $m\text{GCV}^{(t)}$.

After we obtain the estimates $\hat{\mathbf{A}}$ and $\hat{\mathbf{c}}_i$, the estimated coefficient matrix $\hat{\mathbf{C}}$ is constructed as $\hat{\mathbf{C}} = (\hat{\mathbf{c}}_1, \dots, \hat{\mathbf{c}}_n)$, and the estimated factor can be obtained by

$$\hat{\mathbf{F}}^\top = \hat{\mathbf{A}}^\top (\mathbf{Y} - \Phi \hat{\mathbf{C}}).$$

Finally, the signal function component can be estimated by $\hat{\mathcal{X}}_i(u) = \hat{\mathbf{c}}_i^\top \Phi(u)$, where $\Phi(u)$ is defined in (4.4).

4.4 Asymptotic theory

In this section, we study the asymptotic properties of the coefficient estimator $\hat{\mathbf{c}}_i$ with growing sample size and dimension. We state the assumptions in Section 4.4.1 and then provide the asymptotic results of $\hat{\mathbf{c}}_i$ in Section 4.4.2.

4.4.1 Assumptions

In this chapter, the norm of a vector or matrix \mathbf{U} is defined as the Frobenius norm; that is, $\|\mathbf{U}\| = [\text{tr}(\mathbf{U}^\top \mathbf{U})]^{1/2}$. We introduce the matrix

$$D_i(\mathbf{A}) \equiv \frac{1}{p} \Phi^\top M_{\mathbf{A}} \Phi - \frac{1}{p} \Phi^\top M_{\mathbf{A}} \Phi f_i^\top \left(\frac{\mathbf{F}^\top \mathbf{F}}{n} \right)^\top f_i. \quad (4.12)$$

This matrix plays an important role in this chapter. It is used in the proof of consistency of c_i , as can be found in Appendix A. The identifying condition for c_i is that $D_i(A)$ is positive definite for all i , which is stated in Assumption 11.

First, we state the assumptions.

Assumption 10.

$$\sup_u |\phi_k(u)| = O(1), \quad k = 1, \dots, K.$$

The above assumption declares that the basis functions are bounded in norm. This is quite natural as some of the most commonly used basis functions are bounded; for instance, the Fourier basis, B-spline basis, and wavelet basis functions.

Assumption 11.

$$\|c_i\| = O_p(1), \text{ for all } i$$

In this assumption, we assume the smoothing coefficients c_i are bounded uniformly for all i . This assumption is introduced to ensure the uniform consistence of the estimated coefficients \hat{c}_i .

Assumption 12. Let $\mathcal{A} = \{A : A^\top A/p = I\}$. We assume

$$\inf_{A \in \mathcal{A}} D_i(A) > 0$$

This assumption is the identification condition for c_i . The usual assumption for the least squares estimator only contains the first part of (4.12). The second part arises because of the unobservable matrices F and A .

Assumption 13. For some constant M , $\mathbb{E}\|a_j\|^4 \leq M$, $j = 1, \dots, p$ and $\mathbb{E}\|f_j\|^4 \leq M$.

Assumption 14. *The error terms $\epsilon_{ji}, j = 1, \dots, p, i = 1, \dots, n$ are i.i.d. in both directions, with $\mathbb{E}(\epsilon_{ji}) = 0$, and $\text{Var}(\epsilon_{ji}) = \sigma^2$, and $\mathbb{E}|\epsilon_{ji}|^8 \leq M$.*

Assumption 15. *ϵ_{ji} is independent of ϕ_s, f_t , and \mathbf{a}_s for all j, i, s, t .*

We require that the errors are independent in themselves and also of the functional term $\phi(u)$ and factor model terms f_i and \mathbf{a}_j . In order not to mask the main contribution of our method, we use a simplified setting on the error terms to exclude endogeneity. Nevertheless, Assumption 14 can be relaxed and our model can be easily extended to more complicated settings where correlations between the error term and the factor model terms are allowed.

Assumption 16. *The tuning parameter satisfies $\alpha = o(p)$.*

This is conventionally assumed in ridge regression (see, e.g., Knight & Fu (2000)), and assures that the asymptotic bias of the estimator is zero.

Before stating the next assumption, we introduce some notations. Let $\boldsymbol{\omega}_j, j = 1, \dots, p$ denote the j th column of the $K \times p$ matrix $\boldsymbol{\Phi}^\top \mathbf{M}_{A^0}$, and let ψ_{ik} denote the (i, k) th element of the matrix \mathbf{M}_F , where

$$\mathbf{M}_F \equiv \mathbf{I}_n - \mathbf{F} \left(\mathbf{F}^\top \mathbf{F} \right)^{-1} \mathbf{F}^\top. \quad (4.13)$$

Then, for any vector $\mathbf{b} = (b_1, \dots, b_n)^\top$, we can write

$$\frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top \mathbf{M}_{A^0} \mathbf{E} \mathbf{M}_F \mathbf{b} = \frac{1}{\sqrt{np}} \sum_i^n \sum_j^p \boldsymbol{\omega}_j \epsilon_{ji} \sum_k^n \psi_{ik} b_k \equiv \frac{1}{\sqrt{np}} \sum_i^n \sum_j^p \mathbf{x}_{ij}. \quad (4.14)$$

In (4.14), for notational simplicity, we define \mathbf{x}_{ij} as $\boldsymbol{\omega}_j \epsilon_{ji} \sum_k^n \psi_{ik} b_k$. The matrix $\boldsymbol{\Phi}^\top \mathbf{M}_{A^0} \mathbf{E} \mathbf{M}_F$ is of interest because it is the main component that contributes to the asymptotic distribution of the estimators, as shall be seen in the next section.

Let

$$\mathbf{L}_{np} \equiv \frac{\sigma^2}{np} \sum_i^n \sum_j^p \boldsymbol{\omega}_j^\top \boldsymbol{\omega}_j \left(\sum_k^n \psi_{ik} b_k \right)^2. \quad (4.15)$$

We make the following assumption.

Assumption 17. We assume there exists a $K \times K$ matrix \mathbf{L} such that

$$\mathbf{L} \equiv \lim_{n,p} \mathbf{L}_{np}, \quad (4.16)$$

where \mathbf{L}_{np} is defined in (4.15). Let v^2 be the smallest eigenvalue of the matrix \mathbf{L} defined in (4.16).

We then assume that $v^2 > 0$, and that, for all $\varepsilon > 0$,

$$\lim_{n,p \rightarrow \infty} \frac{1}{npv^2} \sum_{i=1}^n \sum_{j=1}^p \mathbb{E} \left[\|\mathbf{x}_{ij}\|^2 \mathbf{1} \left(\|\mathbf{x}_{ij}\|^2 \geq \varepsilon npv^2 \right) \right] = 0.$$

This assumption is the multivariate Lindeberg condition, which is needed in constructing the central limit theorem result in the next section. This is by no means a strong condition; for instance, when the factor model component is ignored, $\boldsymbol{\omega}_j$ is simply $\boldsymbol{\phi}_j$, and $\mathbf{x}_{ij} = \boldsymbol{\phi}_j b_i \epsilon_{ji}$. Since we assume $\boldsymbol{\phi}_j = O(1)$ in Assumption 10, the Lindeberg condition is met.

4.4.2 Asymptotic properties

We use $(\mathbf{c}_i^0, \mathbf{A}^0)$ to denote the true values of parameters. As we have mentioned previously, the identification problem of the latent factor implies that we actually use the estimator $\widehat{\mathbf{A}}$ to estimate a rotation of \mathbf{A}^0 . Based on the objective function (4.6) in Section 4.3, we use a

center-adjusted objective function defined as below.

$$S_{np}(\mathbf{c}_i, \mathbf{A}) = \frac{1}{np} \sum_{i=1}^n \left[(\mathbf{Y}_i - \Phi \mathbf{c}_i)^\top \mathbf{M}_A (\mathbf{Y}_i - \Phi \mathbf{c}_i) + \alpha \mathbf{c}_i^\top \mathbf{R} \mathbf{c}_i \right] - \frac{1}{np} \sum_{i=1}^n \epsilon_i^\top \mathbf{M}_{A^0} \epsilon_i, \quad (4.17)$$

where $\mathbf{M}_A = \mathbf{I}_p - \mathbf{A}\mathbf{A}^\top / p$, satisfying $\mathbf{A}^\top \mathbf{A}_p = \mathbf{I}_r$. The second term on the right-hand side of (4.17) does not contain the unknown \mathbf{A} and \mathbf{c}_i , so the inclusion of this term does not affect the optimization result. This term is only used for center adjusting, so that the resulting objective function has expectation zero. We estimate \mathbf{c}_i^0 and \mathbf{A}^0 by

$$(\hat{\mathbf{c}}_i, \hat{\mathbf{A}}) = \arg \min_{\mathbf{c}_i, \mathbf{A}} S_{np}(\mathbf{c}_i, \mathbf{A}). \quad (4.18)$$

In the following, we establish the theorems for the estimated coefficient matrix $\hat{\mathbf{C}}$. In Theorem 6, the consistency of the matrix $\hat{\mathbf{C}}$ is proved. In Theorem 7, we show the rate of convergence of $\hat{\mathbf{C}}$. Theorem 8 provides the asymptotic distribution of $\hat{\mathbf{C}}$.

Let $\mathbf{P}_U = \mathbf{U}(\mathbf{U}^\top \mathbf{U})^{-1} \mathbf{U}^\top$ for a matrix \mathbf{U} .

Theorem 6. *Under Assumptions 10 - 15, as $n, p \rightarrow \infty$, we have the following statements*

- (i) $\frac{1}{\sqrt{n}} \left\| \mathbf{C} - \hat{\mathbf{C}} \right\| \xrightarrow{p} 0$.
- (ii) $\left\| \mathbf{P}_{\hat{\mathbf{A}}} - \mathbf{P}_{\mathbf{A}^0} \right\| \xrightarrow{p} 0$.

We start by proving the consistency for the vector $\hat{\mathbf{c}}_i$. This consistency is uniform for all $i = 1, \dots, n$. Therefore, we could combine \mathbf{c}_i for all $i = 1, \dots, n$, and we have the result for the coefficient matrix $\hat{\mathbf{C}}$ in (i). The matrix $\hat{\mathbf{C}}$ is of dimension $K \times n$, where K is fixed and the sample size n goes to infinity, so there is a $\frac{1}{\sqrt{n}}$ scale in the result of (i). In the second part of the theorem, note that $\mathbf{P}_A = \mathbf{I}_p - \mathbf{M}_A$, where \mathbf{M}_A is the projection matrix onto the orthogonal supplement of the linear space spanned by the columns of \mathbf{A} .

Thus, $P_{\hat{A}}$ and P_{A^0} represent the spaces spanned by \hat{A} and A^0 , and we show that they are asymptotically the same in (ii).

Next, we obtain the rate of convergence.

Theorem 7. *Under Assumptions 10 - 15, if $p/n \rightarrow \rho > 0$,*

$$\left\| \sqrt{p} \frac{(\mathbf{C} - \hat{\mathbf{C}})}{\sqrt{n}} \mathbf{M}_F \right\| = O_p(1),$$

where \mathbf{M}_F is defined in (4.13).

We study the case when the dimension p and the sample size n are comparable. We achieve rate \sqrt{p} convergence, considering $\frac{\|\mathbf{C} - \hat{\mathbf{C}}\|}{\sqrt{n}}$ on the whole. It is expected that the rate of convergence for smoothing models depends on the number of discrete points p observed on each curve.

Remark 5. *The asymptotic result in Theorem 7 contains a projection matrix \mathbf{M}_F . This matrix projects $\mathbf{C} - \hat{\mathbf{C}}$ onto the space orthogonal to the factor matrix \mathbf{F} . This theorem shows the interplay between \mathbf{C} and \mathbf{F} . When \mathbf{C} and \mathbf{F} are orthogonal, $(\mathbf{C} - \hat{\mathbf{C}})\mathbf{M}_F = \mathbf{C} - \hat{\mathbf{C}}$, and we obtain the rate of convergence for $\mathbf{C} - \hat{\mathbf{C}}$. When \mathbf{C} and \mathbf{F} are not orthogonal, the inference on \mathbf{C} will be affected by the existence of the factor model component.*

We further begin to establish the limiting distribution. It is shown in Appendix A that

$$\left\| \sqrt{p} \frac{(\mathbf{C} - \hat{\mathbf{C}})}{\sqrt{n}} \mathbf{M}_F \right\| = \left\| \left(\frac{1}{p} \Phi^\top \mathbf{M}_{A^0} \Phi \right)^{-1} \frac{1}{\sqrt{np}} \Phi^\top \mathbf{M}_{A^0} \mathbf{E} \mathbf{M}_F \right\| + o_p(1).$$

The limiting distribution is constructed based on the first term on the right-hand side. Let ω_j denote the j th column of the $K \times p$ matrix $\Phi^\top \mathbf{M}_{A^0}$. We then have the following lemma.

Lemma 6. For any vector $\mathbf{b} = (b_1, \dots, b_n)^\top$,

$$\frac{1}{\sqrt{np}} \mathbf{\Phi}^\top \mathbf{M}_{A^0} \mathbf{E} \mathbf{M}_F \mathbf{b} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{L}),$$

where \mathbf{L} is defined in (4.16).

This lemma paves way for the next theorem on asymptotic normality.

Theorem 8. Under Assumptions 10 - 16, if $p/n \rightarrow \rho > 0$, we have for any vector $\mathbf{b} \in \mathbb{R}^n$

$$\sqrt{p} \left(\frac{\mathbf{C} - \widehat{\mathbf{C}}}{\sqrt{n}} \right) \mathbf{M}_F \mathbf{b} \xrightarrow{d} \mathcal{N} \left(\mathbf{0}, \mathbf{Q} (\mathbf{A}^0)^{-1} \mathbf{L} \mathbf{Q} (\mathbf{A}^0)^{-1} \right),$$

where \mathbf{M}_F is defined in Theorem 7, \mathbf{L} is defined in (4.16), and

$$\mathbf{Q} (\mathbf{A}^0) \equiv \frac{1}{p} \mathbf{\Phi}^\top \mathbf{M}_{A^0} \mathbf{\Phi}.$$

The vector \mathbf{b} comes from the same vector in Lemma 6. The asymptotic bias is zero since we assume no serial or cross-sectional correlation in the error terms. This is a simplified setting, which can be extended to allow for weak correlations in errors in both directions. In that case, the asymptotic distribution will include a non-zero bias term.

Remark 6. Theorem 8 shows that the distribution of the coefficient matrix $\widehat{\mathbf{C}}$ relies on the unobserved factor loading matrix \mathbf{A}^0 . However, we are able to find estimators for \mathbf{Q} and \mathbf{L} based on $\mathbf{M}_{\widehat{\mathbf{A}}}$:

$$\begin{aligned} \widehat{\mathbf{Q}} &= \frac{1}{p} \mathbf{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}} \mathbf{\Phi} \\ \widehat{\mathbf{L}} &= \frac{\sigma^2}{np} \sum_i^n \sum_j^p \widehat{\boldsymbol{\omega}}_j^\top \widehat{\boldsymbol{\omega}}_j \left(\sum_k^n \widehat{\psi}_{ik} b_k \right)^2, \end{aligned}$$

where $\hat{\omega}_j$ is the j th column of the $K \times p$ matrix $\Phi^\top \mathbf{M}_{\hat{\Lambda}}$ and $\hat{\psi}_{ik}$ is the (i, k) th element in the matrix $\mathbf{M}_{\hat{\Gamma}}$.

4.5 Statistical inference on covariance matrix estimation

Having presented the model estimation approach and the asymptotic properties of the estimators, we now consider statistical inference with FASM. Our model serves as a dimension reduction technique and avoids the curse of dimensionality problem, rendering making inferences from the model convenient.

Covariance estimation is fundamental in both FDA and high-dimensional data analysis. In these areas, data are of high dimensions, which brings many challenges. In FDA, the number of discrete points on each curve is often larger than the number of curves. Similarly, the dimension p of high-dimensional data is typically of the same order or larger than the sample size n . In this case, the traditional sample covariance estimator no longer works. Dimension reduction by imposing some structure on the data is one of the main ways to solve this problem (see, e.g., Wong et al. (2003), Bickel & Levina (2008) and Fan et al. (2008)). By reducing the dimension of the data with a smoothing model and a factor model in FASM, we propose an alternative covariance matrix estimator.

We consider the covariance matrix of the observed high-dimensional data \mathbf{Y}_i . Let

$$\Sigma_Y \equiv \text{cov}(\mathbf{Y}).$$

Based on the FASM where

$$\mathbf{Y}_i = \Phi \mathbf{c}_i + \mathbf{A} \mathbf{f}_i + \boldsymbol{\epsilon}_i,$$

we obtain

$$\Sigma_Y = \Phi \Sigma_c \Phi^\top + A \Sigma_f A^\top + \Sigma_\epsilon, \quad (4.19)$$

where Σ_c and Σ_f are covariance matrices of the vectors c and F respectively and Σ_ϵ denotes the error variance structure and is a diagonal matrix under Assumption 13. Based on the above equation, we have an estimator

$$\widehat{\Sigma}_Y = \Phi \widehat{\Sigma}_c \Phi^\top + \widehat{A} \widehat{\Sigma}_f \widehat{A}^\top + \widehat{\Sigma}_\epsilon, \quad (4.20)$$

where $\widehat{\Sigma}_c$ and $\widehat{\Sigma}_f$ can be calculated by

$$\begin{aligned} \widehat{\Sigma}_c &= \frac{1}{n-1} \mathbf{C} \mathbf{C}^\top - \frac{1}{n(n-1)} \mathbf{C} \mathbf{1} \mathbf{1}^\top \mathbf{C}^\top \\ \widehat{\Sigma}_f &= \frac{1}{n-1} \mathbf{F} \mathbf{F}^\top - \frac{1}{n(n-1)} \mathbf{F} \mathbf{1} \mathbf{1}^\top \mathbf{F}^\top, \end{aligned}$$

where $\mathbf{1}$ s are vectors containing ones, the dimensions of which depend on the matrices multiplied before and after the vectors. The diagonal error covariance matrix Σ_ϵ is estimated by

$$\widehat{\Sigma}_\epsilon = \text{diag} \left(n^{-1} \widehat{\mathbf{E}} \widehat{\mathbf{E}}^\top \right),$$

where $\widehat{\mathbf{E}}$ is the residual matrix calculated as $\widehat{\mathbf{E}} = \mathbf{Y} - \Phi \widehat{\mathbf{C}} - \widehat{\mathbf{A}} \mathbf{F}^\top$. This type of covariance estimator based on factor models has also been used in previous literature. For example, Fan et al. (2008) employed a multi-factor model where the factors are assumed observable, while Fan et al. (2011) considered an extension to approximate factor models where cross-sectional correlation is allowed in the error terms.

The proposed covariance matrix estimator is built using the sample covariance matrix of the coefficient \hat{c} and the factors \hat{f} . We compare the performance using mean squared error (MSE) of the proposed covariance estimator with the ordinary sample covariance estimator. When the factor structure is ignored, the sample covariance estimator is expected to have larger variance than our estimator. The advantage of the proposed estimator is shown in Section 4.6.4.

4.6 Simulation studies

In this section, we use simulated data to illustrate the superiority of the proposed model. The FASM is compared with the smoothing model in Section 4.6.1 to 4.6.3. In Section 4.6.4, we compare the performance of the covariance matrix estimator introduced in Section 4.5 with the ordinary sample covariance estimator. In Section 4.6.5, we show how the FASM performs when applied to functional data with step jumps.

4.6.1 Data generation

We generate simulated data Y_{ij} , where $i = 1, \dots, n$ and $j = 1, \dots, p$ from the following model:

$$\begin{aligned} Y_{ij} &= \mathcal{X}_i(u_j) + \eta_{ij} + \epsilon_{ji} \\ &= \sum_{k=1}^{13} c_{ik} \phi_k(u_j) + \sum_{k=1}^4 \lambda_{ik} F_{kj} + \epsilon_{ji}, \end{aligned}$$

where $\phi_k(u)$ are chosen as B-spline basis functions of order 4 and the smoothing coefficients c_{ik} are generated from $\mathcal{N}(0, 1.5^2)$. The factors F_{kj} follow $\mathcal{N}(0, 0.5^2)$ and the factor loadings $(\lambda_{i1}, \lambda_{i2}, \lambda_{i3}, \lambda_{i4})^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a 4 by 4 covariance matrix. The random error terms ϵ_{ji} follow $\mathcal{N}(0, 0.5^2)$. We set the multivariate mean term $\boldsymbol{\mu} = \mathbf{0}$ and variance

$\Sigma = \sigma^2 \mathbf{I}_4$. We adjust the value of σ^2 to control the signal-to-noise ratio. When σ^2 is large, the signal-to-noise level is low, and when σ^2 is small, the signal-to-noise level is high.

4.6.2 Estimation

The numeric iteration procedure for finding $(\hat{c}_i, \hat{A}, \hat{f})$ is introduced in Section 4.3. We compare the FASM with the smoothing model, where the factor model component is ignored. The smoothing model can be written as:

$$Y_i = \Phi c_i + \epsilon_i,$$

where the coefficient estimator is calculated as:

$$\hat{c}_i = \left(\Phi^\top \Phi + \alpha R^\top \right)^{-1} \Phi^\top Y_i, \quad i = 1, \dots, n.$$

The tuning parameter α is also chosen using mGCV, as defined in (4.10).

4.6.3 Results

We repeat the simulation setup 100 times and obtain the estimated smooth function $\hat{\mathcal{X}}_i(u) = \hat{c}_i^\top \Phi(u)$. The averaged mean squared error (aMSE) for function estimation is calculated as

$$\text{aMSE} = \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p \left[\mathcal{X}_i(u_j) - \hat{\mathcal{X}}_i(u_j) \right]^2.$$

The results are reported in Table 4.1. With the same size n , increasing the number of points p on the curve decreases the estimation error. However, with the same value for p , increasing the sample size does not decrease the estimation error. This is consistent

with the rate of convergence stated in Section 4.4, where the estimator converges with rate related to p . When σ is large, such that the signal-to-noise ratio is high, the FASM performs better than the smoothing model.

Table 4.1: The aMSE of the function estimates with different sample sizes and dimensions. The size of η is controlled by σ .

Dimension	Size of η	aMSE	
		FASM	Smoothing model
$n = 20, p = 51$	$\sigma = 0$	0.2087	0.0677
	$\sigma = 0.5$	0.1669	0.1177
	$\sigma = 0.75$	0.1774	0.1729
	$\sigma = 1$	0.2151	0.2424
$n = 20, p = 101$	$\sigma = 0$	0.1182	0.0390
	$\sigma = 0.5$	0.0719	0.0661
	$\sigma = 0.75$	0.0893	0.0979
	$\sigma = 1$	0.1163	0.1384
$n = 50, p = 51$	$\sigma = 0$	0.2101	0.0689
	$\sigma = 0.5$	0.1311	0.1207
	$\sigma = 0.75$	0.1522	0.1787
	$\sigma = 1$	0.1943	0.2518
$n = 100, p = 101$	$\sigma = 0$	0.1142	0.0392
	$\sigma = 0.5$	0.0593	0.0674
	$\sigma = 0.75$	0.0794	0.0989

Continued on next page

Dimension	Size of η	aMSE	
		FASM	Smoothing model
	$\sigma = 1$	0.1051	0.1385

4.6.4 Covariance matrix estimation

In this section, we show the finite sample performance of the covariance estimator defined in (4.20). We also calculate the regular sample covariance estimator $\hat{\Sigma}_Y^*$ using

$$\hat{\Sigma}_Y^* = \frac{1}{n-1} (\mathbf{Y} - \bar{\mathbf{Y}}) (\mathbf{Y} - \bar{\mathbf{Y}})^\top,$$

where the $p \times n$ matrix $\bar{\mathbf{Y}}$ is the sample mean matrix whose j th row elements are $\frac{1}{n} \sum_{i=1}^n Y_{ij}$.

Both estimators are compared with the population covariance matrix, which is calculated using (4.19). We calculate the estimation errors under the Frobenius norm as

$$\text{MSE} = \frac{1}{p} \left\| \hat{\Sigma}_Y - \Sigma \right\|^2.$$

We show the MSE results in Table 4.2. It can be seen that the FASM produces smaller MSE values in most cases. With the same sample size n , increasing the dimension p will decrease the estimation error significantly. However, with the same dimension p , when increasing the sample size, the MSE of FASM decreases, but not as fast as the ordinary sample covariance estimator. Thus, it is evident from the simulation results that the proposed covariance estimator performs better especially when the data are of high-dimension or p is large compared with n .

Table 4.2: The MSE of the two covariance estimators with different sample sizes and dimensions. The size of η is controlled by σ .

Dimension	Size of η	MSE	
		FASM	Sample covariance
$n = 20, p = 51$	$\sigma = 0$	0.069	0.100
	$\sigma = 0.5$	0.090	0.143
	$\sigma = 0.75$	0.128	0.198
	$\sigma = 1$	0.218	0.326
$n = 20, p = 101$	$\sigma = 0$	0.075	0.107
	$\sigma = 0.5$	0.089	0.145
	$\sigma = 0.75$	0.118	0.197
	$\sigma = 1$	0.211	0.333
$n = 50, p = 51$	$\sigma = 0$	0.030	0.042
	$\sigma = 0.5$	0.041	0.058
	$\sigma = 0.75$	0.059	0.078
	$\sigma = 1$	0.117	0.122
$n = 100, p = 101$	$\sigma = 0$	0.014	0.019
	$\sigma = 0.5$	0.019	0.027
	$\sigma = 0.75$	0.031	0.038
	$\sigma = 1$	0.066	0.062

4.6.5 Mis-identification of the basis function

We elaborate on the example presented in Section 4.2.2. We generate data from

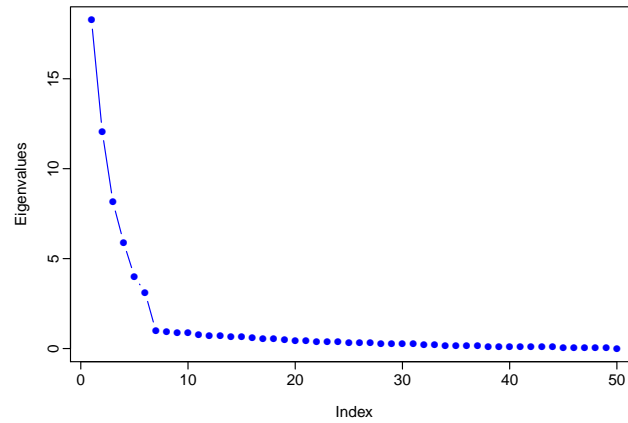
$$Y_{ij} = \sum_{k=1}^7 c_{ik} \phi_k(u_j) + \epsilon_{ji}, \quad i = 1, \dots, n, \quad j = 1, \dots, p,$$

where $\phi_k(u)$ are a set of Fourier basis functions. The first Fourier basis function $\beta_1(u)$ is the constant function; the remainder are sine and cosine pairs with integer multiples of the base period. We generate the Fourier functions with double frequencies in the second half to simulate the change in the basis functions. In particular, when $u \in [0, 0.5]$, $\phi_k(u) = 2 \sin(k\pi u)$, for $k = 2, 4, 6$, and $\phi_k(u) = 2 \cos[(k-1)\pi u]$, for $k = 3, 5, 7$, and when $u \in (0.5, 1]$, $\phi_k(u) = 2 \sin(2k\pi u)$, for $k = 2, 4, 6$, and $\phi_k(u) = 2 \cos[2(k-1)\pi u]$, for $k = 3, 5, 7$. The coefficients c_{ik} are generated from the normal distribution with mean 0 and variance 0.5^2 . The error terms are also drawn from the normal distribution with mean 0 and variance 0.5^2 . The generated Y_{ij} are shown in Figure 4.2 (a). It can be seen that the data exhibit more variation in the second half of the interval.

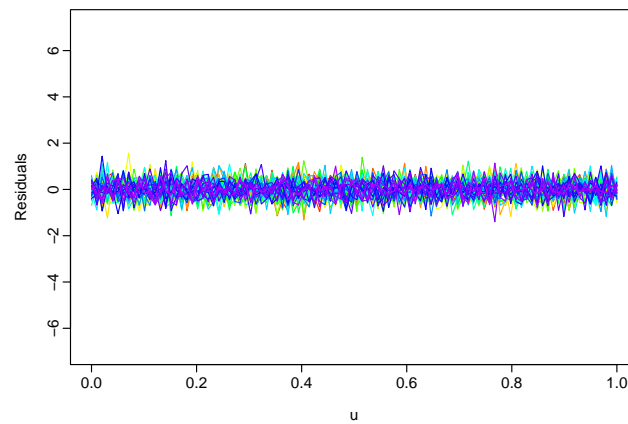
Suppose we were unaware of the change in the frequencies of the basis functions, and used the bases in the first half to fit the data on the whole interval. The residuals of the smoothing model, shown in Figure 4.2 (b), are large in the second half. When the frequency of the basis functions are mis-identified, a smoothing model with the wrong set of bases is inadequate. We conduct principal component analysis on the residuals; the eigenvalues in descending order are shown in Figure 4.4 (a). The residuals preserve a spiked structure, where most of the variation can be explained by six common factors.

We also apply FASM to the same data with the wrong set of basis functions. According to the eigenvalue scree plot, we retain six factors in the model ($r = 6$). The resulting residuals are shown in Figure 4.4 (b). The large residuals in the second part of Figure 4.2

(b) are removed. When the basis functions are mis-identified, the FASM serves as a remedy.



(a) Spikiness of the residuals



(b) Residuals of FASM

Figure 4.4: Applying FASM on the same data

4.6.6 Functional data with step jumps

We study the case where the functional data exhibit a dramatic change in the mean level within a small window. We generate data from the following model

$$Y_{ij} = \mu(u_j) + \sum_{k=1}^7 c_{ik} \phi_k(u_j) + \epsilon_{ji}, \quad i = 1, \dots, n, \quad j = 1, \dots, p,$$

where the basis functions $\phi_k(u)$ are order 4 B-spline bases. The coefficients c_{ik} come from $\mathcal{N}(0, 1.5^2)$ and the error terms from $\mathcal{N}(0, 0.5^2)$. The mean function $\mu(u)$ is generated by a linear combination of 25 B-spline basis functions. Figure 4.5 shows an example of the mean function- there is a sharp increase in the mean function at around $u = 0.5$.

The change in the mean level happens at $u = 0.5$ and δ denotes the amount of change. Figure 4.3 is generated using $\delta = 2$. Figure 4.6 compares the residuals from the smoothing model and the FASM. With the smoothing model, the residuals around the jump are large. In contrast, our model explains the large residuals around the structural break very well. In the aspect of model selection, we consider the trade-off between model fit and model flexibility. We first define a notion of degrees of freedom for the fitted model. We use the same concept as in most textbooks that the degrees of freedom measures the number of parameters estimated from the data that are required to define the model. The degrees of freedom for the smoothing model is calculated by (4.11) of the last step of convergence. The degrees of freedom for the FASM is

$$df = \text{trace} \left[\mathbf{\Phi} (\mathbf{\Phi}^\top \mathbf{M}_{\hat{\Lambda}(t)} \mathbf{\Phi} + \alpha \mathbf{R})^{-1} \mathbf{\Phi}^\top \mathbf{M}_{\hat{\Lambda}(t)} \right] + r,$$

where r is the number of factors retained in the fitted model. The larger the degrees of

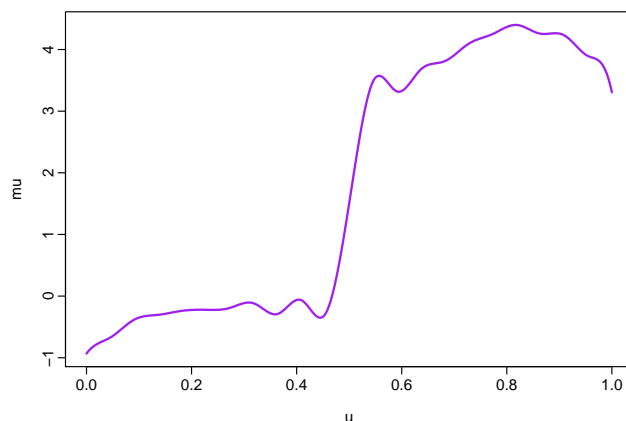
Table 4.3: The trade-off between model fit and flexibility

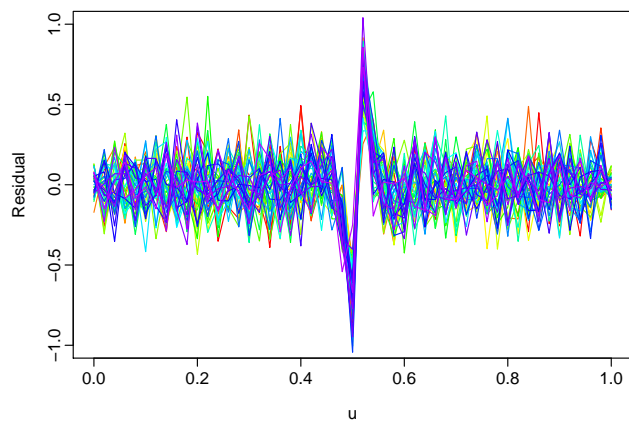
	RMSE		DF	
	Smoothing	FASM	Smoothing	FASM
$\delta = 1$	0.2045	0.1631	10.68	11.15
$\delta = 2$	0.2063	0.1640	17.59	11.03
$\delta = 3$	0.3308	0.1647	14.23	10.94

freedom, the more flexible the fitted models is. To quantify the model fitting, we use

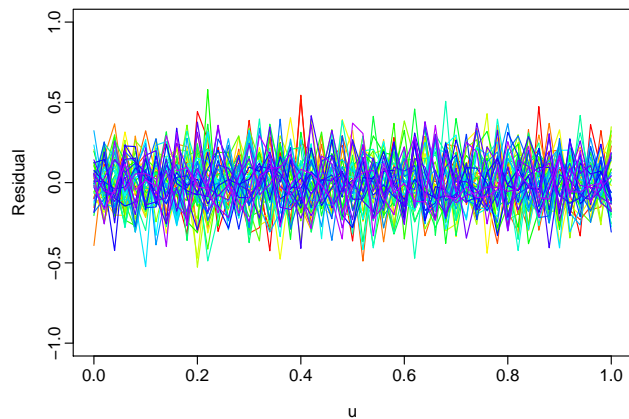
$$\text{RMSE} = \sqrt{\frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p (Y_{ij} - \hat{Y}_{ij})^2},$$

where $\hat{Y}_{ij} = \sum_{k=1}^K \hat{c}_{ik} \phi_k(u_j) + \hat{\eta}_{ij}$. In Table 4.3, we show the simulation results by changing the value of the mean shift δ . The RMSE of the FASM is always smaller than the compared model. The degrees of freedom when $\delta = 1$ are similar. When δ increases, the degrees of freedom is smaller for the proposed model. Therefore, we achieve better fit but less flexibility with the FASM.

Figure 4.5: The mean function $\mu(u)$



(a) Residuals from only applying the smoothing model



(b) Residuals from the proposed model

Figure 4.6: Residual plots of the two models

4.7 Application to weather data

In this section, we apply the FASM to two real data sets. In Section 4.7.1, we compare Canadian yearly temperature and precipitation data and demonstrate the advantages of the FASM when the measurement error is large. In Section 4.7.2, we analyze Australian daily temperature data and demonstrate the necessity of including the factor model because of the spike structure of the data.

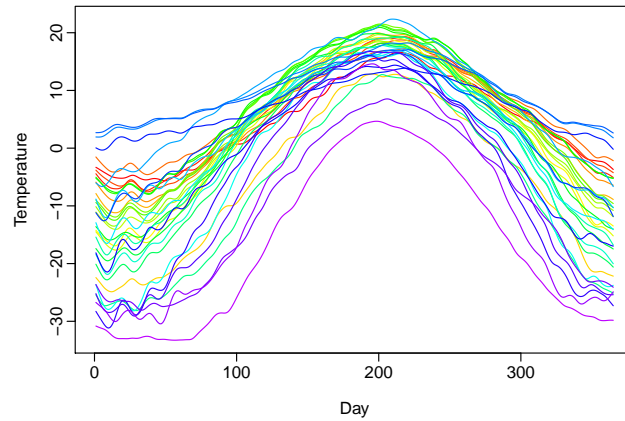
4.7.1 Canadian weather data

In Section 4.2.1, we introduced Canadian weather data. Raw observations of daily temperature and precipitation data are presented in Figure 4.1. We now apply the FASM to these two data sets.

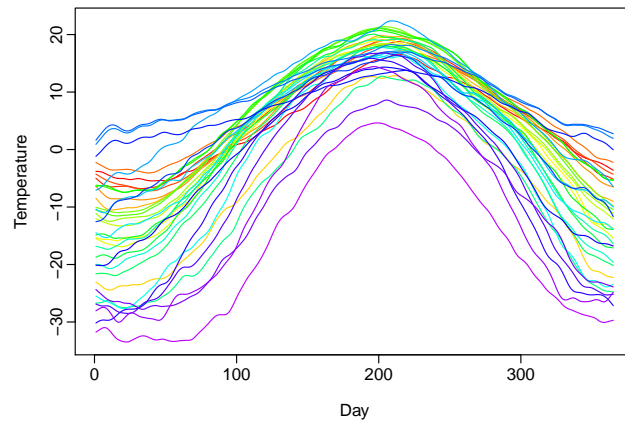
We use order 4 B-spline basis functions with knots at every data point. Thus, when the number of data points is 365, we use 367 basis functions. The number of factors r is chosen with the scree plot showing the fraction of variation explained. For temperature data, we presumed the measurement error is small. The resulting smoothed curves are shown in Figure 4.7. Compared with using the smoothing model alone, the FASM generates similar results. This meets our expectation that when measurement error does not exist, our model should work the same as a simple smoothing model.

From Section 4.2.1, we suspect large measurement errors are contained in the raw log precipitation data. We apply the two models to the log precipitation data; the resulting smoothed curves are presented in Figure 4.8. The plot on the right shows apparently smoother curves, especially at the drop in the blue curve (the 'Victoria' Station) at around day 200. Looking at the residual plots in Figure 4.9, our model mainly explains some extreme residuals left out by solely applying the smoothing model. As in Section 4.5, we also compare the RMSE and degrees of freedom of the two fitted models; they are 0.1933

and 14.41 respectively for the smoothing model and 0.1659 and 12.71 respectively for the proposed model. Thus, in terms of model selection, our model performs better across both model fit and model simplicity.

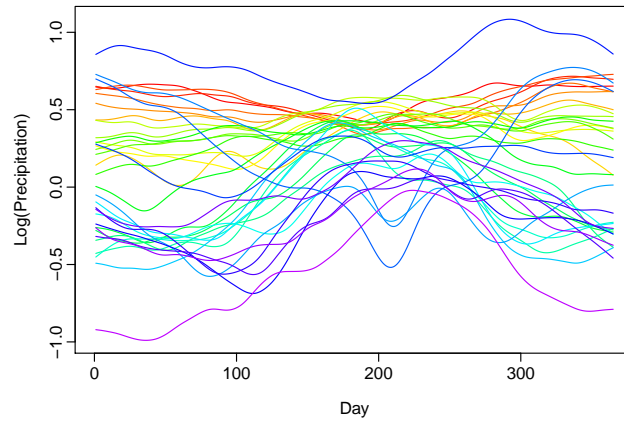


(a) Smoothed temperature curves from basis smoothing with penalty

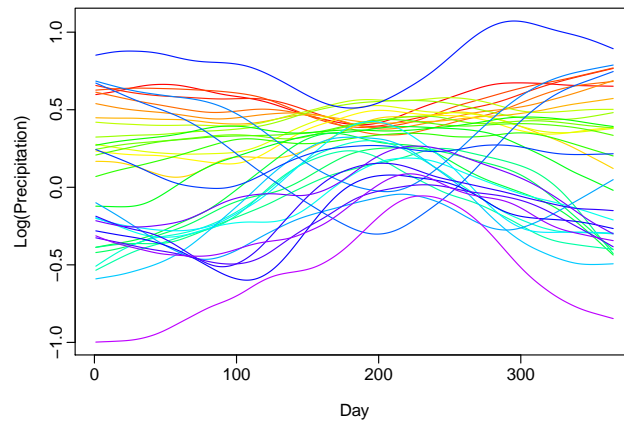


(b) Smoothed temperature curves from the FASM

Figure 4.7: Comparison between the smoothed curves

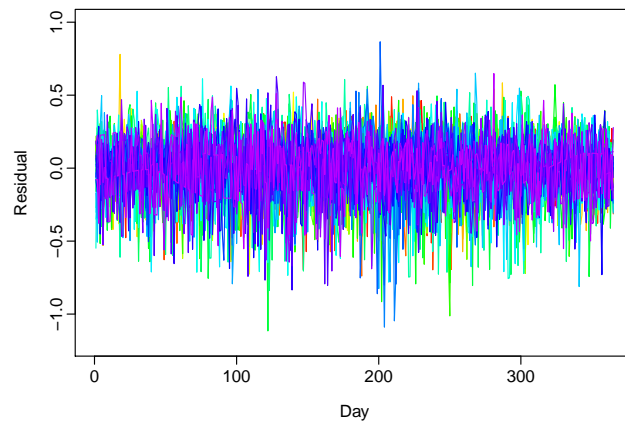


(a) Smoothed log precipitation curves from basis smoothing with penalty

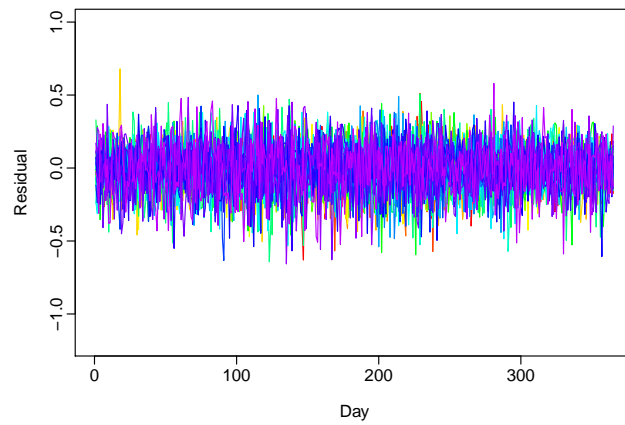


(b) Smoothed log precipitation curves from the FASM

Figure 4.8: Comparison between the smoothed curves



(a) Residuals from basis smoothing with penalty



(b) Residuals from the FASM

Figure 4.9: Comparison between the Residuals

4.7.2 Australian temperature data

In this section, we consider Friday temperature data at Adelaide airport. Data from other week days exhibit similar feature and are not shown here. The data are measured every half an hour from the year 1997 to 2007. The sample size n is 508 and the number of discrete data points from each curve p is 48. The plot of the raw data can be found in Figure 4.10 (a). It can be seen that the data are quite noisy, with extreme values in some of the curves due to large measurement error.

We use order 4 B-spline basis functions with knots at every data point. A penalized smoothing model is fitted to the data, with the tuning parameter selected to minimize the mGCV value. The residuals are shown in Figure 4.10 (b). As can be seen, the smoothing model fails to capture the extreme values contained in the raw data.

We check the spikiness of the residuals in Figure 4.10 (b) by conducting principal component analysis. The eigenvalues in descending order are shown in Figure 4.10 (c). It is evident that the first few eigenvalues are significantly larger than the rest. This means the residuals contain information that can be captured by just a few factors, which calls for a further dimension reduction model on the residuals.

As a comparison, the FASM is also applied to the data. The tuning parameter for the smoothing part is selected based on mGCV at each step of iteration. The number of factors retained in the factor model component is five. The residuals are shown in Figure 4.10 (d). The extreme values are almost all removed from the remaining residuals.

4.8 Conclusion

In this chapter, we propose a factor-augmented smoothing model for functional data. We study raw functional data, which is a mixture of functional curves and high-dimensional

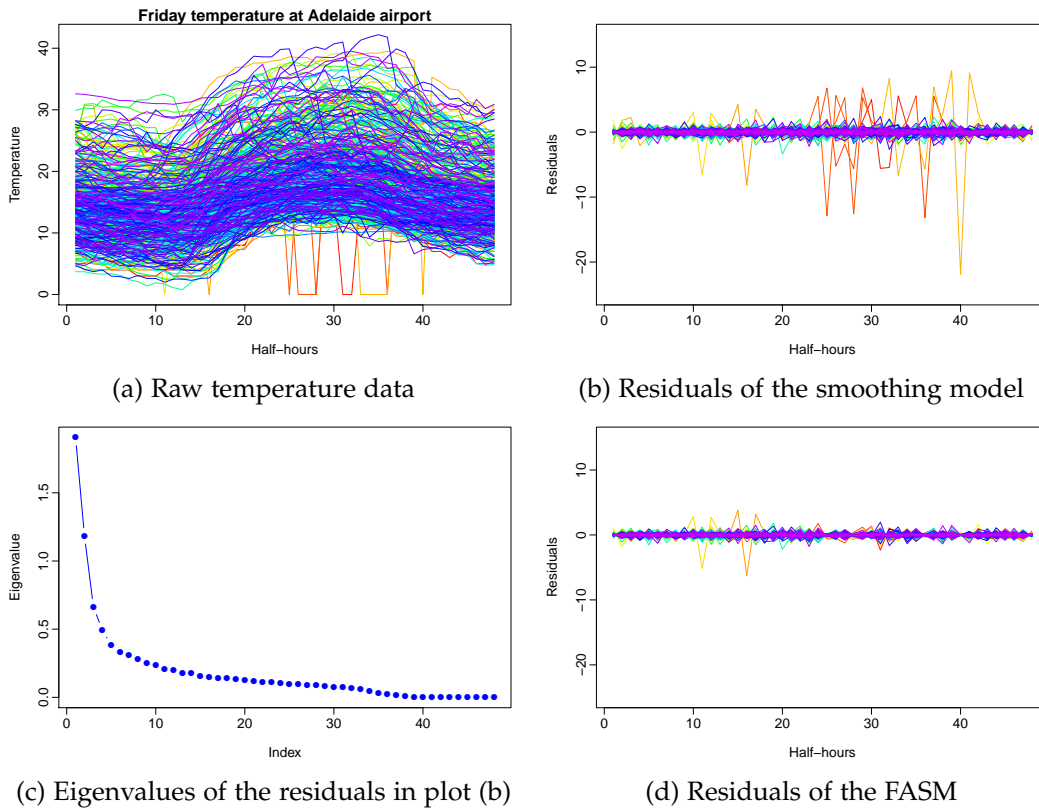


Figure 4.10: Half-hourly Friday temperature data at Adelaide airport

errors. When measurement error is large, a smoothing model alone is inadequate to capture data variation and recover the signal function component. The proposed model incorporates a factor structure into the smoothing model to further explain the large residuals. We propose a numerical iteration approach to simultaneously obtain estimates in the smoothing model and the factor model. The asymptotic distribution of the estimators are given with proofs. Our model also serves as a dimension reduction method on the functional data, easing the path to making inferences. We provide an example of the construction of a covariance estimator for the raw data. Further, we show that the model can be applied in situations where there is mis-identification in the data structure, two examples of which are the wrong selection of smoothing basis functions and the neglect of the step jumps in the mean level of the functions. The advantages of the proposed model are demonstrated in extensive simulation studies, and we also show how our model performs via an application to Canadian weather data and Australian temperature data.

4.9 Appendices

This section contains the proofs for the theorems in the main article. In Appendix A, we provide the proofs for the theorems in Section 4.4. In Appendix B, we include the results of a proposition and its proof. In Appendix C, the lemmas used for the proofs in Appendix A and B are stated as well as their proofs.

4.9.1 Appendix A

Theorem 7 is the main result of the asymptotic theories and the proof of it is lengthy. Thus, we include in the following the outlines for the proof before we show the details.

Outlines for proof of Theorem 2

In Theorem 7, we find the order of convergence of the estimated coefficient matrix $\widehat{\mathbf{C}}$. The difference between $\widehat{\mathbf{C}}$ and \mathbf{C}^0 could be written into three terms:

$$\frac{1}{p} \left(\mathbf{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}} \mathbf{\Phi} + \alpha \mathbf{R}^\top \right) (\widehat{\mathbf{C}} - \mathbf{C}^0) = \frac{1}{p} \alpha \mathbf{R}^\top \mathbf{C}^0 + \frac{1}{p} \mathbf{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}} \mathbf{A}^0 \mathbf{F}^\top + \frac{1}{p} \mathbf{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}} \mathbf{E}. \quad (4.21)$$

The term $\frac{1}{p} (\mathbf{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}} \mathbf{\Phi} + \alpha \mathbf{R}^\top)$ is $O_p(1)$. The first term on the right-hand side of (4.17) comes from the penalty and the order can be found easily from Assumption 15. The third term contains the random error matrix \mathbf{E} and the order can be found using the result in Lemma 15. The second term is the most complicated one and we show in the following proof that it could be further broken down into eight terms. We find the order of each of the eight terms using the lemmas in Appendix C. Most of the terms can be shown to be $o_p(\|\mathbf{C}^0 - \widehat{\mathbf{C}}\|)$ and thus can be omitted. Combining the remaining terms, we arrive at the result

$$\begin{aligned} (\widehat{\mathbf{C}} - \mathbf{C}^0) \mathbf{M}_F &= \mathbf{Q}^{-1} (\widehat{\mathbf{A}}) \frac{1}{p} \alpha \mathbf{R}^\top \mathbf{C}^0 + \mathbf{Q}^{-1} (\widehat{\mathbf{A}}) \frac{1}{p} \mathbf{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}} \mathbf{E} \mathbf{M}_F \\ &\quad + O_p \left(\frac{1}{\min(n, p)} \right) + O_p \left(\frac{\sqrt{n}}{p\sqrt{p}} \right) + O_p \left(\frac{1}{\sqrt{np}} \right), \end{aligned} \quad (4.22)$$

where matrix \mathbf{Q} and \mathbf{M}_F are

$$\mathbf{Q}(\widehat{\mathbf{A}}) = \frac{1}{p} \mathbf{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}} \mathbf{\Phi} \quad \mathbf{M}_F = \mathbf{I}_n - \mathbf{F} (\mathbf{F}^\top \mathbf{F})^{-1} \mathbf{F}^\top.$$

The first term on the right-hand side of (4.22) is $O_p(1)$ using the assumption on the tuning parameter α . We also show the second term is $O_p(1)$ using results from the lemmas. When n and p are of the same order, we are able to show $(\widehat{\mathbf{C}} - \mathbf{C}^0)$ projected on the matrix \mathbf{M}_F is $O_p(1)$.

Next begins the formal proofs.

Proof of Theorem 6

Proof. The concentrated objective function defined in Section 4.4.2 is

$$S_{np}(\mathbf{c}_i, \mathbf{A}) = \frac{1}{np} \sum_{i=1}^n \left[(\mathbf{Y}_i - \Phi \mathbf{c}_i)^\top \mathbf{M}_A (\mathbf{Y}_i - \Phi \mathbf{c}_i) + \alpha \mathbf{c}_i^\top \mathbf{R} \mathbf{c}_i \right] - \frac{1}{np} \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top \mathbf{M}_{A^0} \boldsymbol{\epsilon}_i$$

Assume $\mathbf{c}_i^0 = \mathbf{0}$ for simplicity without loss of generality. From $\mathbf{Y}_i = \Phi \mathbf{c}_i^0 + A^0 \mathbf{f}_i + \boldsymbol{\epsilon}_i = A^0 \mathbf{f}_i + \boldsymbol{\epsilon}_i$, we have

$$\begin{aligned} S_{np}(\mathbf{c}_i, \mathbf{A}) &= \frac{1}{np} \sum_{i=1}^n \left[(A^0 \mathbf{f}_i + \boldsymbol{\epsilon}_i - \Phi \mathbf{c}_i)^\top \mathbf{M}_A (A^0 \mathbf{f}_i + \boldsymbol{\epsilon}_i - \Phi \mathbf{c}_i) + \alpha \mathbf{c}_i^\top \mathbf{R} \mathbf{c}_i \right] - \frac{1}{np} \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top \mathbf{M}_{A^0} \boldsymbol{\epsilon}_i \\ &= \frac{1}{np} \sum_{i=1}^n \mathbf{f}_i^\top A^{0\top} \mathbf{M}_A A^0 \mathbf{f}_i + \frac{1}{np} \sum_{i=1}^n \mathbf{c}_i^\top \Phi^\top \mathbf{M}_A \Phi \mathbf{c}_i - \frac{2}{np} \sum_{i=1}^n \mathbf{f}_i^\top A^{0\top} \mathbf{M}_A \Phi \mathbf{c}_i \\ &\quad + \frac{2}{np} \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top \mathbf{M}_A A^0 \mathbf{f}_i - \frac{2}{np} \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top \mathbf{M}_A \Phi \mathbf{c}_i + \frac{1}{np} \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top (\mathbf{M}_A - \mathbf{M}_{A^0}) \boldsymbol{\epsilon}_i + \frac{\alpha}{np} \sum_{i=1}^n \mathbf{c}_i^\top \mathbf{R} \mathbf{c}_i. \end{aligned}$$

Denote the first three terms in the above equation as

$$\tilde{S}_{np}(\mathbf{c}_i, \mathbf{A}) = \frac{1}{np} \sum_{i=1}^n \mathbf{f}_i^\top A^{0\top} \mathbf{M}_A A^0 \mathbf{f}_i + \frac{1}{np} \sum_{i=1}^n \mathbf{c}_i^\top \Phi^\top \mathbf{M}_A \Phi \mathbf{c}_i - \frac{2}{np} \sum_{i=1}^n \mathbf{f}_i^\top A^{0\top} \mathbf{M}_A \Phi \mathbf{c}_i,$$

Then by Lemma 9,

$$S_{np}(\mathbf{c}_i, \mathbf{A}) = \tilde{S}_{np}(\mathbf{c}_i, \mathbf{A}) + o_p(1).$$

It is easy to see that $\tilde{S}_{np}(\mathbf{c}_i^0 = \mathbf{0}, A^0 H) = 0$ for any $r \times r$ invertible H , because $\mathbf{M}_{A^0 H} = \mathbf{M}_{A^0}$ and $\mathbf{M}_{A^0} A^0 = \mathbf{0}$.

Here we define two matrix operations before further transformations on $\tilde{S}_{np}(\mathbf{c}_i, \mathbf{A})$.

For an $m \times n$ matrix \mathbf{U} and a $p \times q$ matrix \mathbf{V} , the vectorization of \mathbf{U} is defined as

$$\text{vec}(\mathbf{U}) \equiv (u_{1,1}, \dots, u_{m,1}, u_{1,2}, \dots, u_{m,2}, u_{1,n}, \dots, u_{m,n})^\top,$$

and the Kronecker product $\mathbf{U} \otimes \mathbf{V}$ is the $pm \times qn$ block matrix defined as

$$\mathbf{U} \otimes \mathbf{V} \equiv \begin{bmatrix} u_{1,1}\mathbf{V} & \dots & u_{1,n}\mathbf{V} \\ \vdots & & \vdots \\ u_{m,1}\mathbf{V} & \dots & u_{m,n}\mathbf{V} \end{bmatrix}$$

where u_{ij} represents the element on the i th row and j th column of matrix \mathbf{U} .

Next we can further write $\tilde{S}_{n,p}(\mathbf{c}_i, \mathbf{A})$ as

$$\begin{aligned} \tilde{S}_{np}(\mathbf{c}_i, \mathbf{A}) &= \text{vec}(\mathbf{M}_A \mathbf{A}^0)^\top \left(\frac{\mathbf{F}^\top \mathbf{F}}{np} \otimes \mathbf{I}_p \right) \text{vec}(\mathbf{M}_A \mathbf{A}^0) + \frac{1}{n} \sum_{i=1}^n \mathbf{c}_i^\top \left(\frac{1}{p} \mathbf{\Phi}^\top \mathbf{M}_A \mathbf{\Phi} \right) \mathbf{c}_i \\ &\quad - \frac{1}{n} \sum_{i=1}^n 2\mathbf{c}_i^\top \left(\frac{1}{p} \mathbf{f}_i \otimes \mathbf{M}_A \mathbf{\Phi} \right) \text{vec}(\mathbf{M}_A \mathbf{A}^0). \end{aligned}$$

If we denote

$$\begin{aligned} \mathbf{P} &= \frac{1}{p} \mathbf{\Phi}^\top \mathbf{M}_A \mathbf{\Phi}, \\ \mathbf{W} &= \frac{\mathbf{F}^\top \mathbf{F}}{np} \otimes \mathbf{I}_p, \\ \mathbf{V}_i &= \frac{1}{p} \mathbf{f}_i \otimes \mathbf{M}_A \mathbf{\Phi}, \end{aligned}$$

and $\boldsymbol{\gamma} = \text{vec}(\mathbf{M}_A \mathbf{A}^0)$, then we can write

$$\begin{aligned} \tilde{S}_{np}(\mathbf{c}_i, \mathbf{A}) &= \frac{1}{n} \sum_{i=1}^n \left[\mathbf{c}_i^\top \mathbf{P} \mathbf{c}_i + \boldsymbol{\gamma}^\top \mathbf{W} \boldsymbol{\gamma} - 2 \mathbf{c}_i^\top \mathbf{V}_i^\top \boldsymbol{\gamma} \right] \\ &= \frac{1}{n} \sum_{i=1}^n \left[\mathbf{c}_i^\top \left(\mathbf{P} - \mathbf{V}_i^\top \mathbf{W}^{-1} \mathbf{V}_i \right) \mathbf{c}_i + (\boldsymbol{\gamma}^\top - \mathbf{c}_i^\top \mathbf{V}_i^\top \mathbf{W}^{-1}) \mathbf{W} (\boldsymbol{\gamma}^\top - \mathbf{W}^{-1} \mathbf{V}_i \mathbf{c}_i) \right] \\ &\equiv \frac{1}{n} \sum_{i=1}^n \left[\mathbf{c}_i^\top \mathbf{D}_i \mathbf{c}_i + \boldsymbol{\theta}_i^\top \mathbf{W} \boldsymbol{\theta}_i \right]. \end{aligned}$$

In the last equation,

$$\begin{aligned} \mathbf{D}_i &\equiv \mathbf{P} - \mathbf{V}_i^\top \mathbf{W}^{-1} \mathbf{V}_i \\ &= \frac{1}{p} \boldsymbol{\Phi}^\top \mathbf{M}_A \boldsymbol{\Phi} - \frac{1}{p} \boldsymbol{\Phi}^\top \mathbf{M}_A \boldsymbol{\Phi} \mathbf{f}_i^\top \left(\frac{\mathbf{F}^\top \mathbf{F}}{n} \right)^{-1} \mathbf{f}_i \\ \boldsymbol{\theta}_i &\equiv \boldsymbol{\gamma}^\top - \mathbf{W}^{-1} \mathbf{V}_i \mathbf{c}_i. \end{aligned}$$

By Assumption 11 and 12, the matrices \mathbf{D}_i and \mathbf{W} are positive definite for each i . Thus we have $\tilde{S}_{np}(\mathbf{c}_i, \mathbf{A}) \geq 0$. In addition, if either $\mathbf{c}_i \neq \mathbf{c}_i^0$ or $\mathbf{A} \neq \mathbf{A}^0 \mathbf{H}$, then $\tilde{S}_{np}(\mathbf{c}_i, \mathbf{A}) > 0$. Thus $\tilde{S}_{np}(\mathbf{c}_i, \mathbf{A})$ achieves its unique minimum at $(\mathbf{c}_i^0, \mathbf{A}^0)$. Thus we have

$$\hat{\mathbf{c}}_i - \mathbf{c}_i^0 = o_p(1), \quad i = 1, \dots, n$$

Next, we show $\hat{\mathbf{c}}_i$ is consistent uniformly in i .

We can write

$$\begin{aligned} S_{np}(\mathbf{c}_i, \mathbf{A}) - \tilde{S}_{np}(\mathbf{c}_i, \mathbf{A}) &= \frac{2}{np} \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top \mathbf{M}_A \mathbf{A}^0 \mathbf{f}_i - \frac{2}{np} \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top \mathbf{M}_A \boldsymbol{\Phi} \mathbf{c}_i \\ &\quad + \frac{1}{np} \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top (\mathbf{M}_A - \mathbf{M}_{A^0}) \boldsymbol{\epsilon}_i + \frac{\alpha}{np} \sum_{i=1}^n \mathbf{c}_i^\top \mathbf{R} \mathbf{c}_i. \end{aligned}$$

Using Taylor's expansion at \mathbf{c}_i ,

$$\begin{aligned} S_{np}(\mathbf{c}_i, \mathbf{A}) - \tilde{S}_{np}(\mathbf{c}_i, \mathbf{A}) &= \frac{2}{np} \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top \mathbf{M}_A \mathbf{A}^0 \mathbf{f}_i - \frac{2}{np} \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top \mathbf{M}_A \boldsymbol{\Phi} \mathbf{c}_i^0 \\ &\quad + \frac{1}{np} \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top (\mathbf{M}_A - \mathbf{M}_{A^0}) \boldsymbol{\epsilon}_i + \frac{\alpha}{np} \sum_{i=1}^n \mathbf{c}_i^{0\top} \mathbf{R} \mathbf{c}_i^0 \\ &\quad + \left(-\frac{2}{np} \boldsymbol{\epsilon}_i^\top \mathbf{M}_A \boldsymbol{\Phi} + \frac{2\alpha}{np} \mathbf{c}_i^{0\top} \mathbf{R} \right) (\mathbf{c}_i - \mathbf{c}_i^0) + \Delta, \end{aligned}$$

where Δ denotes the small order terms. Then we have

$$\begin{aligned} \left(-\frac{2}{np} \boldsymbol{\epsilon}_i^\top \mathbf{M}_A \boldsymbol{\Phi} + \frac{2\alpha}{np} \mathbf{c}_i^{0\top} \mathbf{R} \right) (\mathbf{c}_i - \mathbf{c}_i^0) &= S_{np}(\mathbf{c}_i, \mathbf{A}) - \tilde{S}_{np}(\mathbf{c}_i, \mathbf{A}) - \frac{2}{np} \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top \mathbf{M}_A \mathbf{A}^0 \mathbf{f}_i \\ &\quad + \frac{2}{np} \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top \mathbf{M}_A \boldsymbol{\Phi} \mathbf{c}_i^0 + \frac{1}{np} \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top (\mathbf{M}_A - \mathbf{M}_{A^0}) \boldsymbol{\epsilon}_i \\ &\quad - \frac{\alpha}{np} \sum_{i=1}^n \mathbf{c}_i^{0\top} \mathbf{R} \mathbf{c}_i^0 + \Delta \end{aligned} \quad (4.23)$$

In the above equation, the right-hand side is $o_p(1)$ uniformly in i . This is because $S_{np}(\mathbf{c}_i, \mathbf{A}) - \tilde{S}_{np}(\mathbf{c}_i, \mathbf{A}) = o_p(1)$ and $S_{np}(\mathbf{c}_i, \mathbf{A})$ and $\tilde{S}_{np}(\mathbf{c}_i, \mathbf{A})$ both consist of summations over i . Furthermore, all other terms on the right-hand side are $o_p(1)$ as proved in Lemma 9 and all contain summations over i . On the left-hand side of (4.23), $-\frac{2}{np} \boldsymbol{\epsilon}_i^\top \mathbf{M}_A \boldsymbol{\Phi}$ is $o_p(1)$ uniformly because $\mathbb{E} \left(\frac{1}{np} \|\boldsymbol{\epsilon}_i^\top \mathbf{M}_A \mathbf{c}_i\| \right) = o(1)$ as shown in Lemma 9 (ii). Moreover, the term $\frac{2\alpha}{np} \mathbf{c}_i^{0\top} \mathbf{R}$ is also $o_p(1)$ uniformly using Lemma 9 (iv) and that we assume \mathbf{c}_i are bounded uniformly in Assumption 11. This leads us to the result that

$$\hat{\mathbf{c}}_i - \mathbf{c}_i^0 = o_p(1), \quad \text{uniformly for all } i = 1, \dots, n$$

Combining the i , we have

$$\frac{\|\widehat{\mathbf{C}} - \mathbf{C}^0\|}{\sqrt{n}} = o_p(1).$$

To prove part (ii), note that the centred objective function satisfies $S_{np}(\mathbf{c}_i^0 = \mathbf{0}, \mathbf{A}^0) = 0$ and, by definition in (4.18), we have $S_{np}(\widehat{\mathbf{c}}_i, \widehat{\mathbf{A}}) \leq 0$. Therefore,

$$0 \geq S_{np}(\widehat{\mathbf{c}}_i, \widehat{\mathbf{A}}) = \widetilde{S}_{np}(\widehat{\mathbf{c}}_i, \widehat{\mathbf{A}}) + o_p(1).$$

Combined with $\widetilde{S}_{np}(\widehat{\mathbf{c}}_i, \widehat{\mathbf{A}}) \geq 0$, it must be true that

$$\widetilde{S}_{np}(\widehat{\mathbf{c}}_i, \widehat{\mathbf{A}}) = o_p(1).$$

This implies that

$$\frac{1}{np} \sum_{i=1}^n \mathbf{F}_i^\top \mathbf{A}^{0\top} \mathbf{M}_{\widehat{\mathbf{A}}} \mathbf{A}^0 \mathbf{F}_i = \text{tr} \left[\frac{\mathbf{A}^{0\top} \mathbf{M}_{\widehat{\mathbf{A}}} \mathbf{A}^0}{p} \frac{\mathbf{F}^\top \mathbf{F}}{n} \right] = o_p(1).$$

Since $\mathbf{F}^\top \mathbf{F}/n = O_p(1)$, it must be true that

$$\frac{\mathbf{A}^{0\top} \mathbf{M}_{\widehat{\mathbf{A}}} \mathbf{A}^0}{p} = \frac{\mathbf{A}^{0\top} \mathbf{A}^0}{p} - \frac{\mathbf{A}^{0\top} \widehat{\mathbf{A}}}{p} \frac{\widehat{\mathbf{A}}^\top \mathbf{A}^0}{p} = o_p(1). \quad (4.24)$$

By Assumption 13, $\mathbf{A}^{0\top} \mathbf{A}^0/p$ is invertible. Thus $\mathbf{A}^{0\top} \widehat{\mathbf{A}}/p$ is also invertible. Next,

$$\|\mathbf{P}_{\widehat{\mathbf{A}}} - \mathbf{P}_{\mathbf{A}^0}\|^2 = \text{tr}[(\mathbf{P}_{\widehat{\mathbf{A}}} - \mathbf{P}_{\mathbf{A}^0})^2] = 2\text{tr}(\mathbf{I}_r - \widehat{\mathbf{A}}^\top \mathbf{P}_{\mathbf{A}^0} \widehat{\mathbf{A}}/p).$$

But (4.24) implies $\widehat{\mathbf{A}}^\top \mathbf{P}_{\mathbf{A}^0} \widehat{\mathbf{A}}/p \rightarrow \mathbf{I}_r$, which means $\|\mathbf{P}_{\widehat{\mathbf{A}}} - \mathbf{P}_{\mathbf{A}}\| \rightarrow 0$. □

Proof of Theorem 7

Proof. Writing the first equation in (4.9) in matrix notation, we have

$$\hat{C} = \left(\Phi^\top M_{\hat{A}} \Phi + \alpha R^\top \right)^{-1} \Phi^\top M_{\hat{A}} Y. \quad (4.25)$$

Substitute $Y = \Phi C^0 + A^0 f^\top + E$ into (4.25) and subtract the matrix C^0 on both sides, we get

$$\begin{aligned} \hat{C} - C^0 &= \left[\left(\Phi^\top M_{\hat{A}} \Phi + \alpha R^\top \right)^{-1} \Phi^\top M_{\hat{A}} \Phi - I_K \right] C^0 \\ &\quad + \left(\Phi^\top M_{\hat{A}} \Phi + \alpha R^\top \right)^{-1} \Phi^\top M_{\hat{A}} A^0 F^\top \\ &\quad + \left(\Phi^\top M_{\hat{A}} \Phi + \alpha R^\top \right)^{-1} \Phi^\top M_{\hat{A}} E, \end{aligned}$$

or

$$\frac{1}{p} \left(\Phi^\top M_{\hat{A}} \Phi + \alpha R^\top \right) (\hat{C} - C^0) = \frac{1}{p} \alpha R^\top C^0 + \frac{1}{p} \Phi^\top M_{\hat{A}} A^0 F^\top + \frac{1}{p} \Phi^\top M_{\hat{A}} E \quad (4.26)$$

We first look at the second term on the right-hand side of (4.26). Recall that $M_{\hat{A}} = I_p - \hat{A} \hat{A}^\top / p$. We have $M_{\hat{A}} \hat{A} = \mathbf{0}$. Thus

$$M_{\hat{A}} A^0 = M_{\hat{A}} \left(A^0 - \hat{A} H^{-1} + \hat{A} H^{-1} \right) = M_{\hat{A}} \left(A^0 - \hat{A} H^{-1} \right),$$

where H is defined in (4.41). Using (4.45), it follows that

$$\begin{aligned} \frac{1}{p} \Phi^\top M_{\hat{A}} A^0 F^\top &= -\frac{1}{p} \Phi^\top M_{\hat{A}} (I_1 + \dots + I_8) \left(\frac{A^{0\top} \hat{A}}{p} \right)^{-1} \left(\frac{F^\top F}{n} \right)^{-1} F^\top \\ &\equiv J_1 + \dots + J_8. \end{aligned} \quad (4.27)$$

In the following, we calculate the order for each from $J1$ to $J8$. Note that $I1$ to $I8$ are defined in (4.43). Before we begin, for simplicity denote

$$\mathbf{G} \equiv \left(\frac{\mathbf{A}^{0\top} \hat{\mathbf{A}}}{p} \right)^{-1} \left(\frac{\mathbf{F}^\top \mathbf{F}}{n} \right)^{-1}. \quad (4.28)$$

We prove in Lemma 10 that $\mathbf{G} = O_p(1)$. We also use the fact that $\|\mathbf{M}_{\hat{\mathbf{A}}}\| = O_p(1)$. Now

$$J1 = -\frac{1}{p} \boldsymbol{\Phi}^\top \mathbf{M}_{\hat{\mathbf{A}}} (I1) \mathbf{G} \mathbf{F}^\top. \quad (4.29)$$

Since $I1 = O_p\left(\frac{\sqrt{p}}{n} \|\mathbf{C} - \hat{\mathbf{C}}\|^2\right)$, using the result from Lemma 7 (i), the term $J1$ is bounded in norm by $O_p\left(\frac{1}{\sqrt{n}} \|\mathbf{C} - \hat{\mathbf{C}}\|^2\right)$. Thus it is also $o_p\left(\|\mathbf{C} - \hat{\mathbf{C}}\|\right)$.

$$J2 = -\frac{1}{p} \boldsymbol{\Phi}^\top \mathbf{M}_{\hat{\mathbf{A}}} (I2) \left(\frac{\mathbf{A}^{0\top} \hat{\mathbf{A}}}{p} \right)^{-1} \left(\frac{\mathbf{F}^\top \mathbf{F}}{n} \right)^{-1} \mathbf{F}^\top \quad (4.30)$$

$$= \frac{1}{p} \boldsymbol{\Phi}^\top \mathbf{M}_{\hat{\mathbf{A}}} \boldsymbol{\Phi} (\hat{\mathbf{C}} - \mathbf{C}) \mathbf{F} (\mathbf{F}^\top \mathbf{F})^{-1} \mathbf{F}^\top. \quad (4.31)$$

For the term $J2$, since it is not a small order term, we keep it as what it is.

Now consider

$$\begin{aligned} J3 &= -\frac{1}{p} \boldsymbol{\Phi}^\top \mathbf{M}_{\hat{\mathbf{A}}} (I3) \left(\frac{\mathbf{A}^{0\top} \hat{\mathbf{A}}}{p} \right)^{-1} \left(\frac{\mathbf{F}^\top \mathbf{F}}{n} \right)^{-1} \mathbf{F}^\top \\ &= \frac{1}{np^2} \boldsymbol{\Phi}^\top \mathbf{M}_{\hat{\mathbf{A}}} \boldsymbol{\Phi} (\hat{\mathbf{C}} - \mathbf{C}) \mathbf{E}^\top \hat{\mathbf{A}} \mathbf{G} \mathbf{F}^\top \end{aligned} \quad (4.32)$$

We take $\mathbf{E}^\top \hat{\mathbf{A}} = \mathbf{E}^\top (\hat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}) + \mathbf{E}^\top \mathbf{A}^0 \mathbf{H}$, where the order of each term can be found in Lemma 11 (i). Again using the result of Lemma 7 (i) and (iii), it can be shown that $J3$ is $o_p\left(\|\mathbf{C} - \hat{\mathbf{C}}\|\right)$.

Next

$$\begin{aligned} \|J4\| &= \left\| -\frac{1}{p} \Phi M_{\hat{A}} I4 \left(\frac{A^{0\top} \hat{A}}{p} \right)^{-1} \left(\frac{F^\top F}{n} \right)^{-1} F^\top \right\| \\ &= O_p \left(\frac{M_{\hat{A}} A^0}{\sqrt{p}} \|C - \hat{C}\| \right). \end{aligned} \quad (4.33)$$

Using Proposition 2, we have $\frac{1}{\sqrt{p}} M_{\hat{A}} A^0 = M_{\hat{A}} \frac{1}{\sqrt{p}} (A^0 - \hat{A}H^{-1}) = o_p(1)$, Thus, $\|J4\| = o_p(\|C - \hat{C}\|)$.

It can also be proven that $\|J5\| = o_p(\|C - \hat{C}\|)$.

Then we consider

$$\begin{aligned} J6 &= -\frac{1}{p} \Phi^\top M_{\hat{A}} I6 G F^\top \\ &= -\frac{1}{np^2} \Phi^\top M_{\hat{A}} A^0 F^\top E^\top \hat{A} G F^\top \\ &= -\frac{1}{np^2} \Phi^\top M_{\hat{A}} (A^0 - \hat{A}H^{-1}) F^\top E^\top \hat{A} G F^\top, \end{aligned}$$

where the last equation comes from $M_{\hat{A}} \hat{A}H^{-1} = \mathbf{0}$. Now

$$\begin{aligned} \|F^\top E^\top \hat{A}\| &= \|E^\top \hat{A}\| \\ &\leq \|E^\top (\hat{A} - A^0 H)\| + \|E^\top A^0 H\| \\ &= O_p \left(\frac{p}{\min(\sqrt{n}, \sqrt{p})} \|C - \hat{C}\| \right) + O_p(\sqrt{n}) + O_p \left(\frac{p}{\sqrt{n}} \right) + O_p(\sqrt{np}) \\ &= O_p \left(\frac{p}{\sqrt{n}} \|C - \hat{C}\| \right) + O_p \left(\frac{p}{\sqrt{n}} \right) + O_p(\sqrt{np}), \end{aligned}$$

using Lemma 11. Thus,

$$\begin{aligned}
\|J6\| &\leq \left\| \frac{1}{np^2} \Phi^\top M_{\hat{A}} (A^0 - \hat{A}H^{-1}) \right\| \left\| F^\top E^\top \hat{A} \right\| \|G\| \|F^\top\| \\
&= -\frac{1}{np^2} \times \left[O_p \left(\frac{p}{\sqrt{n}} \|C - \hat{C}\| \right) + O_p \left(\frac{p}{\min(n, p)} \right) \right] \\
&\times \left[O_p \left(\frac{p}{\sqrt{n}} \|C - \hat{C}\| \right) + O_p \left(\frac{p}{\sqrt{n}} \right) + O_p(\sqrt{np}) \right] \times O_p(\sqrt{n}) \\
&= o_p \left(\|C - \hat{C}\| \right) + O_p \left(\frac{1}{n^2} \right) + O_p \left(\frac{1}{n\sqrt{p}} \right) + O_p \left(\frac{1}{np} \right) + O_p \left(\frac{1}{p\sqrt{p}} \right), \quad (4.34)
\end{aligned}$$

where in the first equation, Proposition 2 is used, and the second equation is a result of calculation on the orders. Next

$$\begin{aligned}
J7 &= -\frac{1}{p} \Phi^\top M_{\hat{A}} I7 G F^\top \\
&= -\frac{1}{np} \Phi^\top M_{\hat{A}} E F \left(\frac{F^\top F}{n} \right)^{-1} F^\top. \quad (4.35)
\end{aligned}$$

This term is is not a small order term so we keep it as what it is. And lastly, the proof of order for the term $J8$ is too long so we show in Lemma 16 that

$$J8 = o_p \left(\|C - \hat{C}\| \right) + O_p \left(\frac{1}{\min(n, p)} \right) + O_p \left(\frac{\sqrt{n}}{\sqrt{p}} \frac{1}{\min(n, p)} \right). \quad (4.36)$$

Collecting terms from $J1$ to $J8$, we can write (4.26) as

$$\left(\frac{1}{p} \Phi^\top M_{\hat{A}} \Phi + \frac{1}{p} \alpha R^\top \right) (\hat{C} - C) = \frac{1}{p} \alpha R^\top C + J1 + \dots + J8 + \frac{1}{p} \Phi^\top M_{\hat{A}} E.$$

Combining the results we have found for $J1, J3, J4, J5, J6$ and $J8$ in Eqs. (4.32)–(4.34)

and (4.36),

$$\begin{aligned} \left(\frac{1}{p} \Phi^\top M_{\hat{A}} \Phi + o_p(1) \right) (\hat{C} - C) - J2 &= \frac{1}{p} \alpha R^\top C + \frac{1}{p} \Phi^\top M_{\hat{A}} E + J7 \\ &+ O_p \left(\frac{1}{\min(n, p)} \right) + O_p \left(\frac{\sqrt{n}}{p\sqrt{p}} \right) + O_p \left(\frac{1}{\sqrt{np}} \right). \end{aligned} \quad (4.37)$$

Substitute $J2$ and $J7$ from Eqs. (4.30) and (4.35) into (4.37), we have

$$\begin{aligned} &\left(\frac{1}{p} \Phi^\top M_{\hat{A}} \Phi + o_p(1) \right) (\hat{C} - C) - \frac{1}{p} \Phi^\top M_{\hat{A}} \Phi (\hat{C} - C) F (F^\top F)^{-1} F^\top \\ &= \frac{1}{p} \alpha R^\top C + \frac{1}{p} \Phi^\top M_{\hat{A}} E - \frac{1}{p} \Phi^\top M_{\hat{A}} E F (F^\top F)^{-1} F^\top \\ &+ O_p \left(\frac{1}{\min(n, p)} \right) + O_p \left(\frac{\sqrt{n}}{p\sqrt{p}} \right) + O_p \left(\frac{1}{\sqrt{np}} \right). \end{aligned} \quad (4.38)$$

We combine the two terms on the left-hand side of (4.38) and also combine the second and third term on the right-hand side of (4.38), then we get

$$\begin{aligned} \frac{1}{p} \Phi^\top M_{\hat{A}} \Phi (\hat{C} - C) \left(I_n - F (F^\top F)^{-1} F^\top \right) &= \frac{1}{p} \alpha R^\top C + \frac{1}{p} \Phi^\top M_{\hat{A}} E \left(I_n - F (F^\top F)^{-1} F^\top \right) \\ &+ O_p \left(\frac{1}{\min(n, p)} \right) + O_p \left(\frac{\sqrt{n}}{p\sqrt{p}} \right) + O_p \left(\frac{1}{\sqrt{np}} \right). \end{aligned}$$

Let $Q(\hat{A}) \equiv \frac{1}{p} \Phi^\top M_{\hat{A}} \Phi$, and $M_F \equiv I_n - F(F^\top F)^{-1}F^\top$. Left multiplying $Q^{-1}(\hat{A})$ to both

sides of the equation above, we have

$$\begin{aligned}
(\widehat{\mathbf{C}} - \mathbf{C})\mathbf{M}_F &= \mathbf{Q}(\widehat{\mathbf{A}})^{-1} \frac{1}{p} \alpha \mathbf{R}^\top \mathbf{C} + \mathbf{Q}(\widehat{\mathbf{A}})^{-1} \frac{1}{p} \boldsymbol{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}} \mathbf{E} \mathbf{M}_F \\
&\quad + O_p\left(\frac{1}{\min(n, p)}\right) + O_p\left(\frac{\sqrt{n}}{p\sqrt{p}}\right) + O_p\left(\frac{1}{\sqrt{np}}\right) \\
&= \mathbf{Q}^{-1}(\mathbf{A}^0) \frac{1}{p} \alpha \mathbf{R}^\top \mathbf{C} + \mathbf{Q}^{-1}(\mathbf{A}^0) \frac{1}{p} \boldsymbol{\Phi}^\top \mathbf{M}_{\mathbf{A}^0} \mathbf{E} \mathbf{M}_F \\
&\quad + O_p\left(\frac{1}{\min(n, p)}\right) + O_p\left(\frac{\sqrt{n}}{p\sqrt{p}}\right) + O_p\left(\frac{1}{\sqrt{np}}\right),
\end{aligned}$$

where in the last equation, we substitute $\mathbf{Q}(\widehat{\mathbf{A}})$ with $\mathbf{Q}(\mathbf{A}^0)$ using Lemma 13 and substitute $\frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}} \mathbf{E}$ with $\frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top \mathbf{M}_{\mathbf{A}^0} \mathbf{E}$ using Lemma 15. Note that $\sqrt{p} O_p\left(\frac{\|\mathbf{c} - \widehat{\mathbf{c}}\|^2}{n}\right)$ in the result of Lemma 15 is dominated by $\sqrt{p} \frac{\|\mathbf{c} - \widehat{\mathbf{c}}\|}{\sqrt{n}}$. Next by multiplying a scale of $\frac{\sqrt{p}}{\sqrt{n}}$,

$$\begin{aligned}
\frac{\sqrt{p}}{\sqrt{n}} (\widehat{\mathbf{C}} - \mathbf{C})\mathbf{M}_F &= \mathbf{Q}(\mathbf{A}^0)^{-1} \frac{1}{p} \alpha \mathbf{R}^\top \mathbf{C} + \mathbf{Q}(\mathbf{A}^0)^{-1} \frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top \mathbf{M}_{\mathbf{A}^0} \mathbf{E} \mathbf{M}_F \\
&\quad + \frac{\sqrt{p}}{\sqrt{n}} \times O_p\left(\frac{1}{\min(n, p)}\right) + \frac{\sqrt{p}}{\sqrt{n}} \times O_p\left(\frac{\sqrt{n}}{p\sqrt{p}}\right) + O_p\left(\frac{1}{\sqrt{np}}\right) \\
&= O_p(1), \tag{4.39}
\end{aligned}$$

when n and p are of the same order, that is $p/n \rightarrow \rho > 0$. □

Proof of Theorem 8

From (4.39), we have when $p/n \rightarrow \rho > 0$,

$$\frac{\sqrt{p}}{\sqrt{n}} (\widehat{\mathbf{C}} - \mathbf{C})\mathbf{M}_F = \mathbf{Q}(\mathbf{A}^0)^{-1} \frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top \mathbf{M}_{\mathbf{A}^0} \mathbf{E} \mathbf{M}_F + o_p(1).$$

Using Lemma 6 we have, for any vector $\mathbf{b} = (b_1, \dots, b_n)^\top$,

$$\frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top M_{A^0} E M_F \mathbf{b} \xrightarrow{d} \mathcal{N}(0, L), \quad (4.40)$$

where L is defined in (4.16).

Multiplying the constant matrix $Q(A^0)^{-1}$ to (4.40), we have the result

$$Q(A^0)^{-1} \frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top M_{A^0} E M_F \mathbf{b} \xrightarrow{d} \mathcal{N}\left(0, Q(A^0)^{-1} L Q(A^0)^{-1}\right).$$

The theorem is thus proved.

4.9.2 Appendix B

In this section, we provide the proposition used in Appendix A along with its proof.

Proposition 2. *Under Assumptions 10 to 13, we have the following statements:*

(i) *The matrix V_{np} defined in (4.8) is invertible and $V_{np} \xrightarrow{p} V$, where the $r \times r$ matrix V is a diagonal matrix consisting of the eigenvalues of $\boldsymbol{\Sigma}_F \boldsymbol{\Sigma}_A$;*

(ii) *Let*

$$\mathbf{H} = (\mathbf{F}^\top \mathbf{F} / n)^{-1} (\mathbf{A}^{0\top} \widehat{\mathbf{A}} / p)^{-1} V_{np}, \quad (4.41)$$

then \mathbf{H} is $r \times r$ invertible matrix and

$$\frac{1}{p} \|\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}\|^2 = O_p\left(\frac{1}{n} \|\mathbf{C} - \widehat{\mathbf{C}}\|^2\right) + O_p\left(\frac{1}{\min(n, p)}\right).$$

Proof. Write the second equation in (4.9) in a matrix form, we have

$$\frac{1}{np}(\mathbf{Y} - \Phi\hat{\mathbf{C}})(\mathbf{Y} - \Phi\hat{\mathbf{C}})^\top \hat{\mathbf{A}} = \hat{\mathbf{A}}\mathbf{V}_{np}.$$

By (4.2), we also have

$$\mathbf{Y} - \Phi\hat{\mathbf{C}} = \Phi(\mathbf{C} - \hat{\mathbf{C}}) + \mathbf{A}^0\mathbf{F}^\top + \mathbf{E}. \quad (4.42)$$

Plugging it in (4.42) and by expanding terms, we obtain

$$\begin{aligned} \hat{\mathbf{A}}\mathbf{V}_{np} &= \frac{1}{np} \left[\Phi(\mathbf{C} - \hat{\mathbf{C}}) + \mathbf{A}^0\mathbf{F}^\top + \mathbf{E} \right] \left[\Phi(\mathbf{C} - \hat{\mathbf{C}}) + \mathbf{A}^0\mathbf{F}^\top + \mathbf{E} \right]^\top \hat{\mathbf{A}} \\ &= \frac{1}{np} \Phi(\mathbf{C} - \hat{\mathbf{C}})(\mathbf{C} - \hat{\mathbf{C}})^\top \Phi^\top \hat{\mathbf{A}} + \frac{1}{np} \Phi(\mathbf{C} - \hat{\mathbf{C}})\mathbf{F}\mathbf{A}^{0\top} \hat{\mathbf{A}} \\ &\quad + \frac{1}{np} \Phi(\mathbf{C} - \hat{\mathbf{C}})\mathbf{E}^\top \hat{\mathbf{A}} + \frac{1}{np} \mathbf{A}^0\mathbf{F}^\top (\mathbf{C} - \hat{\mathbf{C}})^\top \Phi^\top \hat{\mathbf{A}}, \\ &\quad + \frac{1}{np} \mathbf{E}(\mathbf{C} - \hat{\mathbf{C}})^\top \Phi^\top \hat{\mathbf{A}} + \frac{1}{np} \mathbf{A}^0\mathbf{F}^\top \mathbf{E}^\top \hat{\mathbf{A}} \\ &\quad + \frac{1}{np} \mathbf{E}\mathbf{F}\mathbf{A}^{0\top} \hat{\mathbf{A}} + \frac{1}{np} \mathbf{E}\mathbf{E}^\top \hat{\mathbf{A}} \\ &\quad + \frac{1}{np} \mathbf{A}^0\mathbf{F}^\top \mathbf{F}\mathbf{A}^{0\top} \hat{\mathbf{A}} \\ &\equiv I1 + \dots + I9. \end{aligned} \quad (4.43)$$

The above can be rewritten as

$$\hat{\mathbf{A}}\mathbf{V}_{np} - \mathbf{A}^0(\mathbf{F}^\top \mathbf{F}/n)(\mathbf{A}^{0\top} \hat{\mathbf{A}}/p) = I1 + \dots + I8. \quad (4.44)$$

Right multiplying $(\mathbf{F}^\top \mathbf{F}/n)^{-1}(\mathbf{A}^{0\top} \hat{\mathbf{A}}/p)^{-1}$ on each side, we obtain

$$\hat{\mathbf{A}} \left[\mathbf{V}_{np}(\mathbf{A}^{0\top} \hat{\mathbf{A}}/p)^{-1}(\mathbf{F}^\top \mathbf{F}/n)^{-1} \right] - \mathbf{A}^0 = (I1 + \dots + I8)(\mathbf{A}^{0\top} \hat{\mathbf{A}}/p)^{-1}(\mathbf{F}^\top \mathbf{F}/n)^{-1}. \quad (4.45)$$

Note that the matrix in the square brackets is \mathbf{H}^{-1} , but the invertibility of \mathbf{V}_{np} hasn't been proved yet. We can write

$$\frac{1}{\sqrt{p}} \left\| \widehat{\mathbf{A}} \left[\mathbf{V}_{np} (\mathbf{A}^{0\top} \widehat{\mathbf{A}}/p)^{-1} (\mathbf{F}^\top \mathbf{F}/n)^{-1} \right] - \mathbf{A}^0 \right\| \leq \frac{1}{\sqrt{p}} (\|I1\| + \cdots + \|I8\|) \|\mathbf{G}\|, \quad (4.46)$$

where \mathbf{G} is defined in (4.28) and $\|\mathbf{G}\|$ is proved to be $O_p(1)$ in Lemma 10. In the following, we find the order for each term on the right-hand side of (4.46). We repeatedly use results from Lemma 7, where the orders of the matrices Φ, \mathbf{A} and \mathbf{F} are given. The first term

$$\begin{aligned} \frac{1}{\sqrt{p}} \|I1\| &\leq \frac{1}{\sqrt{p}} \frac{1}{np} \|\Phi\| \|(\mathbf{C} - \widehat{\mathbf{C}})(\mathbf{C} - \widehat{\mathbf{C}})^\top\| \|\Phi^\top\| \|\widehat{\mathbf{A}}\| \\ &= O_p \left(\frac{1}{n} \|\mathbf{C} - \widehat{\mathbf{C}}\|^2 \right) = o_p \left(\frac{1}{\sqrt{n}} \|\mathbf{C} - \widehat{\mathbf{C}}\| \right). \end{aligned}$$

For the second term

$$\begin{aligned} \frac{1}{\sqrt{p}} \|I2\| &\leq \frac{1}{\sqrt{p}} \frac{1}{np} \|\Phi\| \|(\mathbf{C} - \widehat{\mathbf{C}})\| \|\mathbf{F}\| \|\mathbf{A}^{0\top}\| \|\widehat{\mathbf{A}}\| \\ &= O_p \left(\frac{1}{\sqrt{n}} \|\mathbf{C} - \widehat{\mathbf{C}}\| \right). \end{aligned}$$

The terms $I3$ to $I5$ are all $O_p(\frac{1}{\sqrt{n}} \|\mathbf{C} - \widehat{\mathbf{C}}\|)$. The proofs are similar to the proof for $I2$ since they are only a switch in the order of the matrices. For the sixth term

$$\frac{1}{\sqrt{p}} \|I6\| \leq \frac{1}{\sqrt{p}} \frac{1}{np} \|\mathbf{A}^0\| \|\mathbf{F}^\top \mathbf{E}^\top\| \|\widehat{\mathbf{A}}\| = O_p \left(\frac{1}{\sqrt{n}} \right),$$

by Lemma 8 (i). Similarly, for the next term

$$\frac{1}{\sqrt{p}} \|I7\| \leq \frac{1}{\sqrt{p}} \frac{1}{np} \|\mathbf{E}\mathbf{F}\| \|\mathbf{A}^{0\top}\| \|\widehat{\mathbf{A}}\| = O_p \left(\frac{1}{\sqrt{n}} \right).$$

For the last term

$$\frac{1}{\sqrt{p}} \|I8\| \leq \frac{1}{\sqrt{p}} \frac{1}{np} \|EE^\top\| \|\hat{\mathbf{A}}\| = O_p\left(\frac{1}{\sqrt{n}}\right) + \left(\frac{1}{\sqrt{p}}\right),$$

where Lemma 8 (iv) is used.

Putting all above together, we have

$$\begin{aligned} \frac{1}{\sqrt{p}} \left\| \hat{\mathbf{A}} \left[\mathbf{V}_{np} (\mathbf{A}^{0\top} \hat{\mathbf{A}}/p)^{-1} (\mathbf{F}^\top \mathbf{F}/n)^{-1} \right] - \mathbf{A}^0 \right\| &= O_p\left(\frac{1}{\sqrt{n}} \|\mathbf{C} - \hat{\mathbf{C}}\|\right) \\ &+ O_p\left(\frac{1}{\min(\sqrt{n}, \sqrt{p})}\right). \end{aligned} \quad (4.47)$$

To show (i), left multiply (4.44) by $\frac{1}{p} \hat{\mathbf{A}}^\top$. Using $\hat{\mathbf{A}}^\top \hat{\mathbf{A}}/p = \mathbf{I}_r$, we have

$$\mathbf{V}_{np} - (\hat{\mathbf{A}}^\top \mathbf{A}^0/p)(\mathbf{F}^\top \mathbf{F}/n)(\mathbf{A}^{0\top} \hat{\mathbf{A}}/p) = \frac{1}{p} \hat{\mathbf{A}}^\top (I1 + \dots + I8) = o_p(1),$$

where the last equality is using Lemma 7 (v) and that $p^{-1/2}(\|I1\| + \dots + \|I8\|) = o_p(1)$ from (4.47). Thus,

$$\mathbf{V}_{np} = (\hat{\mathbf{A}}^\top \mathbf{A}^0/p)(\mathbf{F}^\top \mathbf{F}/n)(\mathbf{A}^{0\top} \hat{\mathbf{A}}/p) + o_p(1).$$

We have shown in (4.24) that $\hat{\mathbf{A}}^\top \hat{\mathbf{A}}^0$ is invertible, thus \mathbf{V}_{np} is invertible. To obtain the limit of \mathbf{V}_{np} , left multiply (4.44) by $\frac{1}{p} \mathbf{A}^{0\top}$ to yield

$$(\mathbf{A}^{0\top} \hat{\mathbf{A}}/p) \mathbf{V}_{np} - (\mathbf{A}^{0\top} \mathbf{A}^0/p)(\mathbf{F}^\top \mathbf{F}/n)(\mathbf{A}^{0\top} \hat{\mathbf{A}}/p) = o_p(1),$$

or

$$(\mathbf{A}^{0\top} \mathbf{A}^0/p)(\mathbf{F}^\top \mathbf{F}/n)(\mathbf{A}^{0\top} \widehat{\mathbf{A}}/p) + o_p(1) = (\mathbf{A}^{0\top} \widehat{\mathbf{A}}/p) \mathbf{V}_{np} \quad (4.48)$$

because $p^{-1} \mathbf{A}^{0\top} (\|I1\| + \dots + \|I8\|) = o_p(1)$. Equation (4.48) shows that the columns of $(\mathbf{A}^{0\top} \widehat{\mathbf{A}}/p)$ are the eigenvectors of the matrix $(\mathbf{A}^{0\top} \mathbf{A}^0/p)(\mathbf{F}^\top \mathbf{F}/n)$, and that \mathbf{V}_{np} consists of the eigenvalues of the same matrix in the limit. Thus, $\mathbf{V}_{np} \xrightarrow{p} \mathbf{V}$, where the $r \times r$ matrix \mathbf{V} is a diagonal matrix consisting of the eigenvalues of $\boldsymbol{\Sigma}_F \boldsymbol{\Sigma}_A$.

For (ii), since \mathbf{V}_{np} is invertible, \mathbf{H} is also invertible we can write (4.47) as

$$\frac{1}{\sqrt{p}} \left\| \widehat{\mathbf{A}} \mathbf{H}^{-1} - \mathbf{A}^0 \right\| = O_p \left(\frac{1}{\sqrt{n}} \left\| \mathbf{C} - \widehat{\mathbf{C}} \right\| \right) + O_p \left(\frac{1}{\min(\sqrt{n}, \sqrt{p})} \right).$$

By right multiplying the matrix \mathbf{H} , we obtain (ii). □

4.9.3 Appendix C

In this section, we state all the lemmas used for previous theorems and propositions, along with the proofs of the lemmas.

Lemma 6 is stated in Section 4.4.2. We provide the proof here.

Proof. For any vector $\mathbf{b} = (b_1, \dots, b_n)^\top$,

$$\frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top \mathbf{M}_{A^0} \mathbf{E} \mathbf{b} = \frac{1}{\sqrt{np}} \sum_i^n \sum_j^p \omega_j \epsilon_{ji} b_i \equiv \frac{1}{\sqrt{np}} \sum_i^n \sum_j^p \mathbf{x}_{ij}$$

where ω_j is the j th column in the matrix $\boldsymbol{\Phi}^\top \mathbf{M}_{A^0}$. Since we assume ϵ_{ji} are i.i.d., the

variance of the above quantity is given by

$$\text{var} \left(\frac{1}{\sqrt{np}} \mathbf{\Phi}^\top \mathbf{M}_{A^0} \mathbf{E} \mathbf{b} \right) = \text{var} \left(\frac{1}{\sqrt{np}} \sum_i^n \sum_j^p \mathbf{x}_{ij} \right) = \frac{1}{np} \sum_i^p \sum_j^n b_j b_i \sigma^2 E \left(\boldsymbol{\omega}_i \boldsymbol{\omega}_j^\top \right).$$

The Lindeberg condition is assumed to hold in Assumption 16. Thus we have central limit theorem result

$$\frac{1}{\sqrt{np}} \mathbf{\Phi}^\top \mathbf{M}_{A^0} \mathbf{E} \mathbf{b} = \frac{1}{\sqrt{np}} \sum_i^n \sum_j^p \mathbf{x}_{ij} \xrightarrow{d} \mathcal{N}(0, \mathbf{L}),$$

where \mathbf{L} is defined in (4.16).

□

Lemma 7. *Under Assumptions 10-12, we have*

$$(i) \frac{1}{\sqrt{p}} \|\mathbf{\Phi}\| = O_p(1)$$

$$(ii) \frac{1}{\sqrt{p}} \|\mathbf{A}\| = O_p(1)$$

$$(iii) \frac{1}{\sqrt{n}} \|\mathbf{F}\| = O_p(1)$$

$$(iv) \frac{1}{\sqrt{np}} \|\mathbf{E}\| = O_p(1)$$

$$(v) \frac{1}{\sqrt{p}} \|\widehat{\mathbf{A}}\| = O_p(1)$$

Proof. In Assumption 10, we assume the basis functions $\phi_k(u)$, $k = 1, \dots, K$ are bounded. The $p \times K$ basis matrix $\mathbf{\Phi}$ contains discrete evaluations on the basis functions so each element is $O_p(1)$, thus $\mathbf{\Phi}$ is of order \sqrt{p} . Similarly, using Assumption 12, we have results (ii) and (iii). Using Assumption 13, we have result (iv). Lastly, (v) is directly from the restriction $\widehat{\mathbf{A}}^\top \widehat{\mathbf{A}}/p = \mathbf{I}_r$. □

Lemma 8. *Under Assumptions 10 to 14, we have*

- (i) $\frac{1}{np} \|\mathbf{EF}\|^2 = O_p(1)$
- (ii) $\frac{1}{np} \|\mathbf{E}^\top \Phi\|^2 = O_p(1)$ and $\frac{1}{np} \|\mathbf{E}^\top \mathbf{A}^0\|^2 = O_p(1)$
- (iii) $\frac{1}{np} \|\mathbf{F}^\top \mathbf{E}^\top \mathbf{A}^0\|^2 = O_p(1)$ and $\frac{1}{np} \|\Phi^\top \mathbf{E}^\top \mathbf{F}\|^2 = O_p(1)$
- (iv) $\|\mathbf{E}^\top \mathbf{E}\|^2 = O_p(n^2p) + O_p(p^2n);$
 $\|\mathbf{EE}^\top\|^2 = O_p(n^2p) + O_p(p^2n);$
 $\|\mathbf{F}^\top \mathbf{E}^\top \mathbf{E}\|^2 = O_p(n^2p) + O_p(p^2n);$
 $\|\Phi^\top \mathbf{E}^\top \mathbf{E}\|^2 = O_p(n^2p) + O_p(p^2n);$
 $\|\Phi^\top \mathbf{E}^\top \mathbf{EA}^0\|^2 = O_p(n^2p) + O_p(p^2n);$
 $\|\mathbf{F}^\top \mathbf{E}^\top \mathbf{EF}\|^2 = O_p(n^2p) + O_p(p^2n).$

Proof. For (i)

$$\begin{aligned} \mathbb{E} \left(\frac{1}{np} \|\mathbf{EF}\|^2 \right) &= \mathbb{E} \left(\frac{1}{np} \sum_{k=1}^p \sum_{i=1}^n \sum_{j=1}^n \epsilon_{ki} \epsilon_{kj} \mathbf{f}_i^\top \mathbf{f}_j \right) \\ &= \frac{1}{np} \sum_{k=1}^p \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}(\epsilon_{ki} \epsilon_{kj}) \mathbb{E}(\mathbf{f}_i^\top \mathbf{f}_j) = O(1), \end{aligned}$$

where the second equation uses the independence between ϵ_{ki} and \mathbf{f}_j assumed in Assumption 14.

The proof of (ii) and (iii) is similar to (i). For (iv),

$$\begin{aligned} \mathbb{E} \left(\|\mathbf{E}^\top \mathbf{E}\|^2 \right) &= \mathbb{E} \left(\sum_{ij}^n \sum_{kl}^p \epsilon_{kj} \epsilon_{lj} \epsilon_{ki} \epsilon_{li} \right) \\ &= \sum_{i \neq j}^n \sum_{k=l}^p \mathbb{E}(\epsilon_{kj}^2) \mathbb{E}(\epsilon_{ki}^2) + \sum_{i=j}^n \sum_{k \neq l}^p \mathbb{E}(\epsilon_{kj}^2) \mathbb{E}(\epsilon_{li}^2) + \sum_{i=j}^n \sum_{k=l}^p \mathbb{E}(\epsilon_{kj}^4) \\ &= O(n^2p) + O(p^2n) + O(np) \\ &= O(n^2p) + O(p^2n), \end{aligned}$$

where Assumption 13 is used. The proof of $\|EE^\top\|$ is the same. The orders of $\|F^\top E^\top E\|$ and $\|F^\top E^\top EF\|^2$ are the same since

$$\mathbb{E} \left(\|F^\top E^\top E\|^2 \right) = \mathbb{E} \left(\sum_{ij}^n \sum_{kl}^p \epsilon_{kj} \epsilon_{lj} \epsilon_{ki} \epsilon_{li} \|f_i\|^2 \right),$$

and

$$\mathbb{E} \left(\|F^\top E^\top EF\|^2 \right) = \mathbb{E} \left(\sum_{ij}^n \sum_{kl}^p \epsilon_{kj} \epsilon_{lj} \epsilon_{ki} \epsilon_{li} \|f_i\|^4 \right),$$

where the order of f_i is assumed to be $O_p(1)$ in Assumption 12. □

Lemma 9. Under Assumptions 10-15,

- (i) $\frac{1}{np} \sum_{i=1}^n \epsilon_i^\top M_A A^0 F_i = o_p(1)$
- (ii) $\frac{1}{np} \sum_{i=1}^n \epsilon_i^\top M_A \Phi c_i = o_p(1)$
- (iii) $\frac{1}{np} \sum_{i=1}^n \epsilon_i^\top (M_A - M_{A^0}) \epsilon_i = o_p(1)$
- (iv) $\frac{\alpha}{np} \sum_{i=1}^n c_i^\top R c_i = o_p(1)$

Proof. We prove (ii). First we have

$$\mathbb{E} \left(\left\| \sum_{i=1}^n \epsilon_i \right\|^2 \right) = \mathbb{E} \left(\sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^p \epsilon_{ik} \epsilon_{jk} \right) = \sum_{i=j}^n \sum_{k=1}^p \mathbb{E} (\epsilon_{ik}^2) = O(np).$$

Since $M_A = I_p - AA^\top/p$, we have

$$\frac{1}{np} \sum_{i=1}^n \epsilon_i^\top M_A \Phi c_i = \frac{1}{np} \sum_{i=1}^n \epsilon_i^\top \Phi c_i - \frac{1}{np^2} \sum_{i=1}^n \epsilon_i^\top AA^\top \Phi c_i. \quad (4.49)$$

The first term on the right of (4.49) is $o_p(1)$ since

$$\begin{aligned}
\mathbb{E} \left(\left\| \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top \boldsymbol{\Phi} \mathbf{c}_i \right\|^2 \right) &= \mathbb{E} \left(\left\| \sum_{j=1}^p \sum_{i=1}^n \epsilon_{ji} \boldsymbol{\phi}_j^\top \mathbf{c}_i \right\|^2 \right) \\
&= \mathbb{E} \left(\sum_t^p \sum_s^n \sum_j^p \sum_i^n \epsilon_{ji} \epsilon_{ts} \mathbf{c}_i^\top \boldsymbol{\phi}_j \boldsymbol{\phi}_t^\top \mathbf{c}_s \right) \\
&= \sum_t^p \sum_s^n \sum_j^p \sum_i^n \mathbb{E}(\epsilon_{ji} \epsilon_{ts}) \mathbb{E}(\mathbf{c}_i^\top \boldsymbol{\phi}_j \boldsymbol{\phi}_t^\top \mathbf{c}_s) \\
&= \sum_j^p \sum_i^n \sigma^2 \mathbb{E}(\mathbf{c}_i^\top \boldsymbol{\phi}_j \boldsymbol{\phi}_j^\top \mathbf{c}_i) \\
&= O(np),
\end{aligned}$$

where the third equation uses Assumption 14; the fourth equation uses the assumption that ϵ_{ji} are independent in both directions.

The second term on the right-hand side of (4.49) is also $o_p(1)$ since

$$\begin{aligned}
\mathbb{E} \left(\left\| \sum_{i=1}^n \boldsymbol{\epsilon}_i^\top \mathbf{A} \mathbf{A}^\top \boldsymbol{\Phi} \mathbf{c}_i \right\|^2 \right) &= \mathbb{E} \left(\left\| \sum_{j=1}^p \sum_{i=1}^n \epsilon_{ji} \mathbf{a}_j^\top \mathbf{A}^\top \boldsymbol{\Phi} \mathbf{c}_i \right\|^2 \right) \\
&= \mathbb{E} \left(\sum_t^p \sum_s^n \sum_j^p \sum_i^n \epsilon_{ji} \epsilon_{ts} \mathbf{c}_i^\top \boldsymbol{\Phi}^\top \mathbf{A} \mathbf{a}_j \mathbf{a}_t^\top \mathbf{A}^\top \boldsymbol{\Phi} \mathbf{c}_s \right) \\
&= \sum_t^p \sum_s^n \sum_j^p \sum_i^n \mathbb{E}(\epsilon_{ji} \epsilon_{ts}) \mathbb{E}(\mathbf{c}_i^\top \boldsymbol{\Phi}^\top \mathbf{A} \mathbf{a}_j \mathbf{a}_t^\top \mathbf{A}^\top \boldsymbol{\Phi} \mathbf{c}_s) \\
&= \sum_j^p \sum_i^n \sigma^2 \mathbf{c}_i^\top \mathbb{E}(\boldsymbol{\Phi}_i^\top \mathbf{A} \mathbf{a}_j \mathbf{a}_j^\top \mathbf{A}^\top \boldsymbol{\Phi}) \mathbf{c}_i \\
&= O(np^3),
\end{aligned}$$

where the third equality uses the independence in Assumption 14 and the last equality uses the results in Lemma 7, where $\boldsymbol{\Phi}$ and \mathbf{A} are both $O_p(\sqrt{p})$.

The proofs for (i) and (iii) are similar. And (iv) is a direct result from Assumption 15. \square

Lemma 10. *Under Assumptions 10-14, we have*

$$\mathbf{G} \equiv \left(\mathbf{A}^{0\top} \widehat{\mathbf{A}}/p\right)^{-1} \left(\mathbf{F}^\top \mathbf{F}/n\right)^{-1} = O_p(1)$$

Proof. The matrix $\mathbf{F}^\top \mathbf{F}/n$ is positive definite by Assumption 12. We have shown in the proof of Theorem 6 in (4.24) that the matrix $\mathbf{A}^{0\top} \widehat{\mathbf{A}}/p$ is invertible, thus is also positive definite. Therefore, $\lambda_{\min} \left(\mathbf{A}^{0\top} \widehat{\mathbf{A}}/p\right) > 0$, and $\lambda_{\min} \left(\mathbf{F}^\top \mathbf{F}/n\right) > 0$, where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue of a matrix. So we have

$$\left(\mathbf{A}^{0\top} \widehat{\mathbf{A}}/p\right)^{-1} = O_p(1), \quad \left(\mathbf{F}^\top \mathbf{F}/n\right)^{-1} = O_p(1).$$

\square

Lemma 11. *We have the following*

(i)

$$\left\| \mathbf{E}^\top (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}) \right\| = O_p \left(\frac{p}{\min(\sqrt{n}, \sqrt{p})} \left\| \mathbf{C} - \widehat{\mathbf{C}} \right\| \right) + O_p(\sqrt{n}) + O_p \left(\frac{p}{\sqrt{n}} \right).$$

(ii)

$$\left\| \mathbf{F}^\top \mathbf{E}^\top (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}) \right\| = O_p \left(\frac{p}{\min(\sqrt{n}, \sqrt{p})} \left\| \mathbf{C} - \widehat{\mathbf{C}} \right\| \right) + O_p(\sqrt{n}) + O_p \left(\frac{p}{\sqrt{n}} \right).$$

Proof. For (i), from Proposition 2, we can write

$$\begin{aligned}\|E^\top(\widehat{A} - A^0H)\| &= \|E^\top(I1 + \dots, I8)G\| \\ &\leq \|E^\top I1G\| + \dots + \|E^\top I8G\| \\ &= \|a1\| + \dots + \|a8\|.\end{aligned}$$

To find the order for each term, the results from Lemma 7 are repeatedly used where the order of the matrices Φ , A , \widehat{A} and F are given.

$$\begin{aligned}\|a1\| &= \left\| E^\top \frac{1}{np} \Phi (C - \widehat{C})(C - \widehat{C})^\top \Phi^\top \widehat{A} G \right\| \\ &\leq \frac{1}{np} \|E^\top \Phi\| \|C - \widehat{C}\|^2 \|\Phi\| \|\widehat{A}\| \|G\| \\ &= O_p\left(\frac{\sqrt{p}}{\sqrt{n}} \|C - \widehat{C}\|^2\right) = o_p\left(\sqrt{p} \|C - \widehat{C}\|\right),\end{aligned}$$

where the order of $\|E^\top \Phi\|$ is from Lemma 8 (ii). The orders of $\|\Phi\|$, $\|\widehat{A}\|$ and $\|G\|$ can be found from Lemmas 7 and 10. Similarly,

$$\begin{aligned}\|a2\| &= \left\| E^\top \frac{1}{n} \Phi (C - \widehat{C}) F \left(\frac{F^\top F}{n}\right)^{-1} \right\| \\ &\leq \frac{1}{n} \|E^\top \Phi\| \|C - \widehat{C}\| \|F\| \left\| \left(\frac{F^\top F}{n}\right)^{-1} \right\| \\ &= O_p\left(\sqrt{p} \|C - \widehat{C}\|\right).\end{aligned}$$

$$\begin{aligned}
\|a3\| &= \left\| \mathbf{E}^\top \frac{1}{np} \boldsymbol{\Phi} (\mathbf{C} - \widehat{\mathbf{C}}) \mathbf{E}^\top \widehat{\mathbf{A}} \mathbf{G} \right\| \\
&\leq \frac{1}{np} \|\mathbf{E}^\top \boldsymbol{\Phi}\| \|\mathbf{C} - \widehat{\mathbf{C}}\| \|\mathbf{E}^\top\| \|\widehat{\mathbf{A}}\| \|\mathbf{G}\| \\
&= O_p \left(\sqrt{p} \|\mathbf{C} - \widehat{\mathbf{C}}\| \right).
\end{aligned}$$

$$\begin{aligned}
\|a4\| &= \left\| \mathbf{E}^\top \frac{1}{np} \mathbf{A}^0 \mathbf{F}^\top (\mathbf{C} - \widehat{\mathbf{C}})^\top \boldsymbol{\Phi}^\top \widehat{\mathbf{A}} \mathbf{G} \right\| \\
&\leq \frac{1}{np} \|\mathbf{E}^\top \mathbf{A}^0\| \|\mathbf{F}^\top\| \|\mathbf{C} - \widehat{\mathbf{C}}\| \|\boldsymbol{\Phi}\| \|\widehat{\mathbf{A}}\| \|\mathbf{G}\| \\
&= O_p \left(\sqrt{p} \|\mathbf{C} - \widehat{\mathbf{C}}\| \right),
\end{aligned}$$

where Lemma 8 (ii) is used.

$$\begin{aligned}
\|a5\| &= \left\| \mathbf{E}^\top \frac{1}{np} \mathbf{E} (\mathbf{C} - \widehat{\mathbf{C}})^\top \boldsymbol{\Phi}^\top \widehat{\mathbf{A}} \mathbf{G} \right\| \\
&\leq \frac{1}{np} \|\mathbf{E}^\top \mathbf{E}\| \|\mathbf{C} - \widehat{\mathbf{C}}\| \|\boldsymbol{\Phi}\| \|\widehat{\mathbf{A}}\| \|\mathbf{G}\| \\
&= O_p \left(\sqrt{p} \|\mathbf{C} - \widehat{\mathbf{C}}\| \right) + O_p \left(\frac{p}{\sqrt{n}} \|\mathbf{C} - \widehat{\mathbf{C}}\| \right),
\end{aligned}$$

where Lemma 8 (iv) is used.

$$\begin{aligned}
\|a6\| &= \|\mathbf{E}^\top \frac{1}{np} \mathbf{A}^0 \mathbf{F}^\top \mathbf{E}^\top \widehat{\mathbf{A}} \mathbf{G}\| \\
&\leq \frac{1}{np} \|\mathbf{E}^\top \mathbf{A}^0 \mathbf{F}^\top \mathbf{E}^\top (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}) \mathbf{G}\| + \frac{1}{np} \|\mathbf{E}^\top \mathbf{A}^0 \mathbf{F}^\top \mathbf{E}^\top \mathbf{A}^0 \mathbf{H} \mathbf{G}\| \\
&\leq \frac{1}{np} \|\mathbf{E}^\top \mathbf{A}^0\| \|\mathbf{F}^\top \mathbf{E}^\top\| \|\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}\| \|\mathbf{G}\| + \frac{1}{np} \|\mathbf{E}^\top \mathbf{A}^0\| \|\mathbf{F}^\top \mathbf{E}^\top \mathbf{A}^0\| \|\mathbf{H} \mathbf{G}\| \\
&= O_p \left(\sqrt{p} \left(\frac{\|\mathbf{C} - \widehat{\mathbf{C}}\|}{\sqrt{n}} \right) \right) + O_p \left(\frac{1}{\min(\sqrt{n}, \sqrt{p})} \right) + O_p(1) \\
&= O_p \left(\sqrt{p} \left(\frac{\|\mathbf{C} - \widehat{\mathbf{C}}\|}{\sqrt{n}} \right) \right) + O_p \left(\frac{1}{\min(\sqrt{n}, \sqrt{p})} \right),
\end{aligned}$$

where the order of $\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}$ is proved in Proposition 2 and the order of other matrix norms can be found in Lemma 8 (i), (ii) and (iii).

$$\begin{aligned}
\|a7\| &= \|\mathbf{E}^\top \frac{1}{np} \mathbf{E} \mathbf{F} \mathbf{A}^{0\top} \widehat{\mathbf{A}} \mathbf{G}\| \\
&= \left\| \frac{1}{n} \mathbf{E}^\top \mathbf{E} \mathbf{F} \left(\frac{\mathbf{F}^\top \mathbf{F}}{n} \right)^{-1} \right\| \\
&\leq \frac{1}{n} \|\mathbf{E}^\top \mathbf{E} \mathbf{F}\| \left\| \left(\frac{\mathbf{F}^\top \mathbf{F}}{n} \right)^{-1} \right\| \\
&= O_p(\sqrt{p}) + O_p \left(\frac{p}{\sqrt{n}} \right),
\end{aligned}$$

where Lemma 8 (iv) is used.

$$\begin{aligned}
\|a_8\| &= \frac{1}{np} \left\| \mathbf{E}^\top \mathbf{E} \mathbf{E}^\top \hat{\mathbf{A}} \mathbf{G} \right\| \\
&\leq \frac{1}{np} \left\| \mathbf{E}^\top \mathbf{E} \mathbf{E}^\top \mathbf{A}^0 \mathbf{H} \mathbf{G} \right\| + \frac{1}{np} \left\| \mathbf{E}^\top \mathbf{E} \mathbf{E}^\top (\hat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}) \mathbf{G} \right\| \\
&\leq \frac{1}{np} \|\mathbf{E}^\top \mathbf{E}\| \|\mathbf{E}^\top \mathbf{A}^0\| \|\mathbf{H}\| \|\mathbf{G}\| + \frac{1}{np} \|\mathbf{E}^\top \mathbf{E}\| \|\mathbf{E}^\top\| \|\hat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}\| \|\mathbf{G}\| \\
&= \frac{1}{np} [O_p(n\sqrt{p}) + O_p(p\sqrt{n})] O_p(\sqrt{np}) \\
&\quad + \frac{1}{np} [O_p(n\sqrt{p}) + O_p(p\sqrt{n})] O_p(\sqrt{np}) \left[O_p\left(\frac{\sqrt{p} \|\mathbf{C} - \hat{\mathbf{C}}\|}{\sqrt{n}}\right) + O_p\left(\frac{\sqrt{p}}{\min(\sqrt{n}, \sqrt{p})}\right) \right] \\
&= O_p(\sqrt{n}) + O_p(\sqrt{p}) + O_p\left(\frac{p}{\sqrt{n}} \|\mathbf{C} - \hat{\mathbf{C}}\|\right) + O_p(\sqrt{p} \|\mathbf{C} - \hat{\mathbf{C}}\|) + O_p(\sqrt{n}) + O_p\left(\frac{p}{\sqrt{n}}\right) \\
&= O_p\left(\frac{p}{\sqrt{n}} \|\mathbf{C} - \hat{\mathbf{C}}\|\right) + O_p(\sqrt{p} \|\mathbf{C} - \hat{\mathbf{C}}\|) + O_p(\sqrt{n}) + O_p\left(\frac{p}{\sqrt{n}}\right),
\end{aligned}$$

where the order of $\hat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}$ is proved in Proposition 2 and the order of other matrix norms can be found in Lemma 8 (i), (ii) and (iii).

Combining all the terms, we have

$$\|\mathbf{E}^\top (\hat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H})\| = O_p\left(\frac{p}{\min(\sqrt{n}, \sqrt{p})} \|\mathbf{C} - \hat{\mathbf{C}}\|\right) + O_p(\sqrt{n}) + O_p\left(\frac{p}{\sqrt{n}}\right).$$

For (ii), multiplying the matrix \mathbf{F}^\top in the front does not change the order, using the fact that $\|\mathbf{F}^\top \mathbf{E}^\top \Phi\|$ is of the same order as $\|\mathbf{E}^\top \Phi\|$ and that $\|\mathbf{F}^\top \mathbf{E}^\top \mathbf{E}\|$ and $\|\mathbf{F}^\top \mathbf{E}^\top \mathbf{E} \mathbf{F}\|$ are of the same order as $\|\mathbf{E}^\top \mathbf{E}\|$, as proved in Lemma 8. \square

Lemma 12. *Under Assumptions 10-14, we have the following*

(i)

$$\frac{1}{p} \Phi^\top (\hat{A} - A^0 H) = O_p \left(\frac{1}{\sqrt{n}} \|C - \hat{C}\| \right) + O_p \left(\frac{1}{\min(n, p)} \right)$$

(ii)

$$\frac{1}{p} A^{0\top} (\hat{A} - A^0 H) = O_p \left(\frac{1}{\sqrt{n}} \|C - \hat{C}\| \right) + O_p \left(\frac{1}{\min(n, p)} \right)$$

(iii)

$$\frac{1}{p} \hat{A}^\top (\hat{A} - A^0 H) = O_p \left(\frac{1}{\sqrt{n}} \|C - \hat{C}\| \right) + O_p \left(\frac{1}{\min(n, p)} \right)$$

(iv)

$$\frac{1}{p} \Phi^\top M_{\hat{A}} (\hat{A} - A^0 H) = O_p \left(\frac{1}{\sqrt{n}} \|C - \hat{C}\| \right) + O_p \left(\frac{1}{\min(n, p)} \right)$$

Proof. For (i), using (4.45)

$$\Phi^\top (\hat{A} - A^0 H) = \Phi^\top (I1 + I2 + \cdots + I8) G \quad (4.50)$$

It can be easily proved that the first five terms in (4.50) are $O_p \left(\frac{p}{\sqrt{n}} \|C - \hat{C}\| \right)$ using the results from Lemma 7 and 8. Recall that $G = (A^{0\top} \hat{A} / p)^{-1} (F^\top F / n)^{-1}$, and that

$\mathbf{G} = O_p(1)$ from Lemma 10. For the sixth term,

$$\begin{aligned}
\Phi^\top I6G &= \frac{1}{np} \Phi^\top A^0 F^\top E^\top \hat{\mathbf{A}} \mathbf{G} \\
&= \frac{1}{np} \Phi^\top A^0 F^\top E^\top (\hat{\mathbf{A}} - A^0 \mathbf{H}) \mathbf{G} + \frac{1}{np} \Phi^\top A^0 F^\top E^\top A^0 \mathbf{H} \mathbf{G} \\
&\leq \frac{1}{np} \|\Phi^\top\| \|A^0\| \|F^\top E^\top (\hat{\mathbf{A}} - A^0 \mathbf{H})\| \|\mathbf{G}\| + \frac{1}{np} \|\Phi^\top\| \|A^0\| \|F^\top E^\top A^0\| \|\mathbf{H}\| \|\mathbf{G}\| \\
&= O_p\left(\frac{\sqrt{p}}{\sqrt{n}} \|\mathbf{C} - \hat{\mathbf{C}}\|\right) + O_p\left(\frac{\sqrt{p}}{\min(\sqrt{n}, \sqrt{p})}\right) + O_p\left(\frac{\sqrt{p}}{\sqrt{n}}\right),
\end{aligned}$$

using the results from Lemma 11. Next,

$$\begin{aligned}
\Phi^\top I7G &= \frac{1}{np} \Phi^\top E F A^{0\top} \hat{\mathbf{A}} \mathbf{G} \\
&= \frac{1}{n} \Phi^\top E F \left(\frac{F^\top F}{n}\right)^{-1} = O_p\left(\frac{\sqrt{p}}{\sqrt{n}}\right),
\end{aligned}$$

where the order of $\Phi^\top E F$ is found in Lemma 8 (iii).

$$\begin{aligned}
\Phi^\top I8G &= \frac{1}{np} \Phi^\top E E^\top \hat{\mathbf{A}} \mathbf{G} \\
&= \frac{1}{np} \Phi^\top E E^\top (\hat{\mathbf{A}} - A^0 \mathbf{H}) \mathbf{G} + \frac{1}{np} \Phi^\top E E^\top A^0 \mathbf{H} \mathbf{G} \\
&\leq \frac{1}{np} \|\Phi^\top E\| \|E^\top (\hat{\mathbf{A}} - A^0 \mathbf{H})\| \|\mathbf{G}\| + \frac{1}{np} \|\Phi^\top E\| \|E^\top A^0\| \|\mathbf{H}\| \|\mathbf{G}\| \\
&= O_p\left(\frac{1}{\sqrt{n}} \|\mathbf{C} - \hat{\mathbf{C}}\|\right) + O_p\left(\frac{1}{\min(\sqrt{n}, \sqrt{p})}\right) + O_p(1) = O_p(1),
\end{aligned}$$

where Lemma 8 (ii) and Lemma 11 (i) are used. Combining the terms, we have proved (i). The proof for (ii) is exactly the same.

For (iii), we can write

$$\widehat{\mathbf{A}}^\top (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}) = (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H})^\top (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}) + (\mathbf{A}^0 \mathbf{H})^\top (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}).$$

The order of the first term on the right can be found in Proposition 2. The order of the second term on the right is proved in (ii). For (iv), we have

$$\Phi^\top M_{\widehat{\mathbf{A}}} (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}) = \Phi^\top (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}) + \frac{1}{p} \Phi^\top \widehat{\mathbf{A}} \widehat{\mathbf{A}}^\top (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}),$$

where the order of the two terms are proved in (i) and (iii). \square

Lemma 13. Define the matrix

$$\mathbf{Q}(\mathbf{A}) = \frac{1}{p} \Phi^\top M_{\mathbf{A}} \Phi.$$

Under Assumptions 10-13, it holds

$$\mathbf{Q}(\widehat{\mathbf{A}})^{-1} - \mathbf{Q}(\mathbf{A}^0)^{-1} = o_p(1).$$

Proof.

$$\begin{aligned} \mathbf{Q}(\widehat{\mathbf{A}}) - \mathbf{Q}(\mathbf{A}^0) &= \frac{1}{p} \Phi^\top M_{\widehat{\mathbf{A}}} \Phi - \frac{1}{p} \Phi^\top M_{\mathbf{A}^0} \Phi = \frac{1}{p} \Phi^\top (M_{\widehat{\mathbf{A}}} - M_{\mathbf{A}^0}) \Phi \\ &= \frac{1}{p} \Phi^\top (\mathbf{P}_{\mathbf{A}^0} - \mathbf{P}_{\widehat{\mathbf{A}}}) \Phi = O_p(\|\mathbf{P}_{\mathbf{A}^0} - \mathbf{P}_{\widehat{\mathbf{A}}}\|) = o_p(1), \end{aligned}$$

using Theorem 6 (ii). In Assumption 11, we have assumed $\inf_{\mathbf{A}} D(\mathbf{A}) > 0$, since the second term in $D(\mathbf{A})$ is nonnegative, we have $\inf_{\mathbf{A}} \mathbf{Q}(\mathbf{A}) > 0$, so the matrix $\mathbf{Q}(\mathbf{A}^0)$ is

invertible. Therefore,

$$\mathbf{Q}(\hat{\mathbf{A}})^{-1} = \left[\mathbf{Q}(\mathbf{A}^0)^{-1} + o_p(1) \right]^{-1} = \mathbf{Q}(\mathbf{A}^0) + o_p(1).$$

□

Lemma 14. Recall \mathbf{H} defined in Proposition 2, then

$$\mathbf{H}\mathbf{H}^\top = \left(\frac{\mathbf{A}^{0\top}\mathbf{A}^0}{p} \right)^{-1} + O_p\left(\frac{1}{\sqrt{n}} \|\mathbf{C} - \hat{\mathbf{C}}\|\right) + O_p\left(\frac{1}{\min(n, p)}\right)$$

Proof. We have from Lemma 12

$$\frac{1}{p}\mathbf{A}^{0\top}(\hat{\mathbf{A}} - \mathbf{A}^0\mathbf{H}) = O_p\left(\frac{1}{\sqrt{n}} \|\mathbf{C} - \hat{\mathbf{C}}\|\right) + O_p\left(\frac{1}{\min(n, p)}\right), \quad (4.51)$$

and

$$\frac{1}{p}\hat{\mathbf{A}}^\top(\hat{\mathbf{A}} - \mathbf{A}^0\mathbf{H}) = \mathbf{I}_r - \frac{1}{p}\hat{\mathbf{A}}^\top\mathbf{A}^0\mathbf{H} = O_p\left(\frac{1}{\sqrt{n}} \|\mathbf{C} - \hat{\mathbf{C}}\|\right) + O_p\left(\frac{1}{\min(n, p)}\right). \quad (4.52)$$

Left multiply (4.51) by \mathbf{H}^\top and sum with the transpose of (4.52) to obtain

$$\mathbf{I}_r - \frac{1}{p}\mathbf{H}^\top\mathbf{A}^{0\top}\mathbf{A}^0\mathbf{H} = O_p\left(\frac{1}{n} \|\mathbf{C} - \hat{\mathbf{C}}\|\right) + O_p\left(\frac{1}{\min(n, p)}\right).$$

Right multiplying by \mathbf{H}^\top and left multiplying by $\mathbf{H}^{\top-1}$, we obtain

$$\mathbf{I}_r - \frac{1}{p}\mathbf{A}^{0\top}\mathbf{A}^0\mathbf{H}\mathbf{H}^\top = O_p\left(\frac{1}{\sqrt{n}} \|\mathbf{C} - \hat{\mathbf{C}}\|\right) + O_p\left(\frac{1}{\min(n, p)}\right).$$

Then left multiplying $(\mathbf{A}^{0\top} \mathbf{A}^0 / p)^{-1}$, we have

$$\mathbf{H}\mathbf{H}^\top = \left(\frac{\mathbf{A}^{0\top} \mathbf{A}^0}{p} \right)^{-1} + O_p \left(\frac{1}{\sqrt{n}} \|\mathbf{C} - \widehat{\mathbf{C}}\| \right) + O_p \left(\frac{1}{\min(n, p)} \right).$$

□

Lemma 15. Under Assumptions 10-14, when $p/n \rightarrow \rho > 0$,

$$\left\| \frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}} \mathbf{E} - \frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top \mathbf{M}_{\mathbf{A}^0} \mathbf{E} \right\| = \sqrt{p} \times O_p \left(\frac{\|\mathbf{C} - \widehat{\mathbf{C}}\|^2}{n} \right) + o_p(1).$$

Proof. Using

$$\mathbf{M}_{\mathbf{A}^0} = \mathbf{I}_p - \mathbf{A}^0 (\mathbf{A}^{0\top} \mathbf{A}^0)^{-1} \mathbf{A}^{0\top}, \quad \mathbf{M}_{\widehat{\mathbf{A}}} = \mathbf{I}_p - (\widehat{\mathbf{A}} \widehat{\mathbf{A}}^\top) / p,$$

we calculate

$$\begin{aligned} \frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top \mathbf{M}_{\mathbf{A}^0} \mathbf{E} - \frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}} \mathbf{E} &= \frac{1}{p\sqrt{np}} \boldsymbol{\Phi}^\top \widehat{\mathbf{A}} \widehat{\mathbf{A}}^\top \mathbf{E} - \frac{1}{p\sqrt{np}} \boldsymbol{\Phi}^\top \mathbf{A}^0 \left(\frac{\mathbf{A}^{0\top} \mathbf{A}^0}{p} \right)^{-1} \mathbf{A}^{0\top} \mathbf{E} \\ &= \frac{1}{p\sqrt{np}} \left\{ \boldsymbol{\Phi}^\top (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}) \mathbf{H}^\top \mathbf{A}^{0\top} \mathbf{E} \right. \\ &\quad + \boldsymbol{\Phi}^\top (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}) (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H})^\top \mathbf{E} \\ &\quad + \boldsymbol{\Phi}^\top \mathbf{A}^0 \mathbf{H} (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H})^\top \mathbf{E} \\ &\quad \left. + \boldsymbol{\Phi}^\top \mathbf{A}^0 \left[\mathbf{H}\mathbf{H}^\top - \left(\frac{\mathbf{A}^{0\top} \mathbf{A}^0}{p} \right)^{-1} \right] \mathbf{A}^{0\top} \mathbf{E} \right\} \\ &\equiv a + b + c + d, \end{aligned}$$

where in the second equality, we substitute $\widehat{\mathbf{A}}$ with $\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H} + \mathbf{A}^0 \mathbf{H}$. So the first term on

the right-hand side of the first equality is broken down into four terms, one of which is combined with the second term in the right-hand side of the first equality.

For notation simplicity, we denote

$$q = \frac{1}{\sqrt{n}} \|\mathbf{C} - \widehat{\mathbf{C}}\| + \frac{1}{\min(\sqrt{n}, \sqrt{p})}, \quad (4.53)$$

which is used to represent the order in the result of Proposition 2.

We calculate each term:

$$\begin{aligned} \|a\| &= \left\| \frac{1}{p\sqrt{np}} \boldsymbol{\Phi}^\top (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}) \mathbf{H}^\top \mathbf{A}^\top \mathbf{E} \right\| \\ &= \frac{1}{p\sqrt{np}} \times O_p(\sqrt{p}) \times O_p(\sqrt{pq}) \times O_p(\sqrt{np}) \\ &= O_p\left(\frac{\|\mathbf{C} - \widehat{\mathbf{C}}\|}{\sqrt{n}}\right) + \left(\frac{1}{\min(\sqrt{n}, \sqrt{p})}\right) = o_p(1), \end{aligned}$$

where the order of $\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}$ is \sqrt{pq} as proved in Proposition 2 and the order of $\|\boldsymbol{\Phi}\|$ $\|\mathbf{A}^\top \mathbf{E}\|$ can be found in Lemma 7 and 8 (ii) respectively. And when $p/n \rightarrow \rho > 0$,

$$\begin{aligned} \|b\| &= \left\| \frac{1}{p\sqrt{np}} \boldsymbol{\Phi}^\top (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}) (\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H})^\top \mathbf{E} \right\| \\ &\leq \frac{1}{p\sqrt{np}} \times O_p(\sqrt{p}) \times O_p(pq^2) \times O_p(\sqrt{np}) \\ &= \sqrt{p} \times \left[O_p\left(\frac{\|\mathbf{C} - \widehat{\mathbf{C}}\|^2}{n}\right) + O_p\left(\frac{1}{\min(n, p)}\right) \right] \\ &= \sqrt{p} \times O_p\left(\frac{\|\mathbf{C} - \widehat{\mathbf{C}}\|^2}{n}\right) + o_p(1), \end{aligned}$$

where again Proposition 2 and Lemma 7 are used. And

$$\begin{aligned}
c &= \frac{1}{p\sqrt{np}} \Phi^\top A^0 H (\hat{A} - A^0 H)^\top E \\
&= \frac{1}{p\sqrt{np}} \Phi^\top A^0 H H^\top (\hat{A} H^{-1} - A^0)^\top E \\
&= \frac{1}{p\sqrt{np}} \Phi^\top A^0 \left[H H^\top - \left(\frac{A^{0\top} A^0}{p} \right)^{-1} \right] (\hat{A} H^{-1} - A^0)^\top E \\
&\quad + \frac{1}{p\sqrt{np}} \Phi^\top A^0 \left(\frac{A^{0\top} A^0}{p} \right)^{-1} (\hat{A} H^{-1} - A^0)^\top E \\
&\equiv c1 + c2,
\end{aligned}$$

where the second equality is using $\hat{A} - A^0 H = (\hat{A} H^{-1} - A^0) H$. In the third equality, we subtract $(A^{0\top} A^0 / p)^{-1}$ from $H H^\top$ and then add it back.

For $c1$, when $p/n \rightarrow \rho > 0$

$$\begin{aligned}
\|c1\| &= \left\| \frac{1}{p\sqrt{np}} \Phi^\top A^0 \left[H H^\top - \left(\frac{A^{0\top} A^0}{p} \right)^{-1} \right] (\hat{A} H^{-1} - A^0)^\top E \right\| \\
&\leq \frac{1}{p\sqrt{np}} \times O_p(\sqrt{p}) \times O_p(\sqrt{p}) \times \left[O_p\left(\frac{1}{\sqrt{n}} \|C - \hat{C}\|\right) + O_p\left(\frac{1}{\min(n, p)}\right) \right] \\
&\quad \times \left[O_p\left(\frac{p}{\min(\sqrt{n}, \sqrt{p})} \|C - \hat{C}\|\right) + O_p(\sqrt{n}) + O_p\left(\frac{p}{\sqrt{n}}\right) \right] \\
&= O_p\left(\frac{\sqrt{p}}{\min(\sqrt{n}, \sqrt{p})} \frac{\|C - \hat{C}\|}{n}\right) + O_p\left(\frac{1}{\sqrt{p}} \frac{\|C - \hat{C}\|}{\sqrt{n}}\right) + O_p\left(\frac{\sqrt{p}}{n} \frac{\|C - \hat{C}\|}{\sqrt{n}}\right) \\
&\quad + O_p\left(\frac{1}{\sqrt{p}} \frac{1}{\min(n, p)}\right) + O_p\left(\frac{\sqrt{p}}{n} \frac{1}{\min(n, p)}\right) \\
&= o_p(1),
\end{aligned}$$

where the order of Φ , A^0 and E are found in Lemma 7; the order of $H H^\top - (A^{0\top} A^0 / p)^{-1}$

is found in Lemma 14; and the order of $\widehat{\mathbf{A}}\mathbf{H}^{-1} - \mathbf{A}^0$ is found in Proposition 2. Now for c_2 , using the same lemmas and proposition,

$$\begin{aligned} \|c_2\| &= \left\| \frac{1}{p\sqrt{np}} \boldsymbol{\Phi}^\top \mathbf{A}^0 \left(\frac{\mathbf{A}^{0\top} \mathbf{A}^0}{p} \right)^{-1} (\widehat{\mathbf{A}}\mathbf{H}^{-1} - \mathbf{A}^0)^\top \mathbf{E} \right\| \\ &\leq \frac{1}{p\sqrt{np}} O_p(\sqrt{p}) \times O_p(\sqrt{p}) \times \left[O_p \left(\frac{p}{\min(\sqrt{n}, \sqrt{p})} \|\mathbf{C} - \widehat{\mathbf{C}}\| \right) + O_p(\sqrt{n}) + O_p \left(\frac{p}{\sqrt{n}} \right) \right] \\ &= O_p \left(\frac{\sqrt{p}}{\sqrt{n}} \frac{\|\mathbf{C} - \widehat{\mathbf{C}}\|}{\sqrt{n}} \right) + O_p \left(\frac{\sqrt{p}}{n} \right), \end{aligned}$$

which is $o_p(1)$ when $p/n \rightarrow \rho > 0$.

And lastly we have

$$\begin{aligned} \|d\| &= \left\| \frac{1}{p\sqrt{np}} \boldsymbol{\Phi}^\top \mathbf{A}^0 \left[\mathbf{H}\mathbf{H}^\top - \left(\frac{\mathbf{A}^{0\top} \mathbf{A}^0}{p} \right)^{-1} \right] \mathbf{A}^{0\top} \mathbf{E} \right\| \\ &\leq \frac{1}{p\sqrt{np}} O_p(\sqrt{p}) \times O_p(\sqrt{p}) \times O_p \left(\frac{1}{\sqrt{n}} \|\mathbf{C} - \widehat{\mathbf{C}}\| + \frac{1}{\min(n, p)} \right) \times O_p(\sqrt{np}) \\ &= O_p \left(\frac{1}{\sqrt{n}} \|\mathbf{C} - \widehat{\mathbf{C}}\| \right) + O_p \left(\frac{1}{\min(n, p)} \right) = o_p(1), \end{aligned}$$

where again Lemma 14 is used.

Thus combining the above terms, we have

$$\left\| \frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top \mathbf{M}_{\widehat{\mathbf{A}}} \mathbf{E} - \frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top \mathbf{M}_{\mathbf{A}^0} \mathbf{E} \right\| = \sqrt{p} \times O_p \left(\frac{\|\mathbf{C} - \widehat{\mathbf{C}}\|^2}{n} \right) + o_p(1)$$

when $p/n \rightarrow \rho > 0$. □

Lemma 16. Recall J_8 defined in (4.27), we have

$$\|J_8\| = o_p\left(\|C - \widehat{C}\|\right) + O_p\left(\frac{1}{\min(n, p)}\right) + O_p\left(\frac{\sqrt{n}}{\sqrt{p}} \frac{1}{\min(n, p)}\right).$$

Proof.

$$\begin{aligned} J_8 &= -\frac{1}{p} \Phi^\top M_{\widehat{A}} I_8 G F^\top \\ &= -\frac{1}{np^2} \Phi^\top M_{\widehat{A}} E E^\top \widehat{A} G F^\top \\ &= -\frac{1}{np^2} \Phi^\top E E^\top \widehat{A} G F^\top + \frac{1}{np^3} \Phi^\top \widehat{A} \widehat{A}^\top E E^\top \widehat{A} G F^\top \\ &\equiv I + II, \end{aligned}$$

where we use $M_{\widehat{A}} = I_p - \widehat{A} \widehat{A}^\top / p$. For I ,

$$I = -\frac{1}{np^2} \Phi^\top E E^\top (\widehat{A} - A^0 H) G F^\top - \frac{1}{np^2} \Phi^\top E E^\top A^0 H G F^\top,$$

then

$$\begin{aligned} \|I\| &\leq \frac{1}{np^2} \left\| \Phi^\top E E^\top \right\| \|\widehat{A} - A^0 H\| \|G\| \|F\| + \frac{1}{np^2} \left\| \Phi^\top E E^\top A^0 \right\| \|H\| \|G\| \|F\| \\ &= \frac{1}{np^2} \times [O_p(p\sqrt{n}) + O_p(n\sqrt{p})] \times O_p(\sqrt{pq}) \times O_p(\sqrt{n}) \\ &\quad + \frac{1}{np^2} \times [O_p(p\sqrt{n}) + O_p(n\sqrt{p})] \times O_p(\sqrt{n}) \\ &= O_p\left(\frac{1}{\sqrt{p}} q\right) + O_p\left(\frac{\sqrt{n}}{p} q\right), \end{aligned}$$

where the order of $\|\Phi^\top E E^\top\|$ and $\|\Phi^\top E E^\top A^0\|$ are found in Lemma 8; the order of $\|\widehat{A} - A^0 H\|$ is from Proposition 2; and the orders of $\|F\|$ and $\|G\|$ are found in Lemma 7 (iii) and Lemma 10 respectively.

For II ,

$$\begin{aligned} II &= \frac{1}{np^3} \Phi^\top \widehat{\mathbf{A}} \left(\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H} + \mathbf{A}^0 \mathbf{H} \right)^\top \mathbf{E} \mathbf{E}^\top \left(\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H} + \mathbf{A}^0 \mathbf{H} \right) \mathbf{G} \mathbf{F}^\top \\ &= \frac{1}{np^3} \Phi^\top \widehat{\mathbf{A}} \left[\left(\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H} \right)^\top \mathbf{E} \mathbf{E}^\top \left(\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H} \right) + \left(\mathbf{A}^0 \mathbf{H} \right)^\top \mathbf{E} \mathbf{E}^\top \left(\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H} \right) \right. \\ &\quad \left. + \left(\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H} \right)^\top \mathbf{E} \mathbf{E}^\top \mathbf{A}^0 \mathbf{H} + \left(\mathbf{A}^0 \mathbf{H} \right)^\top \mathbf{E} \mathbf{E}^\top \mathbf{A}^0 \mathbf{H} \right] \mathbf{G} \mathbf{F}^\top, \end{aligned}$$

then

$$\begin{aligned} \|II\| &\leq \frac{1}{np^3} \|\Phi^\top\| \|\widehat{\mathbf{A}}\| \left[\|\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}\|^2 \|\mathbf{E} \mathbf{E}^\top\| + \|\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}\| \|\mathbf{E} \mathbf{E}^\top \mathbf{A}^0\| + \|\mathbf{A}^0{}^\top \mathbf{E} \mathbf{E}^\top \mathbf{A}^0\| \right] \|\mathbf{G}\| \|\mathbf{F}\| \\ &= \frac{1}{np^2} \{ [O_p(p\sqrt{n}) + O_p(n\sqrt{p})] \times (pq^2 + \sqrt{pq} + 1) \} \times O_p(\sqrt{n}) \\ &= O_p \left[\left(\frac{1}{p} + \frac{\sqrt{n}}{p\sqrt{p}} \right) (pq^2 + \sqrt{pq}) \right], \end{aligned}$$

where the order of $\|\mathbf{E} \mathbf{E}^\top \mathbf{A}^0\|$ and $\|\mathbf{A}^0{}^\top \mathbf{E} \mathbf{E}^\top \mathbf{A}^0\|$ are found in Lemma 8; the order of $\|\widehat{\mathbf{A}} - \mathbf{A}^0 \mathbf{H}\|$ is from Proposition 2; and the orders of $\|\Phi\|$, $\|\mathbf{F}\|$ and $\|\mathbf{G}\|$ are found in Lemma 7 (i), (iii) and Lemma 10 respectively.

Combining I and II , we have

$$\|J8\| = O_p \left[\left(\frac{1}{p} + \frac{\sqrt{n}}{p\sqrt{p}} \right) (pq^2 + \sqrt{pq}) \right].$$

Since $1 = O(\sqrt{pq})$, the term \sqrt{pq} is dominated by pq^2 , thus

$$\begin{aligned}
 \|J8\| &= O_p \left[\left(\frac{1}{p} + \frac{\sqrt{n}}{p\sqrt{p}} \right) pq^2 \right] \\
 &= O_p \left[\left(1 + \frac{\sqrt{n}}{\sqrt{p}} \right) \left(\frac{\|\mathbf{c} - \hat{\mathbf{c}}\|^2}{n} + \frac{1}{\min(n, p)} \right) \right] \\
 &= o_p \left(\|\mathbf{c} - \hat{\mathbf{c}}\| \right) + O_p \left(\frac{1}{\min(n, p)} \right) + O_p \left(\frac{\sqrt{n}}{\sqrt{p}} \frac{1}{\min(n, p)} \right).
 \end{aligned}$$

□

Conclusions and Future Work

In this thesis, we consider three various topics in the field of high-dimensional functional data analysis. In Chapter 2, we propose a model that utilizes the correlation structures between the populations, and also ensures long-term coherence in the forecasts. Based on functional principal component analysis, we propose using a vector error correction model to jointly forecast mortality rates in multiple populations. An algorithm to generate bootstrap prediction intervals is also provided. We compare the proposed model with other forecast models in the previous literature. The superiority of this model is demonstrated through a series of simulation studies and applications to the age- and sex-specific mortality rates in Switzerland and the Czech Republic.

A main focus of this thesis is on the high-dimensional functional data. On the one hand, we consider extending multivariate functional time series to high-dimensional functional time series. By high dimension, we mean we allow the dimension of the multivariate functional time series N to grow with the sample size T . In this sense, the data structure is analogous to the target in conventional high-dimensional data analysis. In Chapter 3, we adopt a twofold dimension reduction model for such data. A dynamic functional principal component analysis is first applied to reduce each functional time series to a vector. We then apply the factor model as a further dimension reduction technique so that only a small number of latent factors are preserved. Classic time series

models can be used to forecast the factors, and conditional forecasts of the functions can be constructed. Asymptotic properties of the approximated functions are established, including both estimation error and forecast error. The proposed method is easy to implement, especially when the dimension of the functional time series is large. We show the superiority of our approach via both simulation studies and an application to Japanese age-specific mortality rates.

On the other hand, functional data by itself is considered to be high-dimensional. Such infinite dimensionality calls for dimension reduction techniques, with functional principal component analysis the most commonly used method to achieve this. This high dimensionality, in turn, means that when measurement error exists, it is also of high dimension and will induce the problem of the "curse of dimensionality." Thus, dimension reduction techniques from the field of multivariate analysis are needed. This is one of the main motivations for the factor-augmented smoothing model proposed in Chapter 4. To retrieve the underlying smooth functions, we impose a factor model structure on the measurement error while smoothing the functional component. Asymptotic theorems are also established to demonstrate the effects of including factor structures on the smoothing results. Specifically, we show that the smoothing coefficients projected on the complement space of the factor loading matrix is asymptotic normal. As a byproduct of independent interest, an estimator for the population covariance matrix of the raw data is presented based on the proposed model. Extensive simulation studies illustrate that these factor adjustments are crucial in improving estimation accuracy and avoiding the "curse of dimensionality." The advantages of our model are also shown using Canadian weather data and Australian temperature data.

The work presented in this thesis leaves several directions open for future research. In Chapter 2 and 3, we consider representing multivariate and high-dimensional functional

time series with low-order principal components under unsupervised settings. When covariates exist, the proposed models can be easily extended to functional linear models that incorporate the covariates. For instance, in Chapter 3, we analyze the Japanese mortality rates for all the prefectures, in which case, the socio-economic features of each prefecture can be included as covariates. Models with functional responses that associate the principal components with predictors are studied in Chiou et al. (2003a,b, 2004). In Chapter 3, the functional principal components obtained from the first dimension reduction step are in a panel data structure. Thus, covariates can be easily incorporated into the second dimension reduction step with factor model. There are many studies on the panel data models with fixed and random effects (see, e.g., Frees 2004, Bai 2009).

With the increasing ability to collect large data, the complexity of data also increases. The factor-augmented smoothing model proposed in Chapter 4 is a good start point for modeling complex data structures. The data we deal with are a mixture of smooth functional curves and high-dimensional measurement error. The factor model component can be regarded as a "boosting" component that improves model accuracy. Extending from this idea, the model can be applied to other data structures. One example is the data that contain change points. Change point is a popular problem in many statistics and econometric topics and has been extensively studied in the multivariate setting. Previous literature on change point in functional data include Berkes et al. (2009), Hörmann & Kokoszka (2010). It is shown in Chapter 4 that our model can be used for modeling functional data with change point in the cross-sectional direction. The model can be modified to account for change point also in the sample direction. Further research can be conducted along this line.

Bibliography

- Ahn, S. C. & Horenstein, A. R. (2013), 'Eigenvalue ratio test for the number of factors', *Econometrica* **81**(3), 1203–1227. (cited on page 105)
- Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE transactions on automatic control* **19**(6), 716–723. (cited on page 106)
- Amengual, D. & Watson, M. W. (2007), 'Consistent estimation of the number of dynamic factors in a large n and t panel', *Journal of Business & Economic Statistics* **25**(1), 91–96. (cited on page 5)
- Anderson, T. (1963), 'The use of factor analysis in the statistical analysis of multiple time series', *Psychometrika* **28**(1), 1–25. (cited on pages 4 and 44)
- Andrews, D. (1991), 'Heteroskedasticity and autocorrelation consistent covariance matrix estimation', *Econometrica* **59**(3), 817–858. (cited on page 51)
- Andrews, D. & Monahan, J. (1992), 'An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator', *Econometrica* **60**(4), 953–966. (cited on page 51)
- Aneiros-Pérez, G., Cao, R. & Vilar-Fernández, J. M. (2011), 'Functional methods for time series prediction: a nonparametric approach', *Journal of Forecasting* **30**(4), 377–392. (cited on page 5)
- Aneiros-Pérez, G. & Vieu, P. (2008), 'Nonparametric time series prediction: A semi-

- functional partial linear modeling', *Journal of Multivariate Analysis* **99**(5), 834–857. (cited on pages 2 and 42)
- Aue, A., Horváth, L. & Pellatt, D. F. (2017), 'Functional generalized autoregressive conditional heteroskedasticity', *Journal of Time Series Analysis* **38**(1), 3–21. (cited on pages 2 and 42)
- Aue, A., Norinho, D. D. & Hörmann, S. (2015), 'On the prediction of stationary functional time series', *Journal of the American Statistical Association* **110**(509), 378–392. (cited on pages 3, 13, 14, and 15)
- Bai, J. (2003), 'Inferential theory for factor models of large dimensions', *Econometrica* **71**(1), 135–171. (cited on pages 4 and 44)
- Bai, J. (2009), 'Panel data models with interactive fixed effects', *Econometrica* **77**(4), 1229–1279. (cited on page 173)
- Bai, J. & Ng, S. (2002), 'Determining the number of factors in approximate factor models', *Econometrica* **70**(1), 191–221. (cited on pages 4, 5, and 105)
- Bai, J. & Ng, S. (2007), 'Determining the number of primitive shocks in factor models', *Journal of Business & Economic Statistics* **25**(1), 52–60. (cited on page 5)
- Bathia, N., Yao, Q. & Ziegelmann, F. (2010), 'Identifying the finite dimensionality of curve time series', *The Annals of Statistics* **38**(6), 3352–3386. (cited on page 5)
- Berkes, I., Gabrys, R., Horváth, L. & Kokoszka, P. (2009), 'Detecting changes in the mean of functional observations', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(5), 927–946. (cited on page 173)

-
- Berrendero, J., Justel, A. & Svarc, M. (2011), 'Principal components for multivariate functional data', *Computational Statistics and Data Analysis* **55**(9), 2619–2634. (cited on page 43)
- Besse, P. C., Cardot, H. & Stephenson, D. B. (2000), 'Autoregressive forecasting of some functional climatic variations', *Scandinavian Journal of Statistics* **27**(4), 673–687. (cited on pages 2 and 41)
- Bickel, P. J. & Levina, E. (2008), 'Regularized estimation of large covariance matrices', *The Annals of Statistics* **36**(1), 199–227. (cited on page 113)
- Boente, G. & Fraiman, R. (2000), 'Kernel-based functional principal components', *Statistics & probability letters* **48**(4), 335–345. (cited on page 6)
- Booth, H., Maindonald, J. & Smith, L. (2002), 'Applying Lee-Carter under conditions of variable mortality decline', *Population Studies* **56**(3), 325–336. (cited on page 9)
- Bosq, D. (2012), *Linear Processes in Function Spaces: Theory and Applications*, Vol. 149, Springer Science & Business Media, New York. (cited on pages 1, 15, and 41)
- Bosq, D. & Blanke, D. (2007), *Inference and Prediction in Large Dimensions*, Vol. 754, John Wiley & Sons. (cited on pages 1 and 41)
- Box, G. E., Jenkins, G. M., Reinsel, G. C. & Ljung, G. M. (2015), *Time Series Analysis: Forecasting and Control, 5th ed.*, John Wiley & Sons. (cited on page 19)
- Brillinger, D. (1981), *Time Series Data Analysis and Theory*, Holder-Day, San Francisco. (cited on pages 4 and 44)
- Cai, T. T. & Yuan, M. (2011), 'Optimal estimation of the mean function based on discretely

- sampled functional data: Phase transition', *The Annals of Statistics* **39**(5), 2330–2355. (cited on page 90)
- Cairns, A. J. G., Blake, D. & Dowd, K. (2006), 'A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration', *Journal of Risk and Insurance* **73**(4), 687–718. (cited on page 9)
- Carter, L. R. & Lee, R. D. (1992), 'Modelling and forecasting us sex differentials in modeling', *International Journal of Forecasting* **8**(3), 393–411. (cited on page 10)
- Castro, B. F. D., Guillas, S. & Manteiga, W. G. (2005), 'Functional samples and bootstrap for predicting sulfur dioxide levels', *Technometrics* **47**(2), 212–222. (cited on page 5)
- Cattell, R. B. (1966), 'The scree test for the number of factors', *Multivariate behavioral research* **1**(2), 245–276. (cited on page 5)
- Chamberlain, G. (1983), 'Funds, factors, and diversification in arbitrage pricing models', *Econometrica* **51**(5), 1305–1323. (cited on pages 4 and 44)
- Chamberlain, G. & Rothschild, M. (1983), 'Arbitrage, factor structure, and mean-variance analysis on large asset markets', *Econometrica* **51**(5), 1281–1304. (cited on page 4)
- Chan, W., Li, J. S. & Li, J. (2014), 'The CBD mortality indexes: modeling and applications', *North American Actuarial Journal* **18**(1), 38–58. (cited on page 9)
- Chiou, J. M. (2012), 'Dynamical functional prediction and classification with application to traffic flow prediction', *The Annals of Applied Statistics* **6**(4), 1588–1614. (cited on page 38)
- Chiou, J. M., Chen, Y. T. & Yang, Y. F. (2014), 'Multivariate functional principal component

-
- analysis: A normalization approach', *Statistica Sinica* **24**(4), 1571–1596. (cited on page 43)
- Chiou, J. M. & Müller, H. G. (2014), 'Linear manifold modelling of multivariate functional data', *Journal of the Royal Society of Statistics. Series B (Statistical Methodology)* **76**(3), 605–626. (cited on page 11)
- Chiou, J. M., Müller, H. G. & Wang, J. L. (2003a), 'A functional multiplicative effects model for longitudinal data, with application to reproductive histories of female medflies', *Statistica Sinica* **13**(4), 1119. (cited on page 173)
- Chiou, J. M., Müller, H. G. & Wang, J. L. (2003b), 'Functional quasi-likelihood regression models with smooth random effects', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(2), 405–423. (cited on page 173)
- Chiou, J. M., Müller, H. G. & Wang, J. L. (2004), 'Functional response models', *Statistica Sinica* **14**(3), 675–693. (cited on page 173)
- Connor, G. & Korajczyk, R. A. (1993), 'A test for the number of factors in an approximate factor model', *the Journal of Finance* **48**(4), 1263–1291. (cited on page 5)
- Cuevas, A. (2014), 'A partial overview of the theory of statistics with functional data', *Journal of Statistical Planning and Inference* **147**, 1–23. (cited on page 1)
- Danesi, I. L., Haberman, S. & Millossovich, P. (2015), 'Forecasting mortality in subpopulations using Lee-Carter type models: A comparison', *Insurance: Mathematics and Economics* **62**, 151–161. (cited on pages 10 and 22)
- Delaigle, A., Hall, P., Huang, W. & Kneip, A. (2020), 'Estimating the covariance of fragmented and other related types of functional data', *Journal of the American Statistical Association* **00**(0), 1–19. (cited on page 6)

- Diebold, F. X. & Mariano, R. S. (1995), 'Comparing predictive accuracy', *Journal of Business & Economic Statistics* **13**(3), 253–263. (cited on page 33)
- Eubank, R. L. (1999), *Nonparametric Regression and Spline Smoothing*, 2nd edn, Marcel Dekker, New York. (cited on pages 6 and 90)
- Fan, J., Fan, Y. & Lv, J. (2008), 'High dimensional covariance matrix estimation using a factor model', *Journal of Econometrics* **147**(1), 186–197. (cited on pages 7, 91, 103, 114, and 115)
- Fan, J. & Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London. (cited on pages 6 and 90)
- Fan, J., Liao, Y. & Mincheva, M. (2011), 'High dimensional covariance matrix estimation in approximate factor models', *The Annals of Statistics* **39**(6), 3320–3356. (cited on pages 7 and 115)
- Faraway, J. J. (2016), 'Does data splitting improve prediction?', *Statistics and Computing* **26**(1), 49–60. (cited on page 21)
- Febrero-Bande, M., Galeano, P. & González-Manteiga, W. (2017), 'Functional principal component regression and functional partial least-squares regression: An overview and a comparative study', *International Statistical Review* **85**(1), 61–83. (cited on page 1)
- Ferraty, F. & Vieu, P. (2006), *Nonparametric Functional Data Analysis: Theory and Practice*, Springer Science & Business Media. (cited on pages 1, 2, and 41)
- Forni, M., Hallin, M., Lippi, M. & Reichlin, L. (2000), 'The generalized dynamic-factor model: Identification and estimation', *Review of Economics and statistics* **82**(4), 540–554. (cited on page 4)

-
- Frees, E. W. (2004), *Longitudinal and panel data: analysis and applications in the social sciences*, Cambridge University Press, Cambridge. (cited on page 173)
- Gneiting, T. & Raftery, A. E. (2007), 'Strictly proper scoring rules, prediction, and estimation', *Journal of the American Statistical Association* **102**(477), 359–378. (cited on page 22)
- Goia, A. & Vieu, P. (2016), 'An introduction to recent advances in high/infinite dimensional statistics', *Journal of Multivariate Analysis* **146**, 1–6. (cited on page 1)
- Golub, G. H., Heath, M. & Wahba, G. (1979), 'Generalized cross-validation as a method for choosing a good ridge parameter', *Technometrics* **21**(2), 215–223. (cited on page 106)
- Golub, G. H. & Loan, C. V. (2012), *Matrix computations*, 4th edn, Johns Hopkins University Press, Baltimore. (cited on pages 67 and 70)
- Granger, C. W. & Joyeux, R. (1980), 'An introduction to long-memory time series models and fractional differencing', *Journal of Time Series Analysis* **1**(1), 15–29. (cited on page 19)
- Green, P. J. & Silverman, B. W. (1999), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall, London. (cited on page 90)
- Hall, P. & Hosseini-Nasab, M. (2006), 'On properties of functional principal components analysis', *Journal of the Royal Statistical Society, Statistical Methodology, Series B* **68**(1), 109–126. (cited on page 42)
- Hall, P., Müller, H. G. & Wang, J. L. (2006), 'Properties of principal component methods for functional and longitudinal data analysis', *The Annals of Statistics* **34**(3), 1493–1517. (cited on page 42)

- Hall, P. & Vial, C. (2006), 'Assessing the finite dimensionality of functional data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(4), 689–705. (cited on pages 5 and 38)
- Hörmann, S., Horváth, L. & Reeder, R. (2013), 'A functional version of the ARCH model', *Econometric Theory* **29**(2), 267–288. (cited on pages 2 and 42)
- Hörmann, S., Kidziński, L. & Hallin, M. (2015), 'Dynamic functional principal components', *Journal of the Royal Statistical Society, Statistical Methodology, Series B* **77**(2), 319–348. (cited on pages 4, 42, and 67)
- Hörmann, S. & Kokoszka, P. (2010), 'Weakly dependent functional data', *The Annals of Statistics* **38**(3), 1845–1884. (cited on pages 39, 71, and 173)
- Horváth, L. & Kokoszka, P. (2012), *Inference for functional data with applications*, Vol. 200, Springer Science & Business Media. (cited on page 1)
- Horvath, L., Kokoszka, P. & Rice, G. (2014), 'Testing stationarity of functional time series', *Journal of Econometrics* **179**(1), 66–82. (cited on page 28)
- Hosking, J. R. (1981), 'Fractional differencing', *Biometrika* **68**(1), 165–176. (cited on page 19)
- Human Mortality Database (2016), *University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany)*. Accessed at 8 March 2016. URL: <http://www.mortality.org>. (cited on page 27)
- Hyndman, R. J. & Booth, H. (2008), 'Stochastic population forecasts using functional data models for mortality, fertility and migration', *International Journal of Forecasting* **24**(3), 323–342. (cited on page 29)

-
- Hyndman, R. J., Booth, H. & Yasmeen, F. (2013), 'Coherent mortality forecasting: the product-ratio method with functional time series models', *Demography* **50**(1), 261–283. (cited on pages 2, 3, 11, and 29)
- Hyndman, R. J. & Khandakar, Y. (2008), 'Automatic time series forecasting: The forecast package for R', *Journal of Statistical Software* **27**(3). (cited on page 14)
- Hyndman, R. J. & Shang, H. L. (2009), 'Forecasting functional time series', *Journal of the Korean Statistical Society* **38**, 199–211. (cited on pages 3 and 5)
- Hyndman, R. J. & Shang, H. L. (2010), 'Rainbow plots, bagplots, and boxplots for functional data', *Journal of Computational and Graphical Statistics* **19**(1), 29–45. (cited on page 28)
- Hyndman, R. J. & Ullah, M. S. (2007), 'Robust forecasting of mortality and fertility rates: A functional data approach', *Computational Statistics and Data Analysis* **51**(10), 4942–4956. (cited on pages 3, 11, 20, 28, 60, 63, and 64)
- Japanese Mortality Database (2017), *National Institute of Population and Social Security Research*. Accessed at 8 March 2016. URL: <http://www.ipss.go.jp/p-toukei/JMD/index-en.html>. (cited on page 62)
- Johansen, S. (1991), 'Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models', *Econometrica* **59**(6), 1551–1580. (cited on page 36)
- Klepsch, J. & Klüppelberg, C. (2017), 'An innovations algorithm for the prediction of functional linear processes', *Journal of Multivariate Analysis* **155**, 252–271. (cited on pages 1 and 41)
- Klepsch, J., Klüppelberg, C. & Wei, T. (2017), 'Prediction of functional ARMA processes

- with an application to traffic data', *Econometrics and Statistics* **1**, 128–149. (cited on pages 2 and 41)
- Knight, K. & Fu, W. (2000), 'Asymptotics for lasso-type estimators', *The Annals of Statistics* **28**(5), 1356–1378. (cited on page 108)
- Koissi, M. C., Shapiro, A. F. & Högnäs, G. (2006), 'Evaluating and extending the Lee–Carter model for mortality forecasting: Bootstrap confidence interval', *Insurance: Mathematics and Economics* **38**(1), 1–20. (cited on page 9)
- Kokoszka, P. & Reimherr, M. (2013), 'Determining the order of the functional autoregressive model', *Journal of Time Series Analysis* **34**(1), 116–129. (cited on pages 1 and 41)
- Kokoszka, P., Rice, G. & Shang, H. L. (2017), 'Inference for the autocovariance of a functional time series under conditional heteroscedasticity', *Journal of Multivariate Analysis* **162**, 32–50. (cited on pages 2 and 42)
- Kowal, D. R., Matteson, D. & Ruppert, D. (2017), 'A bayesian multivariate functional dynamic linear model', *Journal of the American Statistical Association* **112**(518), 733–744. (cited on page 2)
- Lam, C., Yao, Q. & Bathia, N. (2011), 'Estimation of latent factors for high-dimensional time series', *Biometrika* **98**(4), 901–918. (cited on pages 4, 44, 47, 51, and 91)
- Lee, R. D. & Carter, L. R. (1992), 'Modeling and forecasting U. S. mortality', *Journal of the American Statistical Association* **87**(419), 659–671. (cited on page 9)
- Li, D., Robinson, P. M. & Shang, H. L. (2020), 'Long-range dependent curve time series', *Journal of the American Statistical Association* **115**(530), 957–971. (cited on pages 2 and 41)

-
- Li, N. & Lee, R. (2005), 'Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method', *Demography* **42**(3), 575–594. (cited on pages 10 and 19)
- Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T. & Cohen, K. L. (1999), 'Robust principal component analysis for functional data', *Test* **8**(1), 1–73. (cited on page 42)
- Lütkepohl, H. (2005), *New Introduction to Multiple Time Series Analysis*, Springer, New York. (cited on pages 3, 14, 16, and 18)
- M. Hallin, M. & Liška, R. (2007), 'Determining the number of factors in the general dynamic factor model', *Journal of the American Statistical Association* **102**(478), 603–617. (cited on page 5)
- Onatski, A. (2010), 'Determining the number of factors from empirical distribution of eigenvalues', *The Review of Economics and Statistics* **92**(4), 1004–1016. (cited on page 105)
- Onatski, A. (2012), 'Asymptotics of the principal components estimator of large factor models with weakly influential factors', *Journal of Econometrics* **168**(2), 244–258. (cited on page 104)
- Pan, J. & Yao, Q. (2008), 'Modelling multiple time series via common factors', *Biometrika* **95**(2), 365–379. (cited on page 4)
- Panaretos, V. M. & Tavakoli, S. (2013), 'Cramér-Karhunen-Loève representation and harmonic principal component analysis of functional time series', *Stochastic Processes and their Applications* **123**(7), 2779–2807. (cited on page 42)
- Paparoditis, E. (2018), 'Sieve bootstrap for functional time series', *The Annals of Statistics* **46**(6B), 3510–3538. (cited on page 5)

- Peña, D. & Box, G. E. (1987), 'Identifying a simplifying structure in time series', *Journal of the American statistical Association* **82**(399), 836–843. (cited on page 4)
- Peña, D. & Poncela, P. (2006), 'Nonstationary dynamic factor analysis', *Journal of Statistical Planning and Inference* **136**(4), 1237–1257. (cited on page 4)
- Poskitt, D. S. & Sengarapillai, A. (2013), 'Description length and dimensionality reduction in functional data analysis', *Computational Statistics & Data Analysis* **58**, 98–113. (cited on page 5)
- Priestley, M., Rao, T. & Tong, J. (1974), 'Applications of principal component analysis and factor analysis in the identification of multivariable systems', *IEEE Transactions on Automatic Control* **19**(6), 730–734. (cited on pages 4 and 44)
- Ramsay, J. O. & Hooker, G. (2017), *Dynamic Data Analysis: Modeling Data with Differential Equations*, Springer, New York. (cited on page 1)
- Ramsay, J. O. & Silverman, B. W. (2002), *Applied Functional Data Analysis*, Springer, New York. (cited on page 1)
- Ramsay, J. O. & Silverman, B. W. (2005), *Functional Data Analysis*, Springer, New York. (cited on pages 1, 6, 10, and 43)
- Reiss, P. T., Goldsmith, J., Shang, H. L. & Ogden, R. T. (2017), 'Methods for scalar-on-function regression', *International Statistical Review* **85**(2), 228–249. (cited on page 1)
- Renshaw, A. E. & Haberman, S. (2006), 'A cohort-based extension to the Lee-Carter model for mortality reduction factors', *Insurance: Mathematics and Economics* **38**(3), 556–570. (cited on page 9)

-
- Rice, G. & Shang, H. L. (2017), 'A plug-in bandwidth selection procedure for long run covariance estimation with stationary functional time series', *Journal of Time Series Analysis* **38**(4), 591–609. (cited on page 42)
- Rice, J. & Silverman, B. (1991), 'Estimating the mean and covariance structure nonparametrically when the data are curves', *Journal of the Royal Statistical Society. Series B (Methodological)* **53**(1), 233–243. (cited on pages 11 and 38)
- Schwarz, G. (1978), 'Estimating the dimension of a model', *The Annals of Statistics* **6**(2), 461–464. (cited on page 106)
- Shang, H. L. (2014), 'A survey of functional principal component analysis', *AStA Advances in Statistical Analysis* **98**(2), 121–142. (cited on page 1)
- Shang, H. L. & Hyndman, R. J. (2017), 'Grouped functional time series forecasting: An application to age-specific mortality rates', *Journal of Computational and Graphical Statistics* **26**(2), 330–343. (cited on page 3)
- Shibata, R. (1981), 'An optimal selection of regression variables', *Biometrika* **68**(1), 45–54. (cited on page 5)
- Tu, I. P., Chen, H. & Chen, X. (2009), 'An eigenvector variability plot', *Statistica Sinica* **19**, 1741–1754. (cited on page 5)
- Viviani, R., Grön, G. & Spitzer, M. (2005), 'Functional principal component analysis of fMRI data', *Human Brain Mapping* **24**(2), 109–129. (cited on page 42)
- Wahba, G. (1975), 'Smoothing noisy data with spline function', *Numerische Mathematik* **24**(5), 383–393. (cited on page 11)

- Wahba, G. (1990), *Spline Models for Observational data*, Vol. 59, Siam. (cited on pages 6 and 90)
- Wand, M. P. & Jones, C. M. (1995), *Kernel Smoothing*, Chapman & Hall. (cited on pages 6 and 90)
- Wang, J.-L., Chiou, J.-M. & Müller, H.-G. (2016), 'Functional data analysis', *Annual Review of Statistics and Its Application* **3**, 257–295. (cited on page 1)
- Wong, F., Carter, C. K. & Kohn, R. (2003), 'Efficient estimation of covariance selection models', *Biometrika* **90**(4), 809–830. (cited on page 113)
- Wood, S. (1994a), 'Monotonic smoothing splines fitted by cross validation', *SIAM Journal of Statistic Computation* **15**(5), 1126–1133. (cited on page 63)
- Wood, S. N. (1994b), 'Monotonic smoothing splines fitted by cross validation', *SIAM Journal of Scientific Computation* **15**(5), 1126–1133. (cited on page 28)
- Yang, S. S. & Wang, C. (2013), 'Pricing and securitization of multi-country longevity risk with mortality dependence', *Insurance: Mathematics and Economics* **52**(2), 157–169. (cited on pages 3, 10, 11, and 12)
- Yao, F. & Lee, T. C. M. (2006), 'Penalized spline models for functional principal component analysis', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **68**(1), 3–25. (cited on page 39)
- Yao, F., Müller, H. & Wang, J. L. (2005a), 'Functional data analysis for sparse longitudinal data', *Journal of the American Statistical Association* **100**(470), 577–590. (cited on pages 6, 38, 39, and 42)

-
- Yao, F., Müller, H. & Wang, J. L. (2005b), 'Functional linear regression analysis for longitudinal data', *The Annals of Statistics* pp. 2873–2903. (cited on page 6)
- Yao, W. & Li, R. (2013), 'New local estimation procedure for a non-parametric regression function for longitudinal data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(1), 123–138. (cited on page 90)
- Zhang, J. T. & Chen, J. (2007), 'Statistical inferences for functional data', *The Annals of Statistics* **35**(3), 1052–1079. (cited on page 6)
- Zhang, X. & Wang, J. L. (2016), 'From sparse to dense functional data and beyond', *The Annals of Statistics* **44**(5), 2281–2321. (cited on pages 6 and 90)
- Zhou, R., Wang, Y., Kaufhold, K., Li, J. S. H. & Tan, K. S. (2014), 'Modeling mortality of multiple populations with vector error correction models: application to solvency II', *North American Actuarial Journal* **18**(1), 150–167. (cited on pages 3, 10, 11, and 12)
- Zhu, T. & Politis, D. N. (2017), 'Kernel estimates of nonparametric functional autoregression models and their bootstrap approximation', *Electronic Journal of Statistics* **11**(2), 2876–2906. (cited on page 53)