Institute and Faculty of Actuaries

PhD studentship output

Funded by the Institute and Faculty of Actuaries

Gaussian Process and Functional Data Methods for Mortality Modelling

Ruhao Wu

Thesis submitted for the degree of

Doctor of Philosophy

Department of Mathematics

University of Leicester

June 2016

Abstract

Modelling the demographic mortality trends is of great importance due to its considerable impact on welfare policy, resource allocation and government planning. In this thesis, we propose to use various statistical methods, including Gaussian process (GP), principal curve, multilevel functional principal component analysis (MFPCA) for forecasting and clustering of human mortality data. This thesis is actually composed of three main topics regarding mortality modelling. In the first topic, we propose a new Gaussian process regression method and apply it to the modelling and forecasting of age-specific human mortality rates for a single population. The proposed method incorporates a weighted mean function and the spectral mixture covariance function, hence provides better performance in forecasting long term mortality rates, compared with the conventional GPR methods. The performance of the proposed method is also compared with Lee-Miller model and the functional data model by Hyndman and Ullah (2007) in the context of forecasting the French total mortality rates. Then, in the second topic, we extend mortality modelling for a single population independently to that for multiple populations simultaneously, by developing a new framework for coherent modelling and forecasting of mortality rates for multiple subpopulations within one large population. We treat the mortality of subpopulations as multilevel functional data and then a weighted multilevel functional principal component approach is proposed and used for modelling and forecasting the mortality rates. The proposed model is applied to sex-specific data for nine developed countries, and the forecasting results suggest that, in terms of overall accuracy, the model outperforms the independent model (Hyndman and Ullah 2007) and is comparable to the Product-Ratio model (Hyndman et al 2013) but with several advantages. Finally, in the third topic, we introduce a clustering method based on principal curves for clustering of human mortality as functional data. And this innovative clustering method is applied to French total mortality data for exploring its potential features.

Key words: Gaussian process regression, human mortality forecasting, spectral mixture kernel, weighted mean function, coherent forecasts, multilevel functional principal component analysis (MFPCA), time series, life expectancy, functional data clustering, dimension reduction, principal curves

Acknowledgements

It is my great pleasure to take this opportunity to acknowledge all the support I have received during my four years' study as a PhD student at University of Leicester.

First of all, I would like to express my sincere gratitude to my supervisor, Dr. Bo Wang, for his patience and persistent guidance on me throughout the whole period of my PhD study. Without his valuable suggestions and contributions, I won't be able to achieve these results.

Also I would like to thank Professor Alexander Gorban and Dr. Evgeny Mirkes for their beneficial advice on my second project. Many thanks to all my colleagues and friends at Department of Mathematics, including Juxi, Yanshan, Wenyan, Dominic, Sam, Ayo, Sarbaz, Ahmet and Masha, for the kind help they offered to me, both academically and psychologically.

On institutional level, I am grateful for the financial support jointly funded by the Institute and Faculty of Actuaries (IFoA) and the College of Science and Engineering of the University of Leicester (UoL) through a PhD studentship. I also appreciate the managerial and operational support provided by Mr Kevin McIver from IFoA, which ensured our research carried out smoothly.

Finally, I would like to give special thankfulness to my family for their love and faith on me, without which the completion of this thesis is impossible.

Contents

List of Figures

List of Tables

Abbreviations

FDA = Functional data analysis

GP = Gaussian processes

GPR = Gaussian process regression

MFPCA = Multilevel functional principal component analysis

FPCA = Functional principal component analysis

BIC = Bayesian Information Criterion

SE = Squared exponential

MA = Matérn

RQ = Rational quadratic

SM = Spectral mixture

WLS = Weighted least squares

LS = Least squares

MLE = Maximum likelihood estimation

RMSE = Root mean square error

RMSFE = Root mean square forecasting error

KL = Karhunen-Loève

BLUP = Best linear unbiased prediction

MCMC = Markov Chain Monte Carlo

ARIMA = Autoregressive integrated moving average

DF = Degree of freedom

MDS = Multidimensional scaling

HPCC = Hierarchical principal curve clustering

KNN = K_{th} nearest neighbour

NNVE = Nearest neighbour variance estimation

VAR = Vector autoregressive

VARMA = Vector autoregressive moving average

To my family

Chapter 1
Introduction

1.1 Backgrounds

The growing aged population, especially in developed countries, over the past decades, has given rise to significant changes in both social structures and economic conditions. Government as well as the insurance and pension industry need to adjust their existing policies according to the ever increasing aged population. They are obliged to pay billions of pensions and annuities, hence are heavily exposed to so-called “longevity risk”. Assessing and forecasting the demographic mortality trends is therefore of great interests to researchers due to its considerable impact on social welfare, resource allocation and governmental budgeting.

Apart from using biological, medical and behavioural knowledge, statisticians have developed very different, pure mathematical methods to model the mortality patterns. Since Lee and Carter (1992) launched their pioneering work in the modelling and forecasting of mortality for a single population (also called independent mortality forecasting), there has been a surge of interest along the lines of Lee-Carter method, leading to better forecasts of age-specific mortality. Parallel to forecasting mortality for a single population, recently researchers also became aware of the importance of forecasting mortality for multiple populations simultaneously (also called coherent mortality forecasting). Li and Lee (2005) laid down the foundation for the multi-population mortality forecasting as counterparts of single-population mortality forecasting. Their work was followed by several improvements and modifications in context. Apart from mortality forecasting, clustering age-specific mortality data is another important topic. Overall speaking, the development of various statistical methods for mortality forecasting and clustering not only resolves the need for government to plan and budget the allocation of social resources, but also assists life insurance companies and pensions to carry out their business.

1.2 Independent mortality forecasting

Independent mortality forecasting refers to forecasting the age-specific mortality rates for a single population. Lee and Carter (1992) first introduced their statistical model which was then named after them as Lee-Carter model. Lee-Carter model, based on the assumption that the future will be in some sense like the past, is a purely extrapolation method that constructs relatively simple but effective statistical tool to measure the confidence band of future mortality.

The Lee-Carter model, since its publication, has aroused great attention in literature. Several reported post studies, focusing on assessing the long-term prediction performance of Lee-Carter model, suggested some modifications to improve its performance. Lee and Miller (2001) modified the original Lee-Carter model by adjusting the coefficients series such that the fitted life expectancy and the observed life expectancy are equal. Research has also been done to prove that their modification does successfully reduce the bias of the forecasted results. Liu and Yu (2011) proposed the quantile regression method, instead of the original least square method in Lee-Carter model, for the robust estimation of the time-varying index since the quality of forecast relies much on it. Their study has shown that the suggested quantile regression method can improve the forecasting performance of the model, especially when data contain irregular shocks. Renshaw and Haberman (2003) introduced a parallel methodology based on generalized linear modelling, treating time as a known covariate. And there have been a few other modifications and extensions as well; see Bell (1997), Booth et al. (2002) for instance.

While Lee-Carter model and its modifications and extensions provide effective multivariate statistical tools to forecast future mortality, the functional data analysis (FDA), developed in recent decades, offers a new modelling framework on mortality. Ramsay and Silverman (2005) define functional data as set of smooth data curves and the corresponding analysis can then be conducted under certain continuum similarly as conventional multivariate analysis. Under FDA framework, the observed age-specific mortality in a particular year can be regarded as discrete points of a smooth function. Hyndman and Ullah (2007) developed their functional data model, combining nonparametric smoothing, FDA and time series techniques. In their model, the curves of the age-specific mortality are decomposed into a mean function

and the summation of a set of orthonormal basis functions which can be solved as the eigenfunctions of the covariance function of the observed data. This modelling framework of decomposing the random part of the process into several principal components is viewed as a generalization of the Lee-Carter model where only the first principal component is considered. Compared with Lee-Carter model, the advantage of having more principal components in long-term forecasting has been numerically verified by Hyndman and Ullah (2007).

Chiou and Müller (2009) further extended this FDA model based on the assumption that the principal components are evolving over time and introduced a Moving Window Approach to collect observed data curves with respect to the birth year of cohorts falling into that window. With the window moving forward, the mean function and principal components for each year are obtained consecutively, and the prediction of principal components in a future year could be made by functional local linear extrapolations. This method is regarded as of completely nonparametric feature since the coefficients of the principal components are not assumed to have any parametric form.

We develop a new model on a different basis. We consider the mortality of specific age groups over time to follow Gaussian processes (GP). We introduce a new GPR method which incorporates a weighted mean function and the spectral mixture covariance function. The spectral mixture covariance function enables that various covariance structures in mortality rates over time for different age groups can be captured, and the weighted mean function models the long term trend. The combination of these two provides better results in extrapolating mortality rates, compared with the conventional GPR methods. The performance of the proposed method is also compared with Lee-Miller model and the functional data model introduced by Hyndman and Ullah (2007) in the context of forecasting the French total mortality rates.

1.3 Coherent mortality forecasting

All of the work mentioned in the previous section concentrate on forecasting mortality for a single population. However, when dealing with a bundle of populations simultaneously, such individual forecasts may lose its effectiveness, since they tend to result in divergence of life expectancies in long run, although those populations may share extensive similarities in socioeconomic, environmental and biological conditions.

With the globalization carrying on the populations of the world are linked more closely than before. And all the increasing similarities among populations indicate that the mortality gaps between them should at least not widen over time. In fact, numerous studies have pointed to a global convergence in life expectancy in long run (United Nations 1998; White 2002). It then becomes more desirable to model the mortality of multiple populations simultaneously rather than to isolate one population from another. Researchers start to consider models for forecasting mortality of subpopulations within a large population, such as different sex, or different states in a country, and expect the forecasting results to be non-divergent. Contrast to the individual forecasts, such combined forecasts within one big group are named as coherent forecasts (see for example, Li and Lee 2005). Li and Lee (2005) extends the Lee-Carter model to a group of populations for coherent forecasting, by means of splitting the original model into a common factor and a specific factor. Oeppen (2008) applies the methods of compositional data analysis to transform the Lee-Carter model and then extends it to a multiple-decrement life table. And the test using Japanese cause of death data demonstrates a promising result. Hyndman et al (2013) develop an innovative method for coherent mortality forecasting, based on functional time series models. They define product and ratio from the smoothed functional data which are then decomposed individually using functional time series models. The implied coherent functional model for each subpopulation can hence be reverted simply by summing up the functional time series models for product and ratio. Their numerical results show that the forecasting accuracy is homogenized across subpopulations while the life expectancies in long run present convergence.

We propose a new framework for coherent mortality forecasting, namely weighted multilevel functional principal component analysis (MFPCA). We treat the mortality rates of subpopulations within a large population as a set of multilevel functional data (e.g. male and

female mortality data within a country), and then use MFPCA to extract core information from the functional data and analyse them at multilevel scale, so that the model incorporates both overall information from the population as a whole and specific information from the subpopulations. Comparing to the Product-Ratio model (Hyndman et al, 2013), the proposed model possesses several major differences in nature which will be discussed later.

1.4 Clustering mortality (fertility) as functional data

Clustering, as an unsupervised learning process, aims at partitioning a data set into sub-groups so that the instances within a group are similar to each other while they are dissimilar to instances of other groups. Given a set of functional data, it is an interesting and valuable task to identify homogeneity and heterogeneity among curves by using the clustering techniques, which helps to unveil the potential features and patterns of the underlying stochastic process. In the literature, there are several functional clustering methods that have been developed. Tarpey and Kinateder (2003) introduce the concept of principal points for functional Gaussian distributions and used k-means algorithm to estimate the cluster means. In Abraham et al (2003) and Rossi et al (2004), the curves are approximated by the popular B-spline and then k-means and Self-Organized Map are applied respectively to their coefficients for clustering. Peng and Müller (2008) use the k-means algorithm on the principal component scores obtained from functional principal component analysis (FPCA). Gaffney (2004) designs the Curve Clustering Toolbox for MATLAB, which implements a family of clustering algorithms based on Gaussian mixtures combined with spline or polynomial basis approximation. These methods decompose the functional data into different types of basis functions and then cluster on the corresponding FPCA scores (or basis expansion coefficients). They are referred to as hard clustering (or filtering methods). A parallel method to this is soft clustering (or adaptive methods), which is based on a probabilistic modelling of either basis expansion coefficients or FPCA scores. James and Sugar (2003) model on the spline basis expansion of the curves especially adapted for sparsely sample functional data and consider the coefficients to distribute according to a mixture of Gaussian distributions. In Bouveyron and Jacques (2011), the clustering algorithm (named as FunHDDC) is based on a functional latent mixture model, where the appropriate

number of clusters is determined by the Bayesian Information Criterion (BIC). Other examples of soft clustering are given by Fraley and Raftery (2002), Wakefield et al (2003), Booth et al (2008) and Shi and Wang (2008).

A principal curve is intuitively described as one-dimensional smooth curve that pass through the middle of a high dimensional data set. Hastie and Stuetzle (1989), who develop the fundamental framework for principal curves, define principal curves as the self-consistent curves whose point is the average over all points that project there. Since its introduction, there have been many practical usages of principal curves in various fields. Banfield and Raftery (1992) implement a principal-curve method to identify the outlines of ice floe in satellite images. Stanford and Raftery (2000) construct a principal curve clustering algorithm to detect curvilinear features in spatial point patterns. Gorban and Zinovyev (2010) analyse the nonlinear quality of life index for 171 countries by finding a principal curve that goes through a 4D space dataset and the projections of the data points on this curve are used for ranking. Motivated by the above applications, we introduce the functional principal curve clustering method, which, to the best of our knowledge, is the first work to apply principal curve framework to the context of functional data clustering. The proposed method makes use of the nonparametric principal curves to summarize the features of the two-dimensional scores extracted from the original functional data for clustering purpose. A probability model with BIC for principal curves is introduced to automatically and simultaneously find the appropriate number of features and the optimal degree of smoothing. And this principal curve clustering method is then applied to the clustering of French total mortality as well as the Australian fertility.

1.5 Thesis outline

The rest of the thesis is organized as follows.

In chapter 2, we briefly present some fundamental ideas about Gaussian process regression, discuss how this method can be applied in forecasting mortality rates, and introduce the new GPR model with weighted mean function and spectral mixture kernel. The GPR models are then applied to the French total mortality data and their performances are compared. We also

compare the forecasting performance of the proposed model with the Lee-Miller model and the functional data model.

Chapter 3 starts with some backgrounds of functional principal component analysis (FPCA). Then the model and the framework of MFPCA are introduced in details. The MFPCA model is then applied to the sex-specific mortality data of nine developed countries for coherent forecasting, and the performance is compared with the independent model (Hyndman and Ullah, 2007) as well as the Product-Ratio model. We further demonstrate the performance of the model in terms of long-run forecasting convergence in the context of sex-specific mortality data of UK.

In chapter 4, the concept of principal curves is revisited first. After that, the algorithm for clustering functional data under principal curve context is developed in details. The effectiveness of this functional clustering method is verified through simulation study. In the end, this clustering method is applied to the French total mortality and Australian fertility for clustering analysis, in order to discover some potential demographic features of these two countries in the past century.

Some important discussions and future work will be provided in chapter 5.

Chapter 2
Gaussian process regression method for forecasting of mortality rates

2.1 Introduction

Gaussian process regression (GPR) has long been proved to be a powerful and effective Bayesian nonparametric approach for smoothing and interpolation, but less successful for extrapolation. In this chapter, we propose a new model by considering the mortality of specific age groups over time to follow Gaussian processes (GP). A new Gaussian process regression (GPR) method which incorporates a weighted mean function and the spectral mixture covariance function is developed to model the age-specific human mortality rates of selected age groups. After forecasts are made for these selected age groups, the age-specific mortality curve for a particular future year can be obtained by interpolating the forecasted mortality rates to all age groups. The rest of this chapter is organised as follows. In Section 2.2, we briefly present some fundamental ideas about Gaussian process regression, discuss how this method can be applied in forecasting mortality rates, and introduce the new GPR model with weighted mean function and spectral mixture kernel. In Section 2.3, the GPR models are applied to the French total mortality data and their performances are compared. We also compare the forecasting performance of the proposed model with the Lee-Miller model and the functional data model. Conclusion is given in Section 2.4.

2.2 Methodology

2.2.1 Fundamentals of Gaussian process regression

Gaussian process regression has been developed as an effective statistical method for non-linear regression in past decades. Rather than assuming some specific models for an unknown

function $f(x)$, the Gaussian process is a less parametric tool and allows data to release more information for themselves. As Rasmussen & Williams (2006) pointed out, a Gaussian process is a generalization of the Gaussian probability distribution. By definition, a Gaussian process (GP) is a stochastic process that any finite subset throughout its domain follows a multivariate Gaussian (normal) distribution (Shi & Choi, 2010). Consider a nonlinear regression model with noise:

$$y = f(x) + \varepsilon,\ \ \varepsilon \sim N(0, \sigma^2)\ .$$

By Gaussian process method, the unknown function $f(x)$ is treated as a random function and is assumed to have a Gaussian process prior with a mean function $\mu(\cdot)$ and a covariance function $k(\cdot,\cdot)$. The covariance function relates one point to another and is defined as:

$$k(x, x'; \theta) = Cov\big(f(x), f(x')\big),$$

where θ denotes the set of hyper-parameters which need to be estimated. The mean function is often assumed to be zero, but the covariance function is vital to GP since it determines properties of the unknown function $f(x)$ such as smoothness and periodicity. A commonly used covariance function is the squared exponential (SE) kernel which has the following form:

$$k_{SE}(x, x') = \sigma_f^2\ exp[-(x - x')^2/2l^2].$$

In the above kernel function, the hyper-parameter $\theta = \{\sigma_f, l\}$, which can be estimated by empirical Bayesian approach, given the observed data. It can be easily observed that $k_{SE}(x, x')$ approaches its maximum if $x \approx x'$, while $k_{SE}(x, x') \approx 0$ if x is distant from x'.

Given the observed data $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, the joint distribution of $y_1, y_2, \dots, y_n$ is multivariate normal:

$$\boldsymbol{y} = (y_1, y_2, \dots, y_n)^T \sim N_n(\boldsymbol{\mu}, \Psi),$$

where the mean $\boldsymbol{\mu}$ has entries $\mu_i = \mu(x_i)$ and Ψ is an $n \times n$ matrix, whose $(i, j)th$ element is defined as

$$\Psi_{ij} = Cov\big(y_i, y_j\big) = k\big(x_i, x_j; \theta\big) + \sigma^2 \delta_{ij},$$

where δ_{ij} is the Kronecker delta. Therefore the empirical Bayes estimates of θ and σ, denoted as $\hat{\theta}$ and $\hat{\sigma}$, can be obtained by maximizing the marginal log-likelihood

$$l(\theta, \sigma|\mathcal{D}) = -\frac{1}{2}\log(|\Psi|) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^T\Psi^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) - \frac{n}{2}\log(2\pi).$$

For a new input x^*, given the training data $\mathcal{D}$ the predictive distribution of $f(x^*)$ is normal following the multivariate normal distribution properties. Its mean and variance are given as:

$$E(f(x^*)|\mathcal{D}) = \mu(x^*) + \psi^T(x^*)\Psi^{-1}(\boldsymbol{y} - \boldsymbol{\mu}),$$

$$Var(f(x^*)|\mathcal{D}) = k\left(x^*, x^*; \hat{\theta}\right) - \psi^T(x^*)\Psi^{-1}\psi(x^*),$$

where $\psi(x^*) = (k\left(x^*, x_1; \hat{\theta}\right), \dots, k\left(x^*, x_n; \hat{\theta}\right))^T$ is the covariance between $f(x^*)$ and $\boldsymbol{f} = (f(x_1), \dots, f(x_n))$, and Ψ is the covariance matrix of $(y_1, y_2, \dots, y_n)^T$ given in (6). Then the predictive distribution of y^* is also normal, with mean $\hat{y}^*$ equal to the mean of $f(x^*)$ and the variance

$$\hat{\sigma}^{*2} = Var(f(x^*)|\mathcal{D}) + \hat{\sigma}^2.$$

2.2.2 Gaussian process regression models for mortality forecasting

Let $y_x(t)$ denote the log mortality rates for a specific age *x* in year *t*. We assume that there is an underlying function $f_x(t)$ that we are observing with error at discrete points of *t*. Our observations are $\{t_i, y_x(t_i)\},\ x = 1, \dots, n,\ i = 1, \dots, m$ and satisfy

$$y_x(t_i) = f_x(t_i) + \varepsilon_{i,x},$$

where $\varepsilon_{i,x}$ is a sequence of i.i.d normal random variables $N(0, \sigma^2)$. Based on the observations we are interested in forecasting $y_x(t)$ for any x and $t \in [t_{m+1}, t_{m+h}]$.

2.2.2.1 Basic GPR models

The application of GPR models in mortality forecasting is straightforward. For any age x we can build a GPR model for the unknown function $f_x(t)$ as discussed in the previous section, and the forecast of the mortality rates at a future year t^*, $f_x(t^*)$, can then be obtained. As the log mortality rates of specific age groups over time display an overall decreasing trend, a linear function can be used as the mean function of GPR, and three commonly used stationary covariance functions, namely squared exponential (SE), Matern (MA) (with degree of freedom equal to 3/2) and rational quadratic (RQ), will be evaluated in this chapter. The three covariance functions have the following forms:

$$k_{SE}(\tau) = \sigma_{SE}^2(-\tau^2/2l_{SE}^2),$$

$$k_{MA}(\tau) = \sigma_{MA}^2\left(1+\sqrt{3}\tau/l_{MA}\right) exp\left(-\sqrt{3}\tau/l_{MA}\right),$$

$k_{RQ}(\tau) = \sigma_{RQ}^2(1+\tau^2/2\alpha l_{RQ}^2)^{-\alpha}$, and α is non-negative parameter,

where $\tau = t - t'$.

2.2.2.2 GPR model with weighted mean function and SM kernel

In this section we propose a new Gaussian process regression method for the modelling and forecasting of age-specific human mortality rates. The model incorporates a weighted mean function and the spectral mixture covariance function and makes use of the strengths of both of them. As a result the ability of extrapolation by GPR is improved and the numerical example shows that our method provides a stable performance for short term and long term mortality forecasts.

Gaussian process has been proven to be a very effective nonparametric approach for smoothing and interpolation. However, its ability of pattern discovery and extrapolation is still to be developed. Wilson & Adams (2013) introduced spectral mixture (SM) kernels, which is derived by modelling a spectral density − the Fourier transform of a kernel − with a

Gaussian mixture. These kernels can support a broad class of stationary covariance functions and provide a better solution for extrapolation.

According to Bochner's theorem (Bochner, 1959; Stein, 1999), any stationary kernel can be expressed as an integral.

Theorem 2.2.2.2.1 *(Bochner) A complex-valued function k on R^P is the covariance function of a weakly stationary mean square continuous complex-valued random process on R^P if and only if it can be represented as*

$$k(\tau) = \int_{R^P} e^{2\pi i s^T \tau} \varphi(ds),$$

where φ is a positive finite measure.

If φ has a density $S(s)$, then S is named as the spectral density of k, and k and S are Fourier duals:

$$k(\tau) = \int_{R^P} S(s) e^{2\pi i s^T \tau} ds, \tag{2.1}$$

$$S(s) = \int_{R^P} k(\tau) e^{-2\pi i s^T \tau} d\tau.$$

Given these backgrounds, Wilson & Adams (2013) point out that any stationary covariance kernels can actually be approximated to arbitrary precision by using a mixture of Gaussians in the spectral density. Consider a simple case, where

$$\phi(s; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(s-\mu)^2\right\}, \text{ and}$$

$$S(s) = [\,\phi(s) + \phi(-s)]/2.$$

Substituting $S(s)$ into (2.1), we have

$$k(\tau) = exp\{-2\pi^2\tau^2\sigma^2\} cos(2\pi\tau\mu).$$

Considering a mixture of Q Gaussians on R^P, where the q^{th} component has mean vector $\mu_q = (\mu_q^{(1)}, \ldots, \mu_q^{(P)})$ and covariance matrix $M_q = diag(v_q^{(1)}, \ldots, v_q^{(P)})$, and τ_p is the p^{th} component of the P dimensional vector $\tau = t - t'$, then spectral mixture kernel is expressed as

$$k_{SM}(\tau) = \sum_{q=1}^{Q} \omega_q \prod_{p=1}^{P} exp\{-2\pi^2 \tau_p^2 \upsilon_q^{(p)}\} cos(2\pi\tau_p \mu_q^{(p)}),$$

where the weight ω_q specifies the contribution of each mixture component and μ_q and υ_q are hyper-parameters to be optimized. In our case of modelling the log mortality rate $y_x(t)$, $P = 1$.

In the GPR models, the prior mean function tends to have a significant impact on the extrapolative mean since the extrapolation is inclined to move to the prior mean in the long run. Previously, the mean function is modelled by using linear regression on the training data, which means each data point in the past carries equal weight on the mean function. However, in mortality modelling, it is often the case that more recent data tend to have more impact on the results than those in the distant past: the more recent the data point is, the greater influence it tends to have on the future mortality rates. Therefore, we propose to model the prior mean function by assigning different weights to the training data points. The mean function is still modelled by linear regression, but using weighted least squares (WLS) method instead of the original least squares (LS) method. The parameters of the linear mean function is chosen to minimize the error *e*,

$$e = \sum_{i=1}^{m} w_i (y_i - \hat{y}_i)^2,$$

where w_i is the weight of the i^{th} data point, y_i is the observed i^{th} data point and $\hat{y}_i$ is the estimated i^{th} point by linear regression. Here we assume the weights to be equal to the inverse of the time distance to the first year to be forecasted, namely t_0 (in the numerical example later on, t_0 =1991):

$$w_i = 1/(t_0 - t_i),$$

where t_i denotes the year of the i^{th} data point.

The parameter estimation and the prediction for the above model can be performed in the same way as the basic GPR model. It is noted that other weights can be used for the mean function, and if the weights involve tuning parameters, they can be determined by cross validation.

2.3 Applications of GPR method in forecasting French mortality rates

In this section we apply the basic and the proposed GPR models to French total mortality rates and then compare their performance with two of the existing models in the literature, namely the Lee-Miller model (Lee and Miller, 2001) and the functional data model (Hyndman and Ullah, 2007). The data are obtained from the Human Mortality Database (2010), which consist of the observed mortality rates for every one year at each age. We select the years from 1950 to 2010 to avoid the anomalous mortality rates during the first and second world wars. Figure 2.1 shows the log French total mortality for 20, 30, 40 and 50-year groups as univariate time series, observed from 1950 to 2010.

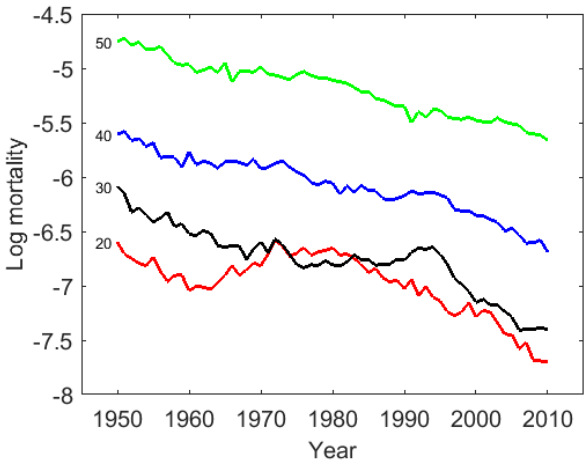


Figure 2.1: Log French total mortality for 20, 30, 40 and 50-year groups observed from 1950 to 2010.

2.3.1 Forecasting comparison based on selected age groups

For testing purpose, we select four age groups 20, 30, 40 and 50 to carry out analysis and compare the forecasting performances of the basic GPR and the GPR with weighted mean and SM kernel. The mortality rates for these four age groups are among the most difficult to model due to the significant variation during the period of study. We split the data of each age group into two parts: the data from 1950 to 1990 as training data and those from 1991 to

2010 as testing data. The basic GPR models with the SE, MA, RQ kernels as well as the GPR with weighted mean and SM kernel are fitted to the training data of each age group separately. The hyper-parameters of each kernel are estimated using maximum likelihood estimation (MLE) where 100 initial values are randomly selected from a possible range and the ones that give the highest marginal likelihood are used as the estimates. Then the mortality forecasts are made for a 20-year horizon and compared with the actual values. As demonstration Figures 2.2 illustrates the forecasting results for the 40-year age group.

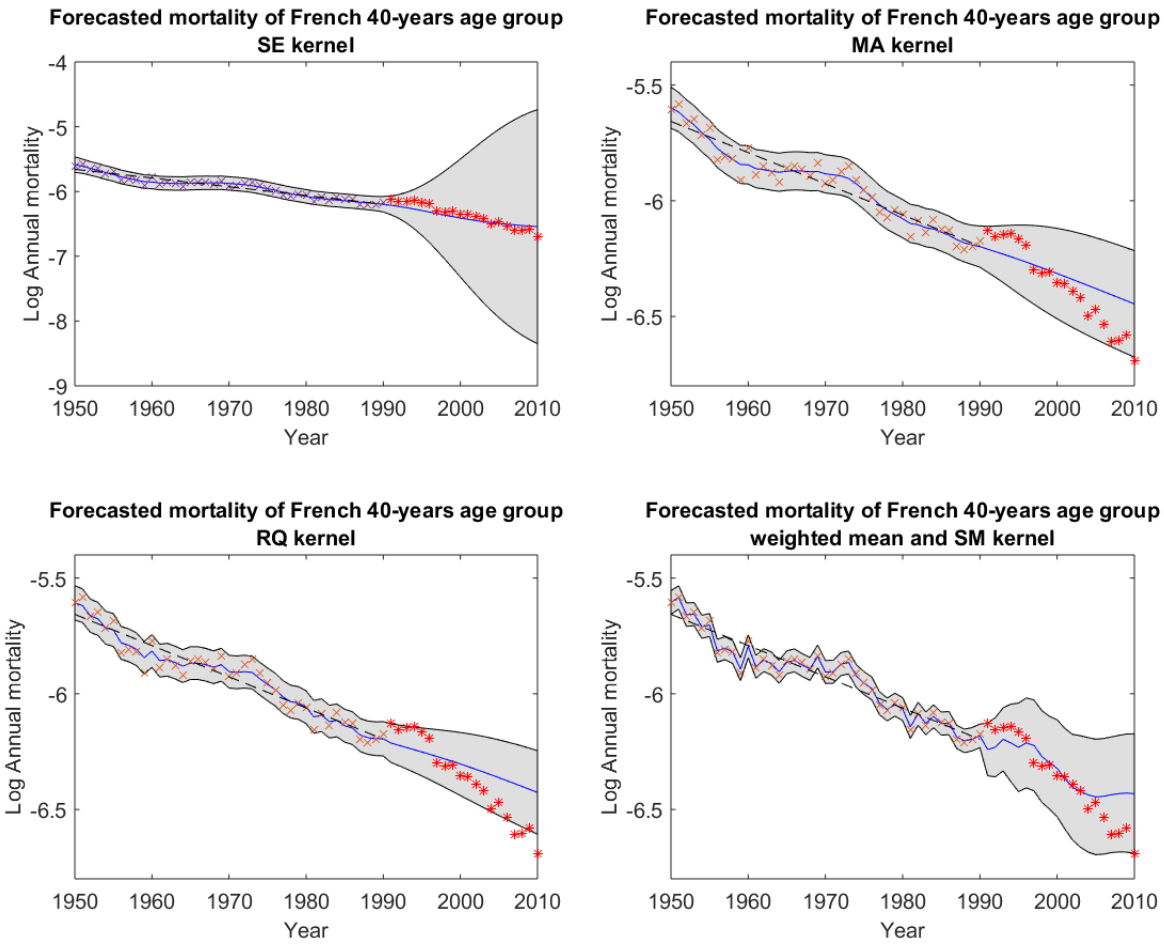


Figure 2.2**:** Forecasting mortality of French 40-year age group using the GPR models with SE, MA & RQ kernels and the GPR with weighted mean and SM kernel. The training data are displayed in blue X-mark while the testing data are displayed in red stars. The blue solid line is the predictive mean, with 95% confidence interval indicated by grey shade. The black dashed line represents the mean function of the GP models.

The root mean squared error (RMSE) between the forecasted values and the actual values for each age group are recorded in Table 2.1.

Age	Basic GPR			WM-SM GPR
	SE	MA	RQ	
20	0.3516	0.4245	0.4808	0.3878
30	0.4752	0.2810	0.2445	0.3217
40	0.1017	0.1111	0.1201	0.1003
50	0.3026	0.0499	0.0540	0.0386

Table 2.1**:** RMSEs of French log mortality for 20, 30, 40 and 50 years group using the GPR with SE, MA and RQ kernels and the GPR with weighted mean function and SM kernel (WM-SM GPR).

It can be observed that, for the basic GPR models, although different kernels perform similarly (in terms of RMSE) in the forecasting for some particular age groups (e.g. 40-years group), the choices of kernels still have a significant impact on forecasting for many other age groups. Taking SE kernel for example, it generates comparatively good result for the 20-years group, but it performs poorly for the 50-years group. On the other hand, although the GPR with weighted mean function and SM kernel (WM-SM GPR) may not provide the best prediction for some age groups (e.g. 20 and 30 year age groups), it does significantly improve the overall forecasting results. The testing results of the above four age groups indicates that, for the basic GPR models choosing an appropriate kernel for a specific age group is of great importance to the accuracy of forecasting, and the proposed model mitigates this difficulty and provides much better overall performance in terms of forecasting accuracy.

2.3.2 Comparison of forecasted mortality curves

We now demonstrate the usefulness of the weighted mean function by comparing the accuracy of forecasted mortality curves in future years using two models: the GPR with SM kernel and weighted mean function and the GPR with SM kernel and unweighted mean function. To construct the mortality curves for a future year, we select 20 specific age groups,

namely, the 0, 1, 2, 5, 10, 12, 15, 18, 20, 22, 25, 28, 30, 40, 50, 60, 70, 80, 90, 100-year groups, fit the GPR models to each of these age groups, and then obtain the mortality curves by interpolating the forecasted mortality rates to all ages. The rationale for age selection is that we want to have dense points in the areas with large variation and sparse in those with small variation. Our experiment shows that there is no significant difference in the results if more or slightly different age groups are used. The figures showing the forecasting results of the above 20 age groups using GPR model (SM kernel) with and without weighted mean are all displayed in appendix. We also compute the forecasting errors of the two GPR models in terms of RMSE, for these 20 selected age groups respectively, and the results are displayed by a table in appendix.

We consider to forecast the mortality curves for 1995 (5-year forecast horizon), 2000 (10-year forecast horizon), 2005 (15-year forecast horizon) and 2010 (20-year forecast horizon), based on the data from 1950 to 1990 as the observation. The results are shown in Figure 2.3-2.6.

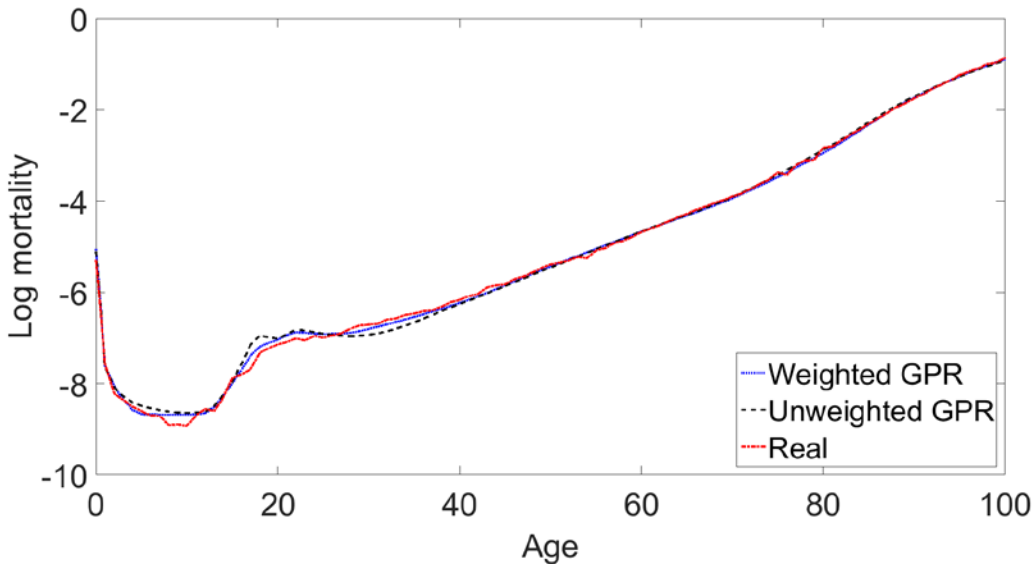


Figure 2.3: Forecasted and real log French total mortality curves for 1995. The black dashed curves represent the forecasted log mortality curves using the SM GPR model with unweighted mean function, while the blue smooth curves represent those using the SM GPR model with weighted mean function. The red curves are the real mortality curves.

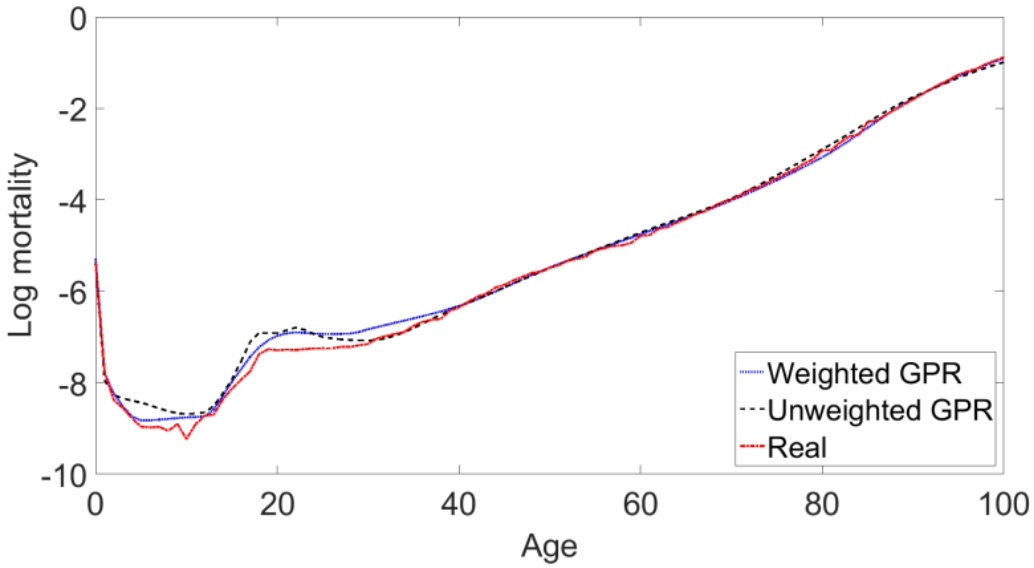


Figure 2.4: Forecasted and real log French total mortality curves for 2000. The black dashed curves represent the forecasted log mortality curves using the SM GPR model with unweighted mean function, while the blue smooth curves represent those using the SM GPR model with weighted mean function. The red curves are the real mortality curves.

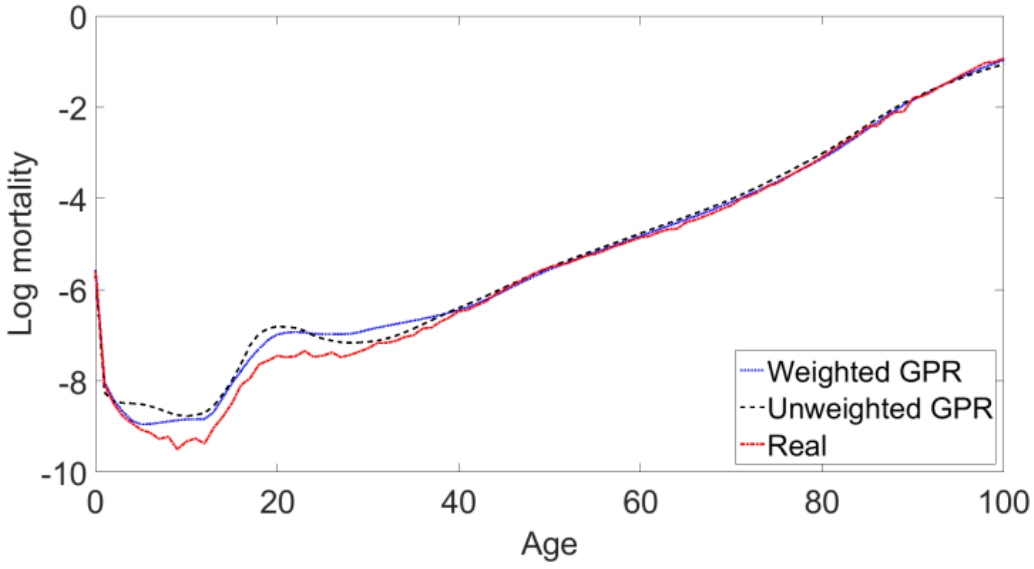


Figure 2.5: Forecasted and real log French total mortality curves for 2005. The black dashed curves represent the forecasted log mortality curves using the SM GPR model with unweighted mean function, while the blue smooth curves represent those using the SM GPR model with weighted mean function. The red curves are the real mortality curves.

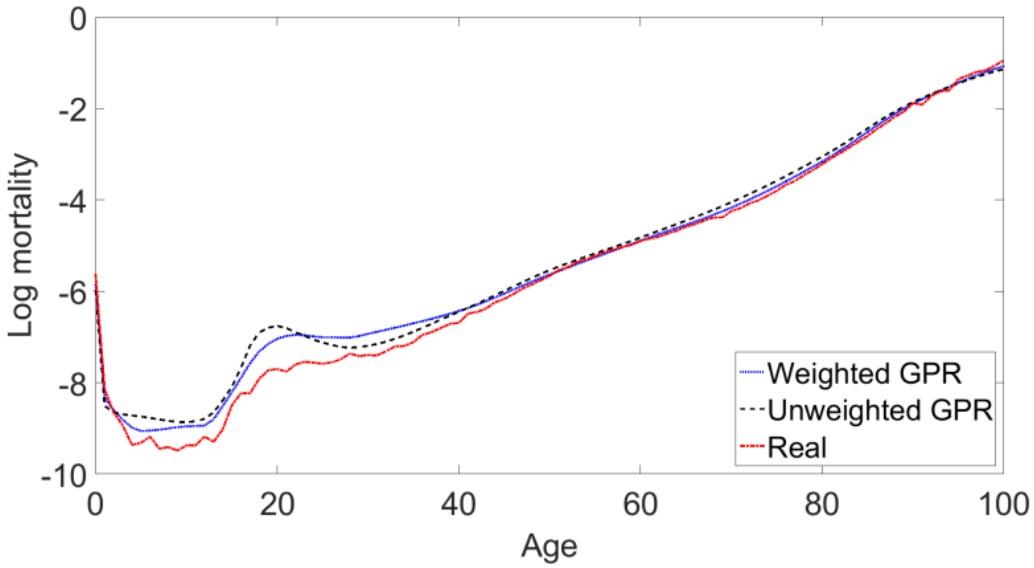


Figure 2.6: Forecasted and real log French total mortality curves for 2010. The black dashed curves represent the forecasted log mortality curves using the SM GPR model with unweighted mean function, while the blue smooth curves represent those using the SM GPR model with weighted mean function. The red curves are the real mortality curves.

It can be observed that, for short-term forecast (with 5-year forecast horizon), the GPR models with and without weighted mean function produces quite similar results, which are both close to the real mortality curve. For long-term forecasts (with 10, 15, 20-year forecast horizons), whilst both of the GPR models give very good forecasts for over 40s, they tend to overestimate the mortality rates of the young age groups (approximately from 2 to 30-year groups). However, the SM GPR model with weighted mean function still provides better results: its forecasted curves are closer to the real ones; it can also be reflected from the RMSEs between the forecasted values and the actual values displayed in Table 2.2. It is obvious that incorporating the weighted mean function into the GPR model gives rise to a significant improvement in forecasting accuracy.

	1995	2000	2005	2010
unweighted	0.1284	0.1936	0.2897	0.3502
weighted	0.0844	0.1517	0.2397	0.2884

Table 2.2: RMSEs of the forecasted log French total mortality curves for 1995, 2000, 2005 and 2010 by SM GPR with unweighted mean function and weighted mean function.

Now we compare the forecast accuracy of the proposed model with some existing models in the literature. We choose the methods of Lee and Miller (2001) and Hyndman and Ullah (2007) as benchmarks for comparison, both of which can be implemented using R package '*demography*'. Our GPR model, Hyndman-Ullah model (also called functional data model, FDM) and Lee-Miller model (LM) are applied to the French total mortality data for years from 1950 to Z (where $Z = 1981, \dots, 1990$) and forecasts are made for up to 20-year horizon, i.e. to forecast the mortality rates for $Z+1, \dots, \min(Z+20, 2010)$. The forecasts are compared with the actual mortality rates (on log scale) and the RMSEs over 20-year horizon for $Z = 1981, \dots, 1990$ are calculated. The average of the 10 RMSEs is calculated and then defined as root mean square forecasting error (RMSFE). The obtained RMSFEs for these three methods are presented in Figure 2.7.

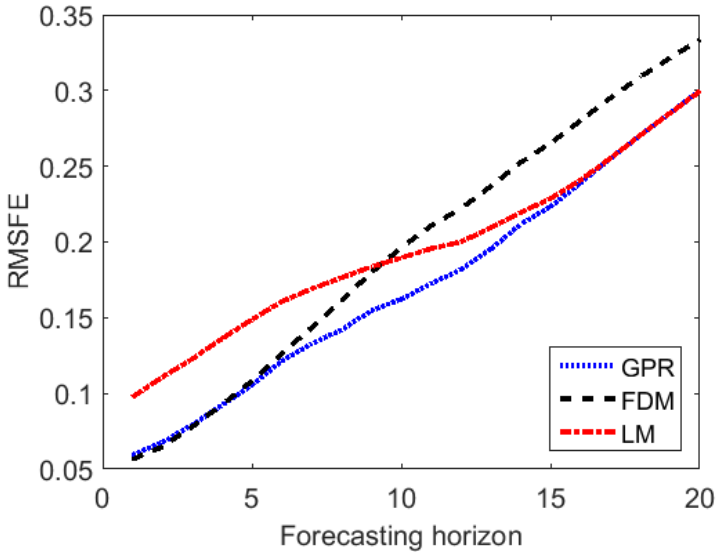


Figure 2.7: Forecasting accuracy of three methods in terms of out-of-sample RMSE (RMSFE).

It can be seen from the above figure that, the performances of our model and the functional data model are indistinguishable in short term (from 1 to 6-year forecasting horizon). But from 7-year horizon onwards, the proposed model substantially outperforms the functional data model. In contrast, the performance of our model is almost equal to that of the Lee-Miller in long term (for 15 to 20-year horizons). However, it has much better accuracy for 1 to 15-year horizons when compared with the Lee-Miller model. In the mid-term, the proposed GPR model is significantly better than both of the other two. Overall, our method provides very stable performance in terms of forecasting error in both short term and long term with much improved forecasting accuracy in mid-term, when compared with the functional data model and the Lee-Miller model, based on the French total mortality data.

2.4 Conclusion

We have introduced Gaussian process regression as a new approach for modelling and forecasting mortality rates. As a Bayesian nonparametric method, Gaussian process models have shown the effectiveness for smoothing and interpolation. We first considered several basic types of GPR models to test their abilities for extrapolation in the context of mortality rates and the numerical examples showed that choosing an appropriate type of the GPR kernel can be vital to the result of extrapolation. To solve the difficulty of manually choosing an appropriate kernel function, we proposed to use the spectral mixture kernel with a weighted mean function in the GPR model in order to capture the future trend more accurately as well as to automatically discover potential patterns from the training data. The numerical examples showed the proposed model improved the forecasting accuracy of mortality rates in long term, compared with the basic GPR methods. The performance of the proposed method was also compared with Lee-Miller model and the functional data model by Hyndman and Ullah (2007). The results demonstrated that our method provides a more stable performance in terms of forecasting errors.

In contrast to the conventional functional data model and Lee-Miller model, which directly act on the historical mortality curves for forecasting, our method provides a different angle to handle this forecasting issue. We focus on modelling the mortality rates of specific age

groups over time and assume that they follow Gaussian processes. After forecasts are made for some age groups, the mortality rates at other ages for a particular future year can be obtained by interpolating the forecasted mortality rates to all age groups. Therefore the forecasting accuracy depends on the choice of the age groups to be modelled. In our example, 20 specific age groups were selected, including 0, 1, 2, 5, 10, 12, 15, 18, 20, 22, 25, 28, 30, 40, 50, 60, 70, 80, 90, 100-years groups. The reason for choosing these age groups is that, the patterns of an age-specific mortality curve tend to be very fluctuated from 0 to 30 years while it increases almost linearly from 30 to 100 years. Hence we need more dense grids for interpolation in the interval from 0 to 30 years and fewer points from 30 years onwards. Our experiment showed that there was no significant difference in the results if more or slightly different age groups were used.

Chapter 3
Coherent mortality forecasting: the weighted multilevel functional principal component approach

3.1 Introduction to coherent mortality modelling

In human mortality modelling, if a population consists of several subpopulations it is always desired to model their mortality rates simultaneously while taking into account the heterogeneity among them. The traditional mortality forecasting methods tend to result in divergent forecasts for subpopulations when independence is assumed. However, under closely related social, economic and biological backgrounds, mortality patterns of these subpopulations are expected to be non-divergent in long run. In this chapter, we propose a new framework for coherent modelling and forecasting of mortality rates for multiple subpopulations within one large population. We treat the mortality of subpopulations as multilevel functional data and then a weighted multilevel functional principal component approach is proposed and used for modelling and forecasting the mortality rates. The proposed model is applied to sex-specific data for nine developed countries, and the forecasting results suggest that, in terms of overall accuracy, the model outperforms the independent model (Hyndman and Ullah 2007) and is comparable to the Product-Ratio model (Hyndman et al 2013) but with several advantages.

3.2 Coherent mortality model based on multilevel functional principal component analysis

3.2.1 A review of the functional principal component analysis (FPCA)

In the past, statisticians focused on solving problems on multivariate statistics when the dimension of data is large. And the extension of data dimension from finite space to infinite space generates the idea of functional data naturally. The recent development in computing and data collection has enabled such extension from the traditional multivariate data to functional data. Functional data is represented by a set of curves belonging to an infinite dimensional space (Ferraty and Vieu, 2006). Under the functional data context, a curve is a random function, which is also considered as a sample from a stochastic process. Functional principal component analysis (FPCA) refers to the statistical methodology of analysing functional data (Shang, 2014). Functional principal component analysis can reveal more statistical information contained in the smoothness and derivatives of the functions, which distinguishes it from the multivariate principal component analysis (Ramsay and Silverman, 2005).

As a natural extension of the multivariate PCA, the core of FPCA technique is based on the Karhunen-Loève (KL) expansion of a stochastic process (Karhunen 1946; Loève 1946). Later on, Rao (1958) and Tucker (1958) applied the KL expansion to functional data. In Dauxois et al (1982), some important asymptotic properties of FPCA elements for the infinite-dimensional data were derived. Since then, there had been many theoretical developments of FPCA mainly in two streams: the linear operator point of view and the kernel operator point of view. The former includes work done by Besse (1992), Mas (2002, 2008) and Bosq (2000). And the latter relates to more recent work by Yao et al (2005), Hall et al (2006) and Shen (2009). Besides, some extensions and modifications of FPCA, including smoothed FPCA, sparse FPCA, multilevel FPCA are also proposed by researchers. The ultimate goal of FPCA is to reduce the infinite dimension of functional data into finite dimensions in principal directions of variation.

Let Y be an L_2-continuous stochastic process defined on some set $\mathcal{T}$ (time interval for instance) and Y is with finite variance $\int_{\mathcal{T}} E(Y^2) < \infty$. Let $\mu(x) = E[Y(x)]_{x\in\mathcal{T}}$ denote the mean function of Y. If $L_2(\mathcal{T})$ is the space of the square-integrable functions defined on $\mathcal{T}$, the covariance function of Y, denoted as K, is defined as:

$$K(x,x') = Cov\big(Y(x),Y(x')\big) = E\big[\big(Y(x)-\mu(x)\big)\big(Y(x')-\mu(x')\big)\big],\ \ x,x' \in \mathcal{T}.$$

And the covariance operator $\mathcal{K}$of Y is induced by:

$$\mathcal{K}: L_2(\mathcal{T}) \to L_2(\mathcal{T}),$$
$$f \to \mathcal{K}f = \int_{\mathcal{T}} K(\cdot,x)f(x)dx.$$

Having covariance operator defined, we are able to carry out the spectral analysis of K:

$$\mathcal{K}\phi = \lambda\phi,$$

to obtain a set of nonnegative eigenvalues $\{\lambda_i\}_{i\geq 1}$ associated with a set of orthonormal eigenfunctions $\{\phi_i\}_{i\geq 1}$, where $\lambda_1 > \lambda_2 > \cdots \geq 0$ and $\int_{\mathcal{T}} \phi_i(x)\,\phi_{i'}(x)dx = 1$ if $i = i'$ and 0 otherwise.

By KL expansion, a stochastic process Y can then be expressed as:

$$Y(x) = \mu(x) + \sum_{k=1}^{\infty} \beta_k \phi_k(x),\ \ x \in \mathcal{T},$$

where $\beta_k = \int_{\mathcal{T}} (Y(x)-\mu(x))\phi_k(x)\,dx$ are uncorrelated random variables with mean zero and variance λ_k. These random variables are called principal component scores. The principal component scores can be intuitively interpreted as the projection of the centred stochastic process $Y(x)-\mu(x)$ in the direction of the k-th eigenfunction ϕ_k. In practice, only the first several eigenfunctions are needed to represent the important features of the set of random

functions. Therefore we usually truncate the expansion at the first N terms to obtain an approximation of $Y(x)$ in L_2 norm:

$$Y(x) = \mu(x) + \sum_{k=1}^{N} \beta_k \phi_k(x), \quad x \in \mathcal{T}.$$

3.2.2 Multilevel FPCA

In practice it is sometimes the case that a set of functional data comprise a number of subsets with strong correlations so that the functional data have a multilevel structure, such as the mortality rates for male and female in a country. The standard FPCA is not suitable for this type of functional data since it ignores the heterogeneity among subgroups. To address the challenges, Di et al (2009) proposes a multilevel FPCA (MFPCA), which combines FPCA and multilevel mixed models. Let $Y_{ij}(x)$ denote a random function for subgroup j within subject i, $i = 1,2, \dots, I$ and $j = 1,2, \dots, J$. Consider a two-way functional ANOVA model

$$Y_{i,j}(x) = \mu(x) + \eta_j(x) + Z_i(x) + W_{i,j}(x),$$

where $\mu(x)$ is the overall mean function, $\eta_j(x)$ is the group-specific mean shift from the overall mean, $Z_i(x)$ is the subject-specific deviation from the group-specific mean, and $W_{i,j}(x)$ is the residual subject and group specific deviation from the subject-specific mean. In such model, $\mu(x)$ and $\eta_j(x)$ are fixed functions while $Z_i(x)$ and $W_{i,j}(x)$ are zero-mean stochastic processes. $Z_i(x)$, termed as the level-one functions, and $W_{i,j}(x)$, the level-two functions, are assumed to be uncorrelated. This model incorporates multiple nested levels of random effect functions, which distinguishes it from the other functional principal component models in the literature.

Both the level-one and the level-two functions can then be decomposed using the Karhunen-Loève (KL) expansion as follows:

$$Z_i(x) = \sum_k \beta_{i,k} \phi_k^{(1)}(x),$$
$$W_{i,j}(x) = \sum_l \gamma_{i,j,l} \phi_l^{(2)}(x),$$

where $\phi_k^{(1)}(x)$ and $\phi_l^{(2)}(x)$ are level-one and level-two eigenfunctions respectively, and $\beta_{i,k}$ and $\gamma_{i,j,l}$ are their corresponding principal component scores. With these expansions, the model can then be expressed as

$$Y_{i,j}(x) = \mu(x) + \eta_j(x) + \sum_k \beta_{i,k}\,\phi_k^{(1)}(x) + \sum_l \gamma_{i,j,l}\,\phi_l^{(2)}(x). \tag{3.1}$$

A set of assumptions are made for the model:

(1) $E(\beta_{i,k}) = 0, var(\beta_{i,k}) = \lambda_k^{(1)}, E(\beta_{i,k_1}\beta_{i,k_2}) = 0$ for any i, $k_1 \neq k_2$;
(2) $\{\phi_k^{(1)}(x): k = 1,2, \dots\}$ is an orthonormal basis of $L^2[0,1]$;
(3) $E(\gamma_{i,j,l}) = 0, var(\gamma_{i,j,l}) = \lambda_l^{(2)}, E(\gamma_{i,j,l_1}\gamma_{i,j,l_2}) = 0$ for any i, j, $l_1 \neq l_2$;
(4) $\{\phi_l^{(2)}(x): l = 1,2, \dots\}$ is an orthonormal basis of $L^2[0,1]$;
(5) $\{\beta_{i,k}: k = 1,2, \dots\}$ are uncorrelated with $\{\gamma_{i,j,l}: l = 1,2, \dots\}$.

The first four assumptions are standard for FPCA while the last one is related to the previously stated assumption that $Z_i(x)$ and $W_{i,j}(x)$ are uncorrelated. Note that the level-one and level-two eigenfunctions are assumed to be orthonormal, but they are not necessarily mutually orthogonal.

To obtain the eigenfunctions (or the principal components) in the model, we need to estimate the covariance functions first. Let $K_T(x_s, x_r) = cov\{Y_{i,j}(x_s), Y_{i,j}(x_r)\}$ be the overall covariance function, $K_B(x_s, x_r) = cov\{Y_{i,j}(x_s), Y_{i,k}(x_r)\}$ the covariance functions between level two units within the same level one unit. And define $K_W(x_s, x_r) \coloneqq K_T(x_s, x_r) - K_B(x_s, x_r)$. It follows from (3.1) that

$$K_T(x_s, x_r) = \sum_{k=1}^{\infty} \lambda_k^{(1)} \phi_k^{(1)}(x_s)\phi_k^{(1)}(x_r) + \sum_{l=1}^{\infty} \lambda_l^{(2)} \phi_l^{(2)}(x_s)\phi_l^{(2)}(x_r),$$

$$K_B(x_s, x_r) = \sum_{k=1}^{\infty} \lambda_k^{(1)} \phi_k^{(1)}(x_s)\phi_k^{(1)}(x_r),$$

$$K_W(x_s, x_r) = \sum_{l=1}^{\infty} \lambda_l^{(2)} \phi_l^{(2)}(x_s)\phi_l^{(2)}(x_r).$$

In practice, each function $Y_{i,j}(x)$ is observed at a set of grid points and we assume that a common grid is used for every subject and subgroup. Then, $\mu(x_s)$, $\eta_j(x_s)$, $K_T(x_s, x_r)$ and $K_B(x_s, x_r)$ can be estimated as follows:

$$\hat{\mu}(x_s) = \frac{1}{IJ}\sum_{i,j} Y_{i,j}(x_s),$$

$$\hat{\eta}_j(x_s) = \frac{1}{I}\sum_{i} Y_{i,j}(x_s) - \hat{\mu}(x_s),$$

$$\widehat{K}_T(x_s, x_r) = \frac{1}{IJ}\sum_{i,j}\{Y_{i,j}(x_s) - \hat{\mu}(x_s) - \hat{\eta}_j(x_s)\}\{Y_{i,j}(x_r) - \hat{\mu}(x_r) - \hat{\eta}_j(x_r)\},$$

$$\widehat{K}_B(x_s, x_r) = \frac{2}{IJ(J-1)}\sum_{i}\sum_{j_1<j_2}\{Y_{i,j_1}(x_s) - \hat{\mu}(x_s) - \hat{\eta}_{j_1}(x_s)\}\{Y_{i,j_2}(x_r) - \hat{\mu}(x_r) - \hat{\eta}_{j_2}(x_r)\}.$$

Since $K_W(x_s, x_r)$ is estimated by the difference between $\widehat{K}_T(x_s, x_r)$ and $\widehat{K}_B(x_s, x_r)$, it may not always be positive definite. This can be solved by trimming pairs of eigenvalue and eigenfunction where the eigenvalue is negative (Hall et al 2008). Consequently the eigenfunctions can be estimated based on the decomposition of covariance functions as follows:

Step 1. Estimate the mean and covariance functions $\hat{\mu}(x_s)$, $\hat{\eta}_j(x_s)$, $\widehat{K}_T(x_s, x_r)$ and $\widehat{K}_B(x_s, x_r)$; set $\widehat{K}_W(x_s, x_r) = \widehat{K}_T(x_s, x_r) - \widehat{K}_B(x_s, x_r)$.

Step 2. Decompose $\widehat{K}_B(x_s, x_r)$ to obtain $\hat{\lambda}_k^{(1)}$, $\hat{\phi}_k^{(1)}(x)$;

Step 3. Decompose $\widehat{K}_W(x_s, x_r)$ to obtain $\hat{\lambda}_l^{(2)}$, $\hat{\phi}_l^{(2)}(x)$; Then trim those pairs of eigenvalue and eigenfunction where the eigenvalue is negative.

Step 4. Estimate the principal component scores for both levels (see below).

As discussed in the previous subsection, in standard FPCA the principal component scores (PC scores) can be estimated straightforwardly by numerical integration, using $\beta_k = \int_{\mathcal{T}} (Y(x) - \mu(x))\phi_k(x)\,dx$. However, estimating the PC scores for multilevel functional data involves extra complication because the two levels of eigenfunctions, namely $\phi_k^{(1)}(x)$ and $\phi_l^{(2)}(x)$, are not necessarily mutually orthogonal. Di et al (2009) assumes that $\beta_{i,k}$ and $\gamma_{i,j,l}$

both follow Gaussian distributions and proposes two parallel methods for estimating the scores. In the first method, the estimated $\mu(x), \eta_j(x), \lambda_k^{(1)}, \lambda_l^{(2)}, \phi_k^{(1)}(x)$ and $\phi_l^{(2)}(x)$ are treated as fixed while the PC scores $\beta_{i,k}$ and $\gamma_{i,j,l}$ are random effects to be estimated. The model can then be written as:

$$\begin{cases} Y_{i,j}(x) = \mu(x) + \eta_j(x) + \sum_{k=1}^{N_1} \beta_{i,k}\phi_k^{(1)}(x) + \sum_{l=1}^{N_2} \gamma_{i,j,l}\phi_l^{(2)}(x) + \varepsilon_{i,j}(x); \\ \beta_{i,k} \sim N\left(0, \lambda_k^{(1)}\right); \quad \gamma_{i,j,l} \sim N\left(0, \lambda_l^{(2)}\right); \quad \varepsilon_{i,j}(x) \sim N(0, \sigma^2), \end{cases}$$

where $\varepsilon_{i,j}(x)$ is i.i.d and appears only when functional data are observed with error (if the functional data is already smoothed or perfectly observed, simply remove this term from the model). And σ^2 is the prior parameter (assuming $1/\sigma^2$ to follow a gamma distribution with mean equal to one and a large variance). It is in fact a linear mixed effect model in nature, therefore the random effects can be estimated by either best linear unbiased prediction (BLUP) or Markov Chain Monte Carlo (MCMC). This model is appropriate for both dense and sparse functional data. However, it faces computational challenges, especially when dealing with large volume of data. Therefore the second method will be adopted in our proposed model.

The intuition behind the second method is to project each mean centred function into the space spanned by each eigenfunction at different levels (Di et al 2009). We denote

$$A_{i,j,k} = \int_0^1 \{Y_{i,j}(x) - \mu(x) - \eta_j(x)\}\phi_k^{(1)}(x)dx = \beta_{i,k} + \sum_{l=1}^{N_2} \gamma_{i,j,l}c_{k,l} + \epsilon_{i,j,k}^{(1)}, \qquad (3.2)$$

$$B_{i,j,l} = \int_0^1 \{Y_{i,j}(x) - \mu(x) - \eta_j(x)\}\phi_l^{(2)}(x)dx = \gamma_{i,j,l} + \sum_{k=1}^{N_1} \beta_{i,k}c_{k,l} + \epsilon_{i,j,l}^{(2)}, \qquad (3.3)$$

where $c_{k,l} = \int_0^1 \phi_k^{(1)}(x)\,\phi_l^{(2)}(x)dx$ is the inner product of two eigenfunctions at different levels, and $\epsilon_{i,j,k}^{(1)}$ and $\epsilon_{i,j,l}^{(2)}$ are the residuals which can be expressed as

$$\epsilon_{i,j,k}^{(1)} = \int_0^1 \{\sum_{l=N_2+1}^{\infty} \gamma_{i,j,l}\phi_l^{(2)}(x) + \varepsilon_{i,j}(x)\}\phi_k^{(1)}(x)dx$$

$$\epsilon_{i,j,k}^{(2)} = \int_0^1 \{\sum_{k=N_1+1}^{\infty} \beta_{i,k}\phi_k^{(1)}(x) + \varepsilon_{i,j}(x)\}\phi_l^{(2)}(x)dx$$

respectively. Equations (3.2) and (3.3) can be rewritten in matrix format. Letting

$$\boldsymbol{A}_{i,j} = (A_{i,j,1}, A_{i,j,2}, \dots, A_{i,j,N_1})^T,\ \boldsymbol{B}_{i,j} = (B_{i,j,1}, B_{i,j,2}, \dots, B_{i,j,N_2})^T,$$

$$\beta_i = (\beta_{i,1}, \beta_{i,2}, \dots, \beta_{i,N_1})^T,\ \ \gamma_{i,j} = (\gamma_{i,j,1}, \gamma_{i,j,2}, \dots, \gamma_{i,j,N_2})^T,$$

$$\epsilon_{i,j}^{(1)} = (\epsilon_{i,j,1}^{(1)}, \epsilon_{i,j,2}^{(1)}, \dots, \epsilon_{i,j,N_1}^{(1)})^T,\ \epsilon_{i,j}^{(2)} = (\epsilon_{i,j,1}^{(2)}, \epsilon_{i,j,2}^{(2)}, \dots, \epsilon_{i,j,N_2}^{(2)})^T,$$

Then (2) and (3) are transformed as follows:

$$\begin{cases} \boldsymbol{A}_{i,j} = \beta_i + C\gamma_{i,j} + \epsilon_{i,j}^{(1)}, \quad \boldsymbol{B}_{i,j} = \gamma_{i,j} + C^T\beta_i + \epsilon_{i,j}^{(2)}, \\ \beta_i \sim N\left(0,\ \Lambda^{(1)}\right), \quad \gamma_{i,j} \sim N\left(0,\ \Lambda^{(2)}\right), \\ \epsilon_{i,j}^{(1)} \sim N\left(0,\ \sigma_1^2\boldsymbol{I}_{N_1}\right), \quad \epsilon_{i,j}^{(2)} \sim N\left(0,\ \sigma_2^2\boldsymbol{I}_{N_2}\right), \end{cases}$$

where $\mathrm{C} = (c_{k,l})_{k,l}$ is an $N_1 \times N_2$ matrix, $\Lambda^{(1)} = diag(\lambda_1^{(1)}, \lambda_2^{(1)}, \dots, \lambda_{N_1}^{(1)})$, and $\Lambda^{(2)} = diag(\lambda_1^{(2)}, \lambda_2^{(2)}, \dots, \lambda_{N_2}^{(2)})$. The residual variance σ_1^2 and σ_2^2 can be estimated from the data. This is also a linear mixed effects model and MCMC is implemented for estimation of the PC scores.

Determining the number of eigenfunctions in the approximation is always important in the context of principal component analysis. A relatively simple solution is to rely on the proportion of variance explained. In the MFPCA model, the cumulative percentage of the total variation explained by the first N_1 components at level one is given by the ratio $\sum_{k=1}^{N_1}\lambda_k^{(1)}/\sum_{k=1}^{\infty}\lambda_k^{(1)}$. Similarly, the cumulative percentage of the total variation explained by the first N_2 components at level two is given by the ratio $\sum_{l=1}^{N_2}\lambda_l^{(2)}/\sum_{l=1}^{\infty}\lambda_l^{(2)}$. Therefore the number of eigenfunction can be determined given a threshold for the cumulative percentage of the total variation explained by the first N_1 (N_2) components at level one (level two). And the proportion of variation explained by the level one and level two against the overall variation of the entire data set can be calculated as $\frac{\sum_{k=1}^{\infty}\lambda_k^{(1)}}{\sum_{k=1}^{\infty}\lambda_k^{(1)}+\sum_{l=1}^{\infty}\lambda_l^{(2)}}$ and $\frac{\sum_{l=1}^{\infty}\lambda_l^{(2)}}{\sum_{k=1}^{\infty}\lambda_k^{(1)}+\sum_{l=1}^{\infty}\lambda_l^{(2)}}$ respectively.

In reality, functional data are usually observed with error. Smoothing is therefore needed before analysis is carried out. In the literature of FPCA, there are three methods commonly used for smoothing: smoothing the data before applying FPCA, introducing a penalty term

(Ramsay and Silverman 2005), and smoothing the covariance function. In our later application on mortality modelling, we will adopt the first method to smooth the raw data first before the MFPCA is conducted. This will be discussed again in the next section.

3.2.3 Weighted MFPCA for coherent mortality forecasting

In this section we propose a weighted MFPCA approach for coherent forecast of the future mortality of a number of subpopulations within one large population. Let $y_{t,j}(x)$ denote the log of the death rate of the j^{th} subpopulation for age *x* in year *t*. And we assume that there is an underlying function $f_{t,j}(x)$ that we are observing with error at discrete points of x. Suppose we have observed $\{x_i, y_{t,j}(x_i)\},\ t = 1, \dots, n, i = 1, \dots, p, j = 1, \dots, m$. Then,

$$y_{t,j}(x_i) = f_{t,j}(x_i) + \sigma_{t,j}(x_i)e_{t,j,i},$$

where $e_{t,j,i}$ is i.i.d standard normal random variables and $\sigma_{t,j}(x_i)$ allows the amount of noise to vary with age x.

In demographic modelling, it is often the case that more recent data tend to have more impact on the results than those in the distant past. Hence, we propose to incorporate the concept of weight into the MFPCA model to allow recent mortality data to affect forecasting result more significantly, as discussed below.

The overall mean function $\mu(x)$ is estimated using a weighted average:

$$\hat{\mu}(x) = \sum_{t,j}(w_t/m)f_{t,j}\,(x),$$

where $f_{t,j}(x)$ is the smoothed function from $y_{t,j}(x)$, m is the total number of subgroups, and $w_t = \kappa(1-\kappa)^{n-t}$ is a geometrically decaying weight with $0 < \kappa < 1$. The larger κ is, the faster the weight for the past years is decaying over time. Therefore, κ represents people's perception on how past data should be weighted. The parameter κ can be determined by cross validation or specified *a priori*.

The subgroup-specific shift from the overall mean, $\eta_j(t)$, is then estimated as:

$$\hat{\eta}_j(t) = \sum_t w_t f_{t,j}(x) - \hat{\mu}(x).$$

The mean-adjusted functional data are denoted as $\hat{f}^*_{t,j}(x) = f_{t,j}(x) - \hat{\mu}(x) - \hat{\eta}_j(t)$. Note that the weights are incorporated into the mean-adjusted data $\hat{f}^*_{t,j}(x)$ in order to calculate the weighted functional principal components. Similar to Hyndman and Ullah (2007), the weighted multilevel functional principal components $\hat{\phi}^{(1)}_k(x)$ and $\hat{\phi}^{(2)}_l(x)$ can be obtained as follows. We first discretize $\hat{f}^*_{t,j}(x)$ on a dense grid of p equally spaced points $\{x_1, \dots, x_p\}$ and denote the discretized $\hat{f}^*_{t,j}(x)$ as an $nm \times p$ matrix F. Then the weights are incorporated by multiplying a weight matrix W to F, where $W = diag(\underbrace{\frac{w_1}{m}, \frac{w_1}{m}, \dots}_{m}, \dots, \underbrace{\frac{w_n}{m}, \frac{w_n}{m}, \dots}_{m})$. Thus F is transformed into $F^* = WF$. Based on the weighted mean-adjusted matrix F^*, the covariance functions $\widehat{K}_T(x_s, x_r)$ and $\widehat{K}_B(x_s, x_r)$ can be estimated using the methods discussed in the previous subsection and $\widehat{K}_W(x_s, x_r) = \widehat{K}_T(x_s, x_r) - \widehat{K}_B(x_s, x_r)$ is calculated accordingly. The weighted multilevel functional principal components $\hat{\phi}^{(1)}_k(x)$ and $\hat{\phi}^{(2)}_l(x)$ can then be obtained by decomposing $\widehat{K}_B(x_s, x_r)$ and $\widehat{K}_W(x_s, x_r)$. And finally, the PC scores at two levels can be directly estimated using either of the methods described before. Therefore, the entire weighted MFPCA model for coherent forecasting of the mortality of subpopulations is given as:

$$y_{t,j}(x_i) = \mu_j(x_i) + \sum_{k=1}^{N_1} \beta_{t,k} \phi^{(1)}_k(x_i) + \sum_{l=1}^{N_2} \gamma_{t,j,l} \phi^{(2)}_l(x_i) + \sigma_{t,j}(x_i) e_{t,j,i},$$

where $\mu_j(x_i) = \mu(x_i) + \eta_j(x_i)$ is the mean of the subpopulation j. Forecasts can then be achieved by extrapolating the level-one and level-two scores $\beta_{t,k} = \{\beta_{t,1}, \dots, \beta_{t,N_1}\}$ and $\gamma_{t,j,l} = \{\gamma_{t,j,1}, \dots, \gamma_{t,j,N_2}\}, j = 1, \dots m$, using time series models. Since the level-one scores are independent we assume independent possibly non-stationary autoregressive integrated moving average (ARIMA) models for each of $\{\beta_{t,1}, \dots, \beta_{t,N_1}\}$. As for the level-two scores $\gamma_{t,j,l}$, it is noted that, for the same order l $(l = 1, \dots, N_2)$, the set of scores for different subpopulations $\{\gamma_{t,1,l}, \dots, \gamma_{t,m,l}\}$ share the same basis $\phi^{(2)}_l(x)$, which implies that the scores $\{\gamma_{t,1,l}, \dots, \gamma_{t,m,l}\}$ are not independent and hence multivariate time series models are more appropriate. However, in order to avoid model and computation complexity, a univariate autoregressive moving average (ARMA) model with stationary restriction is used for each of

the level-two scores in our numerical examples. The stationary constraint is to ensure the coherence in mortality forecasting.

Let $\hat{\beta}_{(t+h),k}$ denote the h-step ahead forecast of $\beta_{(t+h),k}$ and $\hat{\gamma}_{(t+h),j,l}$ denote the h-step ahead forecast of $\gamma_{(t+h),j,l}$. Then the h-step ahead forecast of $y_{t,j}(x)$ is obtained as:

$$\hat{y}_{(t+h),j}(x) = \hat{\mu}_j(x) + \sum_{k=1}^{N_1} \hat{\beta}_{(t+h),k}\hat{\phi}_k^{(1)}(x) + \sum_{l=1}^{N_2} \hat{\gamma}_{(t+h),j,l}\hat{\phi}_l^{(2)}(x).$$

Due to the fact that each component in the model is uncorrelated with each other, the forecasting variance can be obtained by adding up the variance of each component:

$$var\{y_{(t+h),j}(x)\} = \hat{\sigma}^2_{\mu_j}(x) + \sum_{k=1}^{N_1} u_{(t+h),k}\{\hat{\phi}_k^{(1)}(x)\}^2$$

$$+ \sum_{l=1}^{N_2} v_{(t+h),j,l}\{\hat{\phi}_l^{(2)}(x)\}^2 + \{\sigma_{(t+h),j}(x)\}^2,$$

where $\hat{\sigma}^2_{\mu_j}(x)$ denotes the variance of the smoothed means and can be estimated from the smoothing method used; $u_{(t+h),k} = var\{\hat{\beta}_{(t+h),k}\}$ and $v_{(t+h),j,l} = var\{\hat{\gamma}_{(t+h),j,l}\}$ can be obtained from the time series models; and the observational variance $\{\sigma_{(t+h),j}(x)\}^2$ can be estimated from the historical data (see Hyndman and Ullah 2007).

3.3 Applications

In this section we evaluate our weighted MFPCA model by applying it to some sex-specific mortality data. We first consider the coherent forecasting of the sex-specific mortality rates in the UK, and forecast the male and female life expectancies in the UK with thirty years horizon to see if the forecasting result appears to be non-divergent. We then compare the forecasting accuracy of our model with that of the Product-Ratio model as well as the independent model using the sex-specific mortality rates of nine different countries.

3.3.1 Coherent forecasting for the male and female mortality in the UK

The sex-specific mortality data in the UK are obtained from the Human Mortality Database (2010). The data provide the observed mortality rates for every one year at per age. We select the years from 1950 to 2010 and the ages from 0 to 100 to avoid the anomalous mortality rates during the first and second world wars and the erratic rates above 100. The mortality curves are first smoothed by using the weighted penalized regression splines with a monotonicity constraint (Hyndman and Ullah, 2007). The smoothed curves are shown in Figure 3.1.

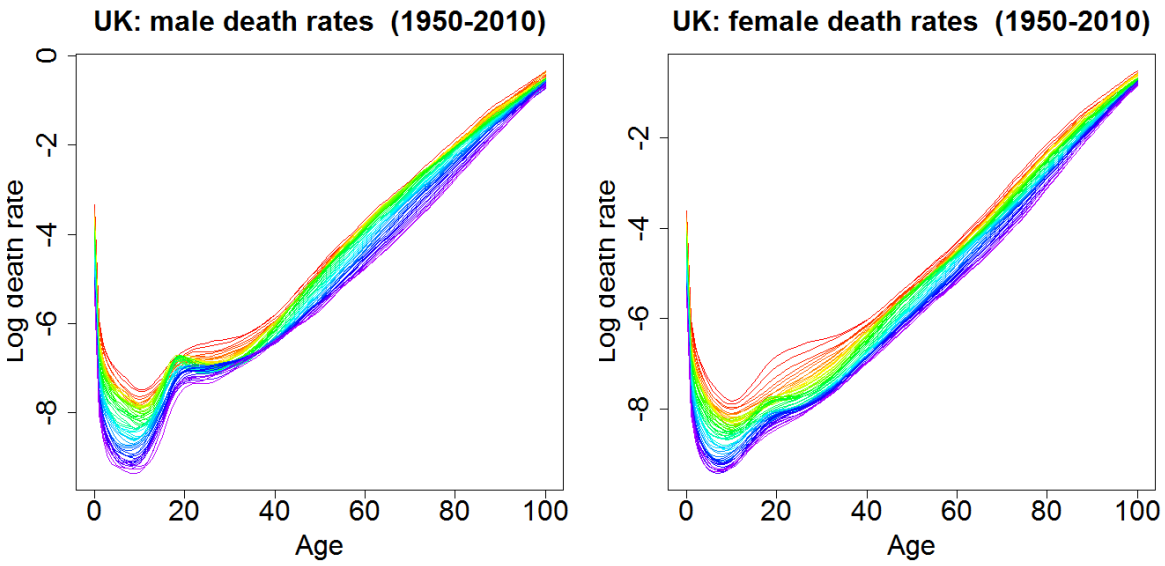


Figure 3.1: The smoothed log death rates for male and female in the UK from 1950 to 2010, viewed as functional data series.

The weighted MFPCA model is then fitted to the data. The weight parameter κ is set as $\kappa = 0.05$ as suggested by Hyndman et al (2013) in their studies, thus giving a weight of $0.05(0.95) = 0.0475$ to the most recent year, $0.05(0.95)^2 = 0.0451$ to the year before that, and so on. The numbers of principal components for both level-one and level-two are $N_1 = N_2 = 3$ since the first three PCs have accounted for more than 85% of the variation. The level-one and level-two mean functions, functional principal components as well as their corresponding scores are estimated as discussed in the previous section. The time series modelling and forecasting for the scores are performed using the R package 'forecast'

(Hyndman and Khandakar, 2008). The estimates and the forecasts with 30-year horizon are shown in Figures 3.2 and 3.3 respectively.

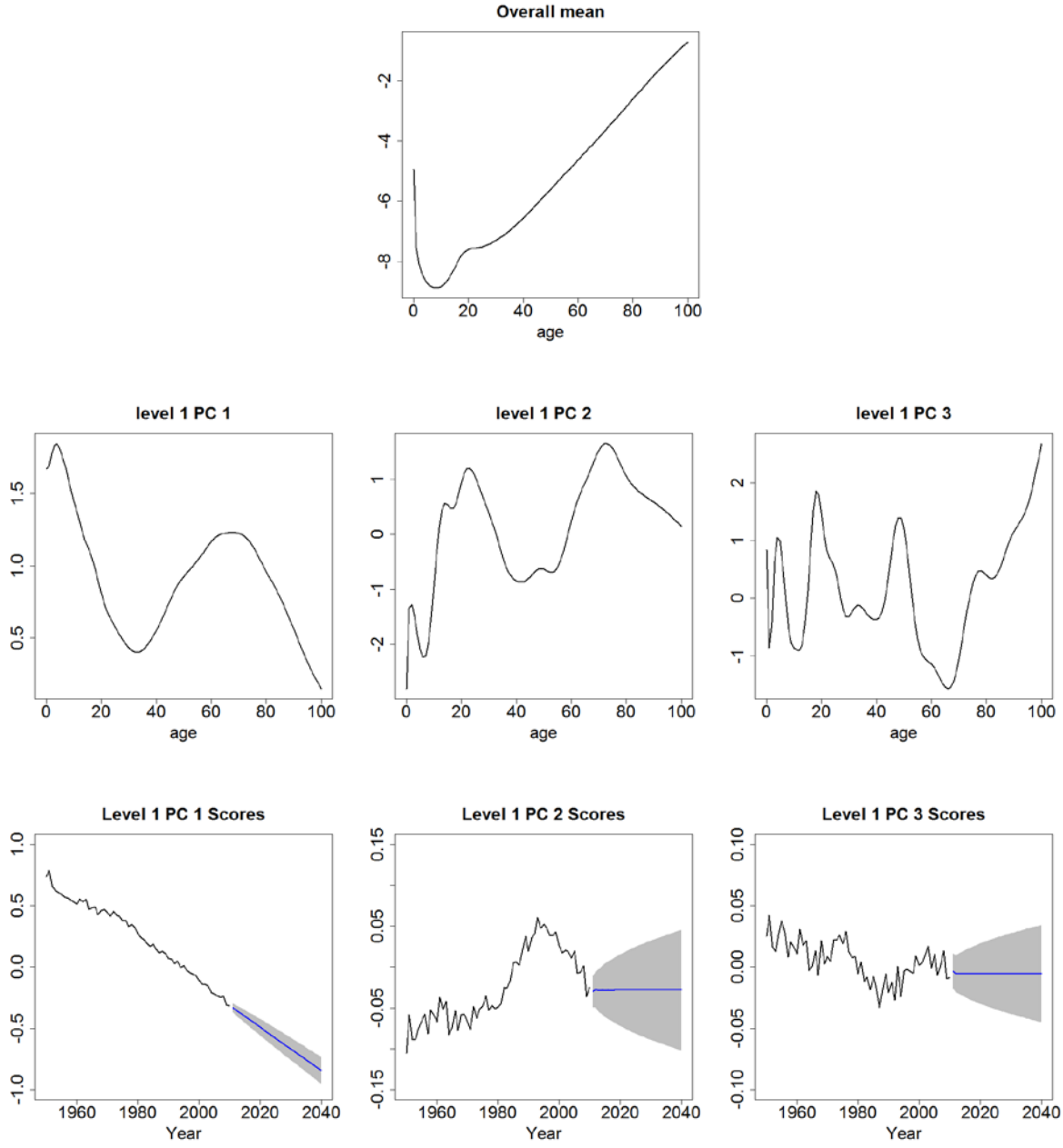


Figure 3.2**:** Level-one decomposition: the overall mean function, the first three level-one functional principal components and their corresponding scores with 30-year forecast horizon and 80% confidence interval using ARIMA models without restriction.

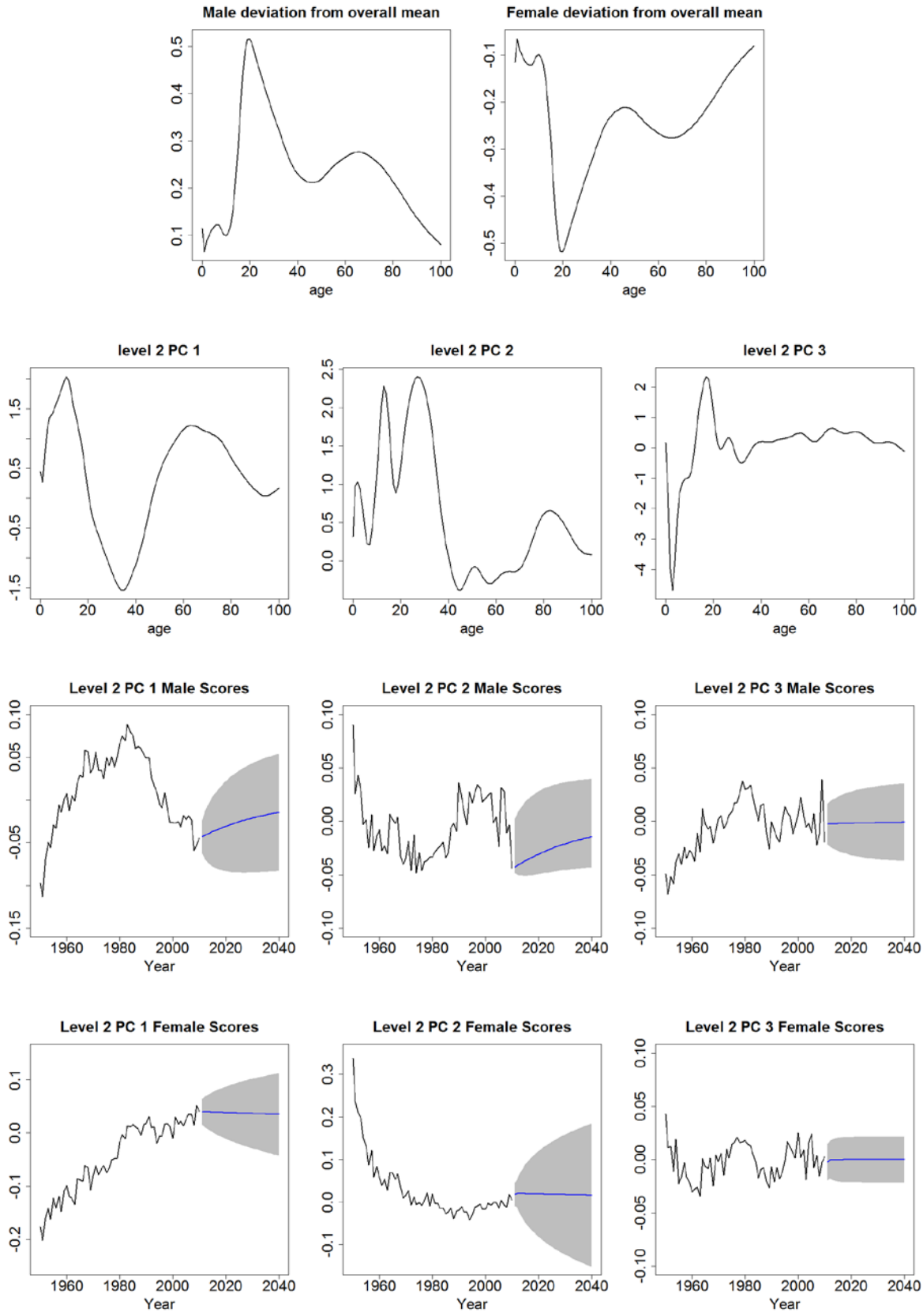

Male deviation from overall mean
Female deviation from overall mean
age
level 2 PC 1
level 2 PC 2
level 2 PC 3
Level 2 PC 1 Male Scores
Level 2 PC 2 Male Scores
Level 2 PC 3 Male Scores
Level 2 PC 1 Female Scores
Level 2 PC 2 Female Scores
Level 2 PC 3 Female Scores
Year

Figure 3.3: Level-two decomposition: the sex-specific deviations from the overall mean function, the first three level-two functional principal components and their corresponding scores with 30-year forecast horizon and 80% confidence interval using stationary ARMA models.

Figure 3.2 displays the level-one components which are shared by both male and female subpopulations. As can be seen in the figure, the first principal component of level-one models the degree of variation of mortality among different age groups, which is very large for child age groups and relatively small for twenties and old age groups. The second and third principal components present more complicated patterns and are difficult to interpret. The result shows that the first three principal components explain around 97.7%, 1.6% and 0.3% of the level one variation respectively, and the level-one variation takes up 94% of the total variation.

Figure 3.3 displays the level-two components which are specific for male and female subpopulations (except for the level-two principal components which are shared by both male and female). It can be seen that the male and female deviations from the overall mean are complementary to each other, and such pattern can be verified if the number of subgroups is two, because of the method used for calculating the mean and deviations. The first three principal components explain around 61.8%, 17.2% and 7.0% of the level-two variation respectively, and the level-two variation takes up 6% of the total variation. As mentioned before, univariate stationary ARMA models are used to model the level-two scores in order to reduce model and computation complexity, despite that they are theoretically dependent.

Figure 3.4 shows the 30-year forecasts of the male and female life expectancies at birth by our proposed model as well as the independent model (Hyndman and Ullah, 2007). It can be observed that the independent model produces a more divergent forecasting result than our coherent model does. The life expectancies of male and female in 2040 forecasted by the independent model are 82.5 years and 86.8 years respectively, in contrast to 83.1 years and 86.6 years by our weighted MFPCA model. It can also be viewed that the forecasted life expectancies of male and female in 2010 are 78.4 years and 82.3 years respectively, indicating a 3.9 years sex gap. This gap is increased by the independent model to 4.3 years in 2040 while it is decreased by our coherent model to 3.5 years. Hence, our coherent model demonstrates slow convergence in forecasted life expectancies in the UK.

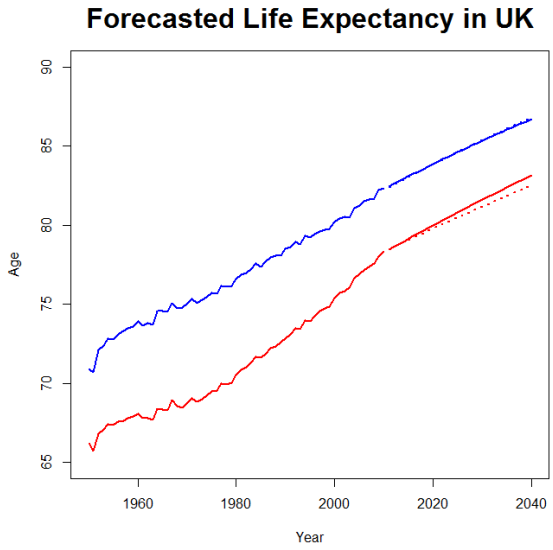


Figure 3.4: 30-year forecasted life expectancies for male and female in the UK by the weighted MFPCA model (solid lines) and the independent model (dotted lines). Blue lines represents female group, while red lines represent male group.

3.3.2 Comparing accuracy with the Product-Ratio model and the independent model

We now compare the forecasting accuracy of our proposed model with those of the Product-Ratio model and the independent model for the UK mortality rates. We use the UK male and female mortality data from 1950 to 1973+t as observations and forecast the mortality rates for years 1973+t + 1, …, 1973+t + 30, for $t = 0, \dots, 9$. Thus, ten sets of forecasts with 1 to 30 year horizons are obtained and compared with the actual values. For a specific forecast horizon h ($h = 1, \dots, 30$), the out-of-sample root mean square forecast error (RMSFE) for the j^{th} subpopulation is defined as:

$$RMSFE_j(h) = \sqrt{\frac{1}{10p}\sum_{t=0}^{9}\sum_{i=1}^{p}\{y_{(1973+t+h),j}(x_i) - \hat{y}_{(1973+t+h),j}(x_i)\}^2}.$$

For the UK male and female mortality data, the values of the out-of-sample RMSFE are obtained using the weighted MFPCA model, the Product-Ratio model and the independent model and are illustrated in the top panels of Figure 3.5.

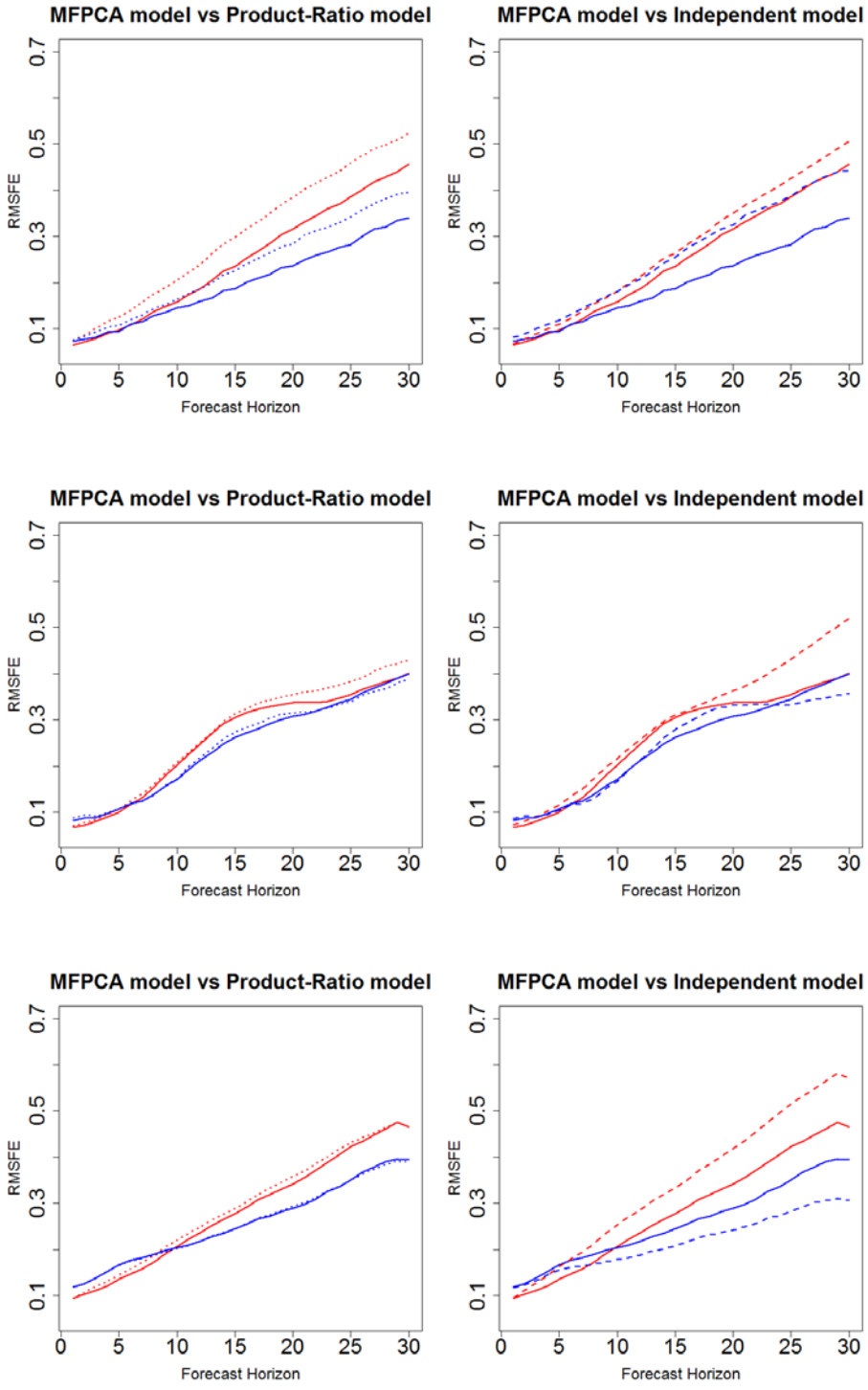


Figure 3.5**:** Out-of-sample RMSFEs for male and female of three countries using the weighted MFPCA model (solid lines), the Product-Ratio model (dotted lines) and the independent model (dashed lines). Top panels: the UK; the middle: Italy; the bottom: Australia. The left panel shows the comparison between the weighted MFPCA and the Product-Ratio model whilst the right panel is between the weighted MFPCA and the independent model. Blue lines represents female group and red lines represent male group.

It can be observed that the MFPCA model presents a considerably higher forecasting accuracy than the Product-Ratio model and the independent model, for both male and female groups. The average RMSFE (over forecast horizon and sex) by the MFPCA model is 0.1916, compared with 0.2672 by the Product-Ratio model and 0.2685 by the independent model. The numerical results show that, in overall terms, the MFPCA model performs considerably better than the Product-Ratio model and the independent model for the UK mortality forecast.

The same experiment is further performed to the male and female mortality data of eight other developed countries, including Australia, USA, France, Japan, Spain, Canada, Netherlands and Italy. As demonstration the out-of-sample RMSFEs for Italy and Australia are shown in the lower panels of Figure 12.

It can be seen from the above figures that the RMSFE by the MFPCA model is slightly smaller than that of the Product-Ratio model for Italian male for almost all forecast horizons. For Italian female, the MFPCA model has a marginally smaller RMSFE for forecast horizons from 1 to 20. After 20 year horizon, the RMSFE of the MFPCA model exceeds that of the Product-Ratio model marginally. And the RMSFE of the MFPCA model is slightly smaller than that of the independent model for all forecast horizons, for both Italian male and female. Overall, for the Italian case the MFPCA model performs slightly better than the Product-Ratio model (average RMSFE 0.2512 vs 0.2572) while it performs much better than the independent model (average RMSFE 0.2512 vs 0.2694) in terms of forecast accuracy.

For Australian male and female the RMSFEs of the MFPCA model are almost the same as those by the Product-Ratio model. Compared with the independent model, the forecast error by the MFPCA model is lowered for Australian male, at the cost of the increase in error for the female. In other words, the MFPCA model homogenizes the forecast error across male and female, compared with the independent model. Overall, for the Australian case the MFPCA model performs almost as well as the Product-Ratio model (average RMSFE 0.2774 vs 0.2757) while it performs marginally better than the independent model (average RMSFE 0.2774 vs 0.2806) in terms of forecast accuracy.

The values of the average RMSFE by the three models for all nine countries are given in Table 3.1.

	MFPCA	Product-Ratio	Independent
AUS	0.2774	0.2757	0.2806
USA	0.1568	0.1247	0.1614
UK	0.1916	0.2672	0.2685
FRA	0.2483	0.2188	0.2362
JPN	0.3616	0.3551	0.3614
ESP	0.2855	0.2766	0.3404
CAN	0.2353	0.2039	0.2451
NLD	0.2415	0.2383	0.2851
ITA	0.2512	0.2572	0.2694

Table 3.1**:** The average RMSFEs by the weighted MFPCA model, the Product-Ratio model and the independent model for nine developed countries.

It can be observed that the MFPCA model outperforms the independent models in most of the cases, and is comparable to the Product-Ratio model in terms of average RMSFE. For the UK mortality the MFPCA model presents a significant advantage over the other two models.

Further analysing the experimental result, we notice that the MFPCA model presents a significant advantage over the other two models in minimizing the short-term RMSFE (average RMSFE for 1 to 10-year horizon). With respect to short-term RMSFE, the MFPCA model outperforms the other two models for 7 out of 9 countries. The values of the short-term RMSFE by the three models for all nine countries are given in Table 3.2.

	MFPCA	Product-Ratio	Independent
AUS	0.1548	0.1592	0.1617
USA	0.0927	0.0886	0.0927
UK	0.1065	0.1271	0.1243
FRA	0.1115	0.1116	0.1139
JPN	0.1262	0.1507	0.1522
ESP	0.1615	0.1618	0.1801
CAN	0.1406	0.1305	0.1425
NLD	0.1547	0.1586	0.1697
ITA	0.1182	0.1220	0.1243

Table 3.2: The short-term RMSFEs (average of 1 to 10-year horizon) by the weighted MFPCA model, the Product-Ratio model and the independent model for nine developed countries.

3.4 Conclusion

In this chapter, we proposed a new model for coherent forecasting of mortality rates among several subpopulations. The model is based on the multilevel functional principal component analysis (MFPCA) framework and is incorporated with weights which allow more recent data to have more significant impact on forecasting result. The mortality curves of different subpopulations are treated as a set of multilevel functional data, and the principal components and their corresponding scores are obtained at two levels and forecasts are made by extrapolating these scores using time series models. Using the sex-specific mortality data of nine developed countries, we demonstrated the usefulness of the proposed model and compared the results with the Product-Ratio model and the independent model.

The MFPCA model consists of a group mean, a decomposition of level-one functions and a decomposition of level-two functions. The level-one functions govern the common properties of the entire population while the level-two functions involve properties which are specific to subpopulations. Therefore the model possesses a simple and explicit form which makes

modelling procedure easy and interpretable. Since the level-one functions are largely dominated by their first principal component, the age pattern of change at this level is relatively fixed. However, the governance of level-two functions are separated into several principal components, which gives a flexible age pattern of change at level two. In such way, the MFPCA model provides more flexibility in the age pattern of change compared with the independent model which contains only a single level of principal components.

It can be seen that the MFPCA model has a similar structure to the Product-Ratio model. For instance, the Product-Ration model is composed of a group mean, a decomposition of product function and a decomposition of ratio function. The product function represents the geometric mean of subpopulation death rates while the ratio function is the ratio of the death rates of a specific subpopulation to the geometric mean. Despite these similarities, the MFPCA model has several advantages over the Product-Ratio model in its nature. The core of the MFPCA method is to find the variations of multilevel functional data at different levels and then decompose them using KL expansion. There is no need to pre-processing the data as done in the Product-Ratio method (calculating the product functions and the ratio functions) so the functional data are allowed to speak more for themselves within the MFPCA framework. The Product-Ratio model assumes the subpopulations have equal variances, which is not needed in the MFPCA model. Moreover, within the MFPCA framework, the percentage of variance explained by each principal component at both levels can be calculated explicitly and easily, which cannot be achieved by the Product-Ratio model. Knowing these percentages of variance explained, we are able to explicitly evaluate the importance of every principal component at both levels.

Coherence is another important issue to be discussed. Hyndman et al (2013) defines coherence as the convergence of the ratios of the forecast age-specific death rates from any two subpopulations to appropriate constants. In our model, coherence boils down to the convergence of the level-two functions of any two subpopulations to appropriate constants, which is ensured by applying stationary ARMA models to the level-two principal component scores. The convergence of the level-two principal component scores to certain constants under stationary ARMA model guarantees that the long-term forecasts of the level-two functions also converge to their age-specific constants. As the level-two functions converge to constants, they gradually lose ability to affect the change of mortality. The change of

mortality is then entirely dominated by level-one functions. Since level-one functions are commonly shared by different subpopulations, their impact on the changes of mortality of those subpopulations is equal. The forecasted mortality differences among subpopulations are thus constrained, leading to a similar constraint on the forecasted life expectancies as well.

Chapter 4
Clustering mortality (fertility) as functional data using principal curve method

4.1 Clustering functional data

Functional data is collection of data represented by curves rather than data points alone. The clustering of functional data, often used as a preliminary step for functional data exploration, involves extra complexity since the dimension of the data takes values into an infinite dimensional space. In this chapter, we propose an innovative clustering method for functional data, with the aid of the principal curves. The proposed method borrows the strength of nonparametric principal curves to effectively detect the potential features of the two-dimensional scores extracted from the original functional data for clustering purpose. A probability model with Bayesian Information Criterion (BIC) for principal curves is constructed to automatically find the appropriate number of features and the optimal degree of smoothing. We make use of this approach for clustering human mortality and fertility as functional data.

4.2 Principal curve method for clustering functional data

4.2.1 Principal curves

A principal curve is a smooth, one dimensional nonparametric curve that passes through the "middle" of a p-dimensional data set. Different from the principal component, which is linear summarization of data, a principal curve allows for nonlinearity in summarizing the data and it is actually an extension of the first principal component. The fundamental idea was introduced by Hastie and Stuetzle (1989) (hereafter HS), using the concept of self-

consistency, which means that each point of the curve is the average over all points that project there. According to HS, the projection index $\lambda_f(x)$: $\boldsymbol{R}^p \rightarrow \boldsymbol{R}^1$ is defined as:

$$\lambda_f(x) = sup_\lambda\{\lambda: \|x - f(\lambda)\| = inf_\mu \|x - f(\mu)\|\},$$

where x is a sample from a random vector X in $\boldsymbol{R}^p$ with density h and λ_f is the function that projects points in $\boldsymbol{R}^p$ orthogonally onto the curve f. The projection index is the value of λ for which $f(\lambda)$ is closest to x. If there are multiple values of λ, the largest one will be selected. Then, the curve is called principal curve of h if it satisfies:

$$E\big(X\big|\lambda_f(X) = \lambda\big) = f(\lambda),$$

for almost all λ. A principal curve is parameterized by λ, the arc length along the curve.

The algorithm for finding a principal curve involves starting with the first principal component as a smooth curve and then repeating the projection and conditional expectation from the above definitions until the convergence is achieved. And the conditional expectation is usually replaced by a scatterplot smoother, where the conditional expectation at λ_i is estimated by averaging all the observations x_k in the sample for which the corresponding λ_k is close to λ_i. This algorithm, motivated by the definition of finding principal curves from sample, is considered to estimate a population quantity that minimizes a population criterion. On the other hand, HS also provides an alternative algorithm to fit the principal curves by using cubic smoothing splines, which minimizes data-dependent criterion. This spline-smoothing algorithm will be discussed in more details in next sub-section. Besides this classical definition given by HS, there exist some other definitions, for example, by Tibshirani (1992) and Kégl et al (2000). These algorithms share something in common: they all start with a straight line (very often the first principal component) and then evolve the line(s) until it converges to the middle of data satisfactorily. This family of algorithms are called “top-down” strategies.

Parallel to that, there is another family of algorithms called “bottom-up” strategies. Instead of starting with a global initial line, the “bottom-up” method constructs the principal curve by taking into account the data in a local neighbourhood of the current point considered in every step. Hence it is able to handle complex data structure such as spirals or branched curves which “top-down” method is often unable to handle. Delicado (2001) first introduces a principal curve approach that belongs to this family. Later on, Einbeck et al (2005) adopt Delicado’s concept and simplified his algorithm to be fast and efficient in computation.

Einbeck's approach can be considered as a simple and fast approximation to Delicado's algorithm and is called the *local principal curve* method.
In this chapter, we adopt the classical HS algorithm as a stable method to construct principal curves for our functional clustering approach.

4.2.2 The spline-smoothing algorithm by HS

Assume X is a random variable in $\boldsymbol{R}^p$ and $x_i \in \boldsymbol{R}^p, i = 1, \dots, n$ are samples from X. According to HS, the criterion for defining principal curves in this context is: find $f(\lambda)$ and $\lambda_i \in [0,1]$ $(i = 1, \dots, n)$ so that the penalized least squares

$$D^2(f, \lambda) = \sum_{i=1}^{n} \|x_i - f(\lambda_i)\|^2 + \eta \int_0^1 \|f''(\lambda)\|^2 d\lambda$$

is minimized over all f with the penalty (smoothing parameter) η. And the spline-smoothing algorithm designed on this criterion is given as following:

(1) Given f, minimizing $D^2(f, \lambda)$ over λ_i which only involves the squared distances part and this is the usual projection step. Then rescale λ_i to lie in $[0,1]$.

(2) Given λ_i, split the penalized least squares into p expressions, one for each coordinate function. Then smooth each coordinate separately against λ_i using a cubic spline smoother with parameter η.

It is suggested by HS that if a minimum of the penalized least squares exists, it must be a cubic spline in each coordinate. It can be difficult to guess the smoothing parameter η under some circumstances. We choose to use an alternative method which employs the degree of freedom (DF) to determine the amount of smoothing. And the DF of a cubic spline is given by the trace of the implicit smoother matrix.

4.2.3 Principal curve clustering algorithm for functional data

As mentioned in Chapter 3, based on Karhunen- Loève (KL) expansion, a stochastic process $Y(t)$ can be expanded as:

$$Y(t) = \mu(t) + \sum_{k=1}^{\infty} \beta_k \phi_k(t), \quad t \in \mathcal{T}.$$

Based on this expansion, we truncate the first N terms of principal components which account for most part of the total variation to approximate $Y(t)$:

$$Y(t) = \mu(t) + \sum_{k=1}^{N} \beta_k \phi_k(t), \quad t \in \mathcal{T},$$

where the cumulative percentage of the overall variation explained by the first N components is given by the ratio $\sum_{k=1}^{N} l_k / \sum_{k=1}^{\infty} l_k$. To achieve a good approximation, we require the cumulative percentage of variation of the first N terms to exceed 95%, in order to determine the appropriate number of components.

In reality, the observations are very often with noise. Hence, the FPCA model of smoothed random curves $\boldsymbol{Y}(t) = \{Y_1(t), \dots, Y_q(t)\}$ can be expressed as:

$$Y_s(t) = \mu(t) + \sum_{k=1}^{N} \beta_{s,k} \phi_k(t) + e_s(t), \quad t \in \mathcal{T}, s = 1,2, \dots, q,$$

where $e_s(t)$ is normally distributed noise term.

This decomposition enables us to collect the set of principal component scores $\beta_s = (\beta_{s,1}, \dots, \beta_{s,N})$ to represent the features of the s_{th} random function, on which the principal curve clustering algorithm can be applied. However, since principal curve clustering is based on extracting and detecting the curvilinear features of point patterns on two-dimensional plane, we will require one more step to transform the principal component scores $\beta_s = (\beta_{s,1}, \dots, \beta_{s,N})$ in $\boldsymbol{R}^N$ into $\alpha_s = (\alpha_{s,1}, \alpha_{s,2})$ in $\boldsymbol{R}^2$ before the algorithm for clustering can be applied. To this end the multidimensional scaling (MDS) will be adopted to realize such transformation. Multidimensional scaling aims to find a projection of given original objects for which one has a distance matrix into 2 or 3 dimensional (Euclidean) space for best visualization (Peng and Müller, 2008). The projected points in 2 or 3 dimensional space represent the original objects in such a way that their distances match with the original distances or dissimilarities, according to some target criterion. With the aid of MDS projection, we are finally able to obtain the 2-dimensional scores $\alpha_s = (\alpha_{s,1}, \alpha_{s,2})$, $s = 1,2, \dots, q$, which can approximately represent the features of the original functional data.

Stanford and Raftery (2000) propose a clustering algorithm based on principal curves for multivariate data, which contains three main steps: denoising, initial clustering and hierarchical principal curve clustering (HPCC). The first step aims to separate the feature points from potential background or feature noise. In the second step, a model-based clustering is used to for initial clustering of the feature points. The third step, hierarchical principal curve clustering, aims to combine potential feature clusters and find the appropriate

number of clusters. Stanford and Raftery (2000) introduce a clustering criterion based on a weighted sum of the squared distances about the curve (V_{about}) and the squared distances along the curve (V_{along}), and also developed a probability model with Bayesian Information Criterion (BIC) for principal curves to overcome the overfitting problem.

The issue with the above algorithm is that, in the final step, it relies on the BIC values for choosing the appropriate number of final clusters as well as the clustering criterion for step-by-step merging, which makes the algorithm very complex. Also, for the clustering criterion, the choice of weights allocated between V_{about} and V_{along} arouses extra difficulty. In this sub-section we develop a different probability model for principal curves and refine the merging criterion to rely on BIC only. Our new algorithm for hierarchical principal curve clustering is described as follows.

Suppose X is a random variable in $\boldsymbol{R}^p$ with a set of observations $\boldsymbol{X} = \{x_1 \dots x_n\}$, and D is a partition consisting of clusters $D_1, \dots D_M$. It is assumed that the feature points are distributed normally about the true underlying feature. At the same time, it is also assumed that the projections of all the feature points spread uniformly along the summation of the lengths of all principal curves. That is, their projections onto the corresponding principal curves form a uniform distribution $U\left(0, \sum v_j\right)$, where v_j is the length of the jth curve. And the orthogonal distances from points to the curve form a normal distribution $N(0, \sigma_j^2)$. The M clusters are then combined in a mixture model and the unconditional probability of a point belonging to the jth feature is denoted by $\pi_j (j = 1, \dots M)$.

Let θ denote the entire set of parameters, then the likelihood is

$$L(\boldsymbol{X}|\theta) = \prod_{i=1}^{N} L(x_i\,|\theta),$$

where $L(x_i|\theta) = \sum_{j=1}^{M} \pi_j\, L(x_i|\theta, x_i \in D_j)$ and $\sum_{j=1}^{M} \pi_j = 1$. For feature clusters,

$$L\left(x_i\middle|\theta, x_i \in D_j\right) = \frac{1}{\sum v_j}\left[\frac{1}{\sqrt{2\pi\sigma_j^2}}\, exp\left(\frac{-\|x_i - f(\lambda_{ij})\|^2}{2\sigma_j^2}\right)\right],$$

where $\|x_i - f(\lambda_{ij})\|$ is the Euclidean distance from the point x_i to its projection $f(\lambda_{ij})$.

Having defined the probability model for the principal curves, we utilize the BIC to determine the optimal number of features and the smoothness (DF) of the curves simultaneously. Under the BIC context, each combination of number of features and degree of freedom is considered as a specific model for the data. And two models compete with each

other through comparing the BIC values. In our case, the BIC for a model with M features is defined as:

$$BIC = 2\log\big(L(\boldsymbol{X}|\theta)\big) - V\log(N),$$

where $V = M(DF + 2) + M - 1$ is the number of parameters. There are totally M features, each involving two parameters ν_j and σ_j, plus the degree of freedom (DF). The mixing proportions π_j contribute to the other $M - 1$ parameters. The larger the BIC is, the more the model is favoured by the data. Conventionally, differences of 2-6 in BIC values show positive evidence for one model, differences of 6-10 release strong signal for one model, while differences greater than 10 indicate very strong evidence (Kass and Raftery, 1995).

We start with the denoised and initially clustered FPC scores (with at least seven points in each cluster, recommended by Stanford and Raftery (2000)), and then look for every possible pair of clusters to check if they can be merged. The merging criterion is based on the BIC value: for each pair of clusters, we are actually facing the choices between a model with two features and one with a single feature. We fit principal curve(s) to the data in both situations, estimate the unknown parameters and finally calculate the BIC values within a reasonable range of DF for one-feature and two-feature models respectively. As has been mentioned above, if the maximum BIC value of one-feature model exceeds that of two-feature model by 2 or more, within the chosen range of DF, there is positive evidence for merging these two clusters. We keep looking for pairs of clusters for all possible mergences until the algorithm stops with no more clusters can be merged. Compared with the original algorithm of Stanford and Raftery (2000), our modified clustering algorithm avoids the complexity of first determining the appropriate number of clusters and then performing step-by-step mergences according to a specific criterion until the desired number of clusters is reached. It can iteratively perform the mergence based on BIC values and automatically arrive at the appropriate number of clusters.

The complete procedure of the functional data clustering algorithm is then summarized as follows.

(1) Smooth the functional data observed at discrete points using appropriate smoothers.

(2) Decompose the smoothed functional data using FPCA and collect the scores $\beta_s = \big(\beta_{s,1}, \dots, \beta_{s,N}\big),\ s = 1,2, \dots, q$ of the first N principal components which make up over 95% of the total variation.

(3) Use the multidimensional scaling (MDS) to transform N dimensional scores $\beta_s = (\beta_{s,1}, \dots, \beta_{s,N})$ into 2-dimensional scores $\alpha_s = (\alpha_{s,1}, \alpha_{s,2}),\ s = 1,2, \dots, q$.

(4) Treat $\alpha_1, \dots, \alpha_q$ as q points in $\boldsymbol{R}^2$. Remove the estimated noise points and make an initial clustering of α_s into several initial clusters.

(5) Seek for all possible pairs of individual clusters. For each pair, based on the probability model, consider it as either one-feature model or two-feature model and calculate the BIC values within a selected range of DF respectively.

(6) Check if the maximum BIC of one-feature model exceeds that of two-feature model by 2 or more. If so, conduct the mergence. Otherwise, leave them as two individual clusters.

(7) Repeat (5) and (6) until no more pair of individual clusters can be merged.

(8) Obtain the finalized clusters of α_s and then convert them to the clusters of the functional data accordingly.

4.3 Simulation study

4.3.1 Case one: Semicircle scores

We simulate the functional data $Y(t)$ using the mean function $\mu(t) = 4\sin(\frac{\pi t}{5})$, the first eigenfunction $\phi_1(t) = -\cos(\frac{\pi t}{10})$, and the second eigenfunction $\phi_2(t) = -\sin(\frac{\pi t}{10})$ within the interval $0 \le t \le 10$. The FPC scores ($\beta_{s,1}$,$\beta_{s,2}$) on 2D plane form two offset semicircles with random Gaussian noise added (the underlying two semicircles follow $\beta_{s,1} = \sqrt{1-(\beta_{s,2}+0.5)^2} - 0.15 + 0.1\varepsilon,\ \beta_{s,1} = -\sqrt{1-(\beta_{s,2}-0.5)^2} + 0.15 + 0.1\varepsilon,\ \varepsilon \sim N(0,1)$ respectively).

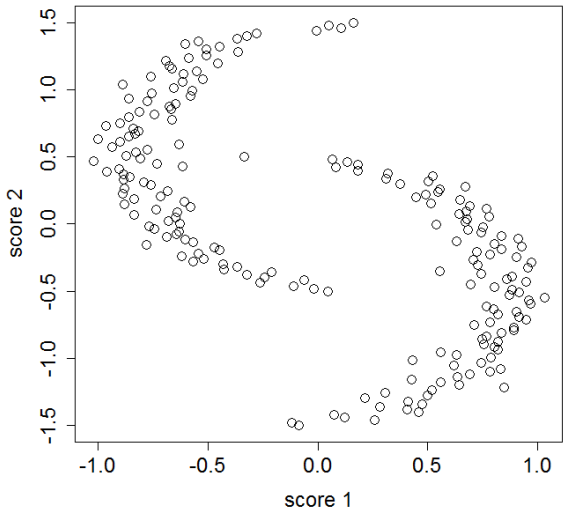


Figure 4.1: Bivariate plot of simulated principal component scores, with x-axis representing the first principal component scores and y-axis representing the second principal component scores.

200 FPC scores are generated (see Figure 4.1) and the trajectories are simulated from $Y_s(t) = \mu(t) + \sum_{k=1}^{2} \beta_{s,k}\phi_k(t)$, $s = 1,2,\ldots,200$, with equally spaced t, as is shown in Figure 4.2.

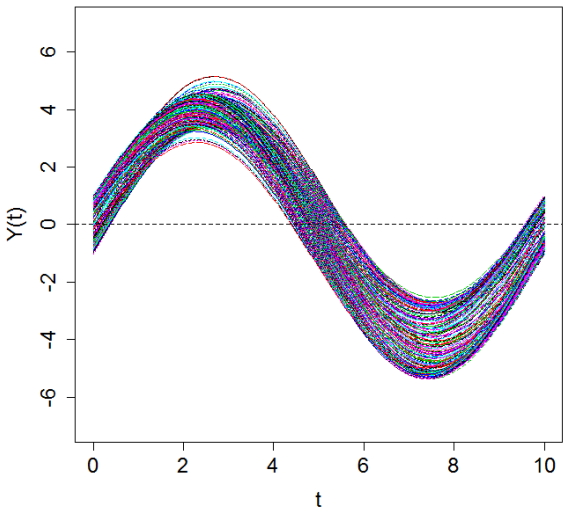


Figure 4.2: 200 random functions simulated from the designated mean function, eigenfunctions and semicircle scores.

The FPCA is then performed and the corresponding principal component scores are obtained, as shown in Figure 4.3. It is noticed that through FPCA, the scale of the original simulated scores are reduced and the axes are rotated to the directions which maximum variance occurs. However, the basic shape of the pattern (two offset semicircles) remains unchanged. As the first two principal components have already explained 100% of the total variance, the multidimensional scaling step is not needed.

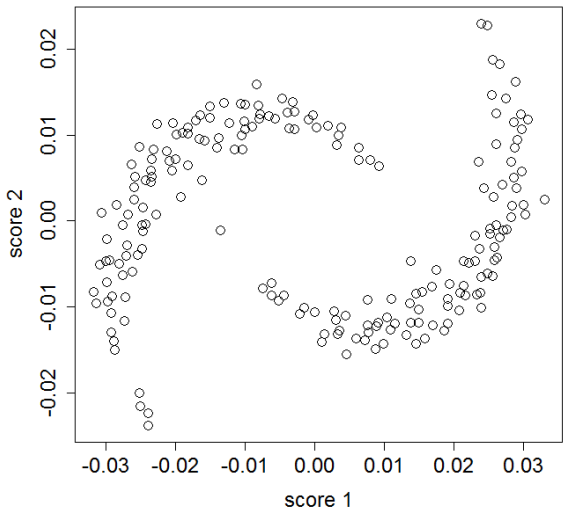


Figure 4.3: Bivariate plot of the first two principal component scores after applying FPCA on the simulated functional data.

The denoising and initial clustering are conducted using K_{th} nearest neighbour clutter removal (KNN) or robust covariance estimation by the nearest neighbour variance estimation (NNVE) and model-based clustering (mclust) and the set of principal component scores is divided into seven initial clusters as illustrated by Figure 4.4. The R package “Prabclus”, “covRobust” and “mclust” contain the implementations of the KNN, NNVE and model-based clustering algorithms and they are available to download from website.

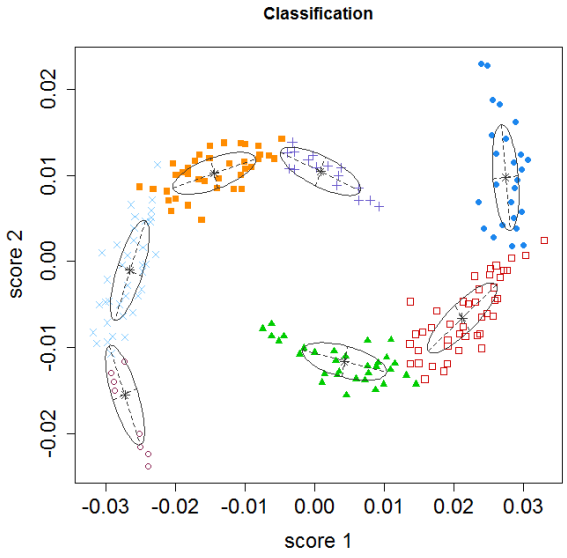
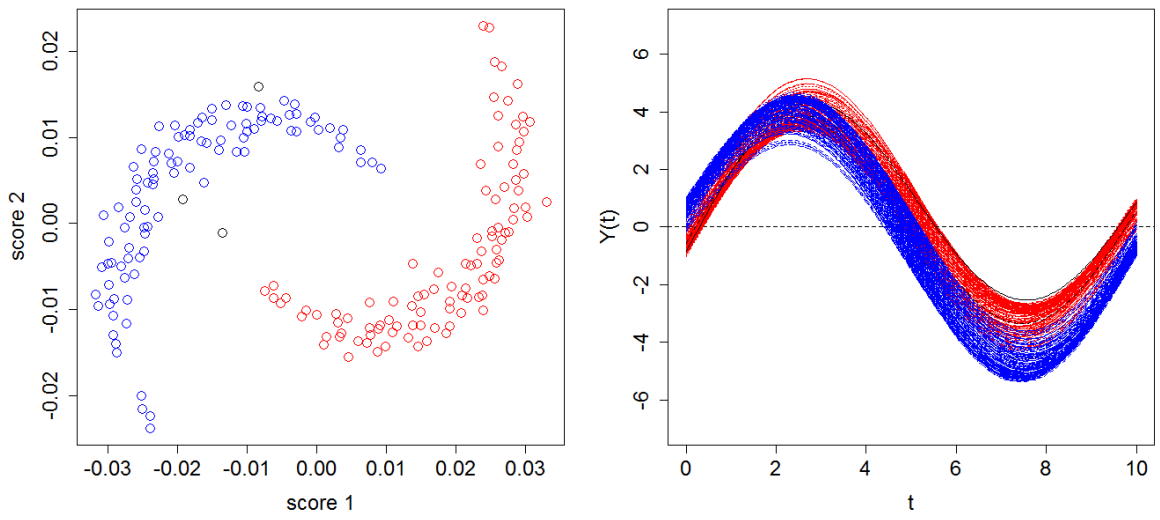


Figure 4.4: Initial clustering of the scores after the removal of potential noise.

Starting with these seven initial clusters, the merging process is performed iteratively for each possible pair of clusters and the BIC values are compared repeatedly to check if that pair of clusters can be merged. Finally, the algorithm stops automatically when two clusters (two semicircles) are recognized, as shown in Figure 4.5(a). And the clustering of the functional data is demonstrated in Figure 4.5(b).



(a) (b)

Figure 4.5: Final clustering of the scores and the functional data. Cluster-one feature is in red and cluster-two feature is in blue. The black stands for noise.

Another issue to be discussed is the range of DF selected for the probability model. We choose the range of DF based on the following rules: the lower bound of DF is always 2; for one-feature model, we set the upper bound of DF as $2/3$ of the total number of points in the single cluster; for two-feature model, we set the upper bound of DF as $2/3$ of the total number of points in the smaller cluster of two. And DF is always capped by 20.

We now compare the performance of our principal curve clustering method with some other functional clustering methods. We choose functional k-means (Peng and Müller 2008) of filtering method and FunHDDC (Bouveyron and Jacques 2011) of adaptive method as our benchmarks. The functional k-means clustering algorithm is to apply the k-means algorithm to FPCA scores. The appropriate number of clusters is determined by the "elbow" in the sum of squared distances scree plot which is shown in Figure 4.6.

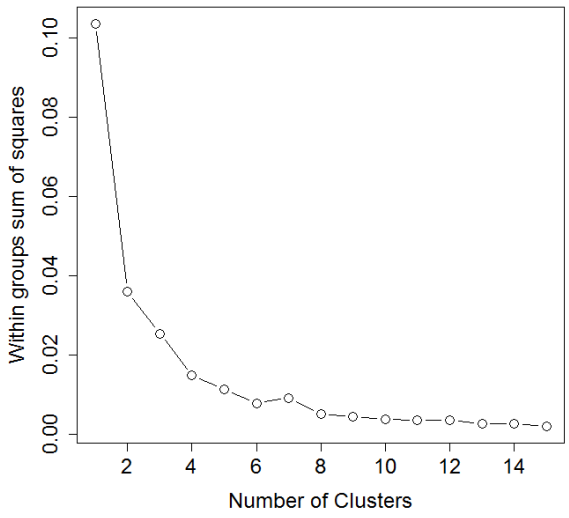


Figure 4.6: The scree plot of the sum of squared distances for different number of clusters by k-means method.

The plot indicates 4 as the appropriate number of clusters. In this way, it fails to recognize the inherent semicircle features, which can be accurately recognized by our principal curve clustering method.

The FunHDDC algorithm is based on a functional latent mixture model and the appropriate number of clusters for a dataset can be determined by the BIC criterion. Figure 4.7 shows the

BIC values by FunHDDC on the simulated functional data set with respect to the total number of clusters. It can be observed that the BIC value increases until $k = 6$ and then stabilizes, which suggests 6 as the optimal number of clusters. However, this betrays the inherent feature of semicircles.

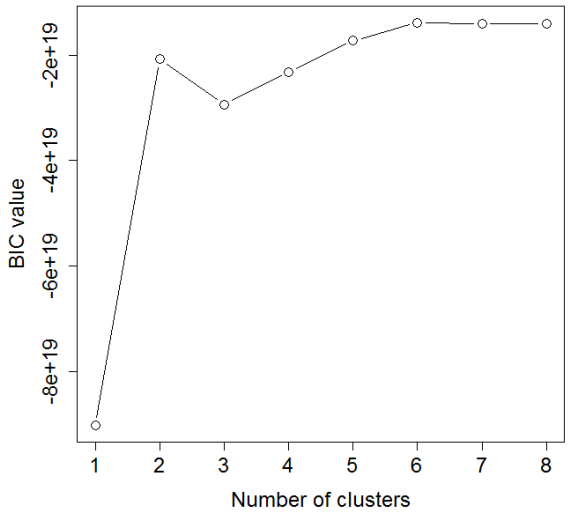


Figure 4.7: The plot of BIC values for different number of clusters by FunHDDC method.

We further investigate the performances of these alternative functional clustering methods by specifying the total number of clusters as 2. Given such condition, k-means method and FunHDDC method are applied to the same set of the simulated functional data and the clustering results are reflected from the corresponding 2D FPCA scores, which are displayed in Figure 4.8. It can be observed that, even given the correct number of clusters, these two methods still cannot produce accurate classification (with the points at the end of semicircles misclassified) while the principal curve clustering method can.

We also consider the curve clustering method developed by Gaffney (2004) using the Curve Clustering Toolbox for MATLAB. This method does not provide a means to determine the number of clusters, so we specify as 2. The clustering result appears to be the same as that by FunHDDC.

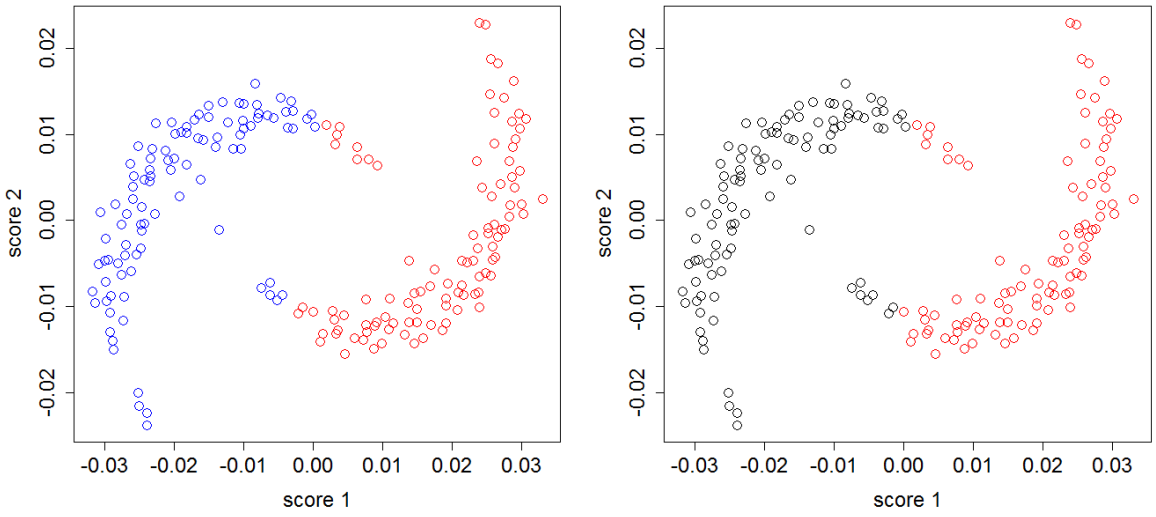


Figure 4.8: Final clustering results reflected by the FPCA scores. The left panel shows the clustering result produced by FunHDDC and the Curve Clustering Toolbox. The right panel shows the clustering result produced by the k-means method.

4.3.2 Case two: Sinusoidal scores

Same as in case one, we specify the mean function $\mu(t) = 4\sin(\frac{\pi t}{5})$, the first eigenfunction $\phi_1(t) = -\cos(\frac{\pi t}{10})$, and the second eigenfunction $\phi_2(t) = -\sin(\frac{\pi t}{10})$ within the interval $0 \leq t \leq 10$. However, this time the FPC scores $(\beta_{s,1},\beta_{s,2})$ on 2D plane are designed to form a sinusoidal pattern with random Gaussian noise added (the underlying sine curve follows $\beta_{s,1} = \sin(\beta_{s,2}) + 0.15\varepsilon,\ \varepsilon \sim N(0,1)$).

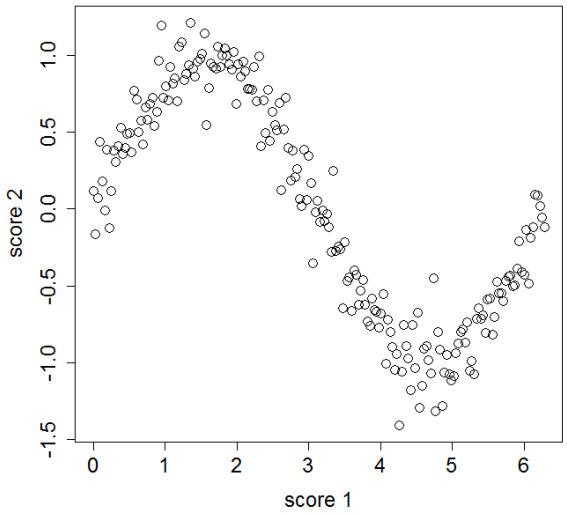


Figure 4.9: Bivariate plot of simulated principal component scores, with x-axis representing the first principal component scores and y-axis representing the second principal component scores.

200 FPC scores are generated (see Figure 4.9) and the trajectories are simulated from $Y_s(t) = \mu(t) + \sum_{k=1}^{2} \beta_{s,k}\phi_k(t)$, $s = 1,2,\ldots,200$, with equally spaced t. See Figure 4.10.

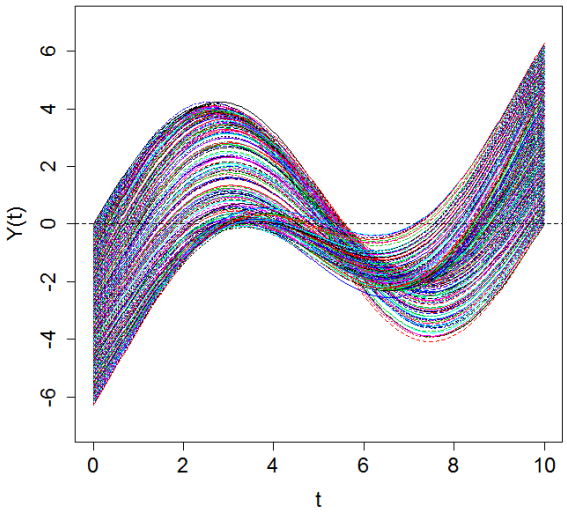


Figure 4.10: 200 random functions simulated from the designated mean function, eigenfunctions and sinusoidal scores.

The FPCA is then performed on the simulated functional data and the corresponding principal component scores are obtained, as shown in Figure 4.11. It is noticed that through FPCA, the scale of the original simulated scores are reduced and the axes are rotated to the directions which maximum variance occurs. However, the basic shape of the pattern (one sinusoidal curve) remains unchanged. As the first two principal components have already explained 100% of the total variance, the multidimensional scaling step is not needed.

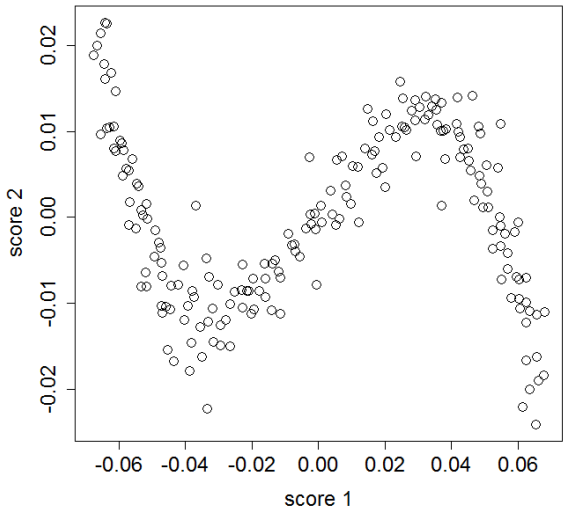


Figure 4.11: Bivariate plot of the first two principal component scores after applying FPCA on the simulated functional data.

The denoising and initial clustering are conducted using K_{th} nearest neighbour clutter removal (KNN) or robust covariance estimation by the nearest neighbour variance estimation (NNVE) and model-based clustering (mclust) and the set of principal component scores is divided into five initial clusters as illustrated by Figure 4.12.

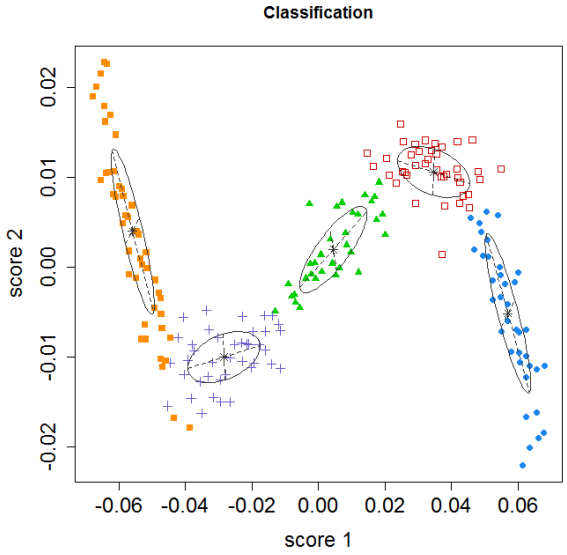
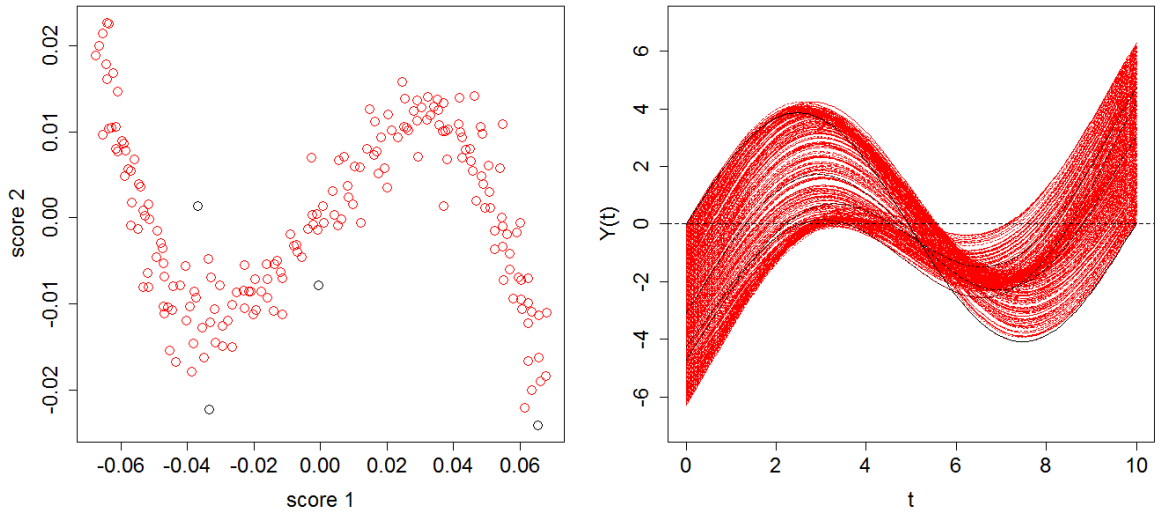


Figure 4.12: Initial clustering of the scores after the removal of potential noise.

Starting with five initial clusters, the merging process is performed iteratively for each possible pair of clusters and the BIC values are compared repeatedly to see if that pair of clusters can be merged. Finally, the algorithm stops automatically when a single cluster (one complete sine pattern) is achieved, as shown in Figure 4.13(a). And the clustering of the functional data is demonstrated in Figure 4.13(b).



(a) (b)

Figure 4.13: Final clustering of the scores and the functional data. One sinusoidal feature (in red) with noise (in black) is identified.

Alternatively, we apply the functional k-means method and the FunHDDC method to the simulated functional data for clustering. For the k-means method, the "elbow" in the sum of squared distances scree plot suggests that 3 clusters is optimal (refer to Figure 4.14). For the FunHDDC method, the BIC value stabilizes at $k = 6$, indicating that 6 clusters is optimal (see Figure 4.15). Obviously, neither the k-means method nor the FunHDDC method can figure out the potential sinusoidal feature while our principal curve clustering method can. This case further illustrates the superiority of our method in finding curvilinear features for clustering purposes.

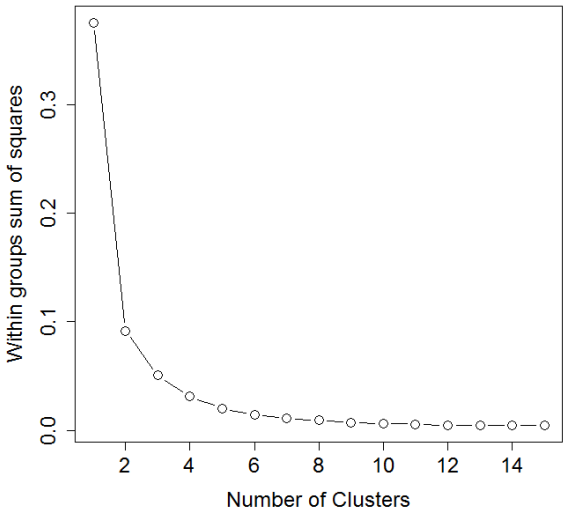


Figure 4.14: The scree plot of the sum of squared distances for different number of clusters by k-means.

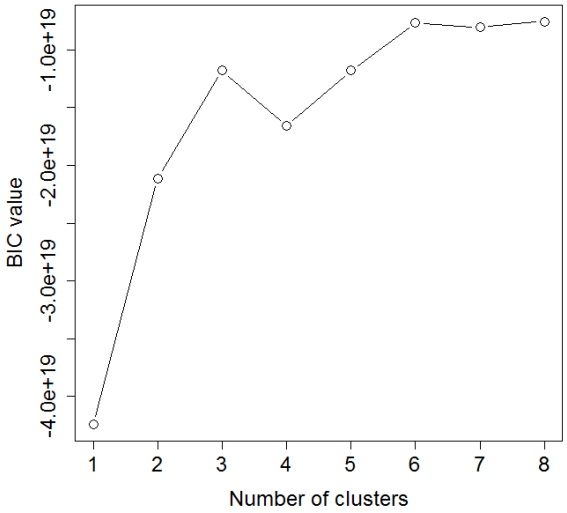


Figure 4.15: The plot of BIC values for different number of clusters by FunHDDC method.

4.4 Empirical study

In this section, we demonstrate the capability of our principal curve clustering approach by applying it to two sets of real functional data. We consider the age-specific mortality data and age-specific fertility data of our interest due to their importance in actuarial science study.

4.4.1 French mortality

We first consider the French total mortality data from 1899 to 2012. The data are obtained from the Human Mortality Database (2012). The age-specific mortality rates are represented in log scale and smoothed data are shown in Figure 4.16.

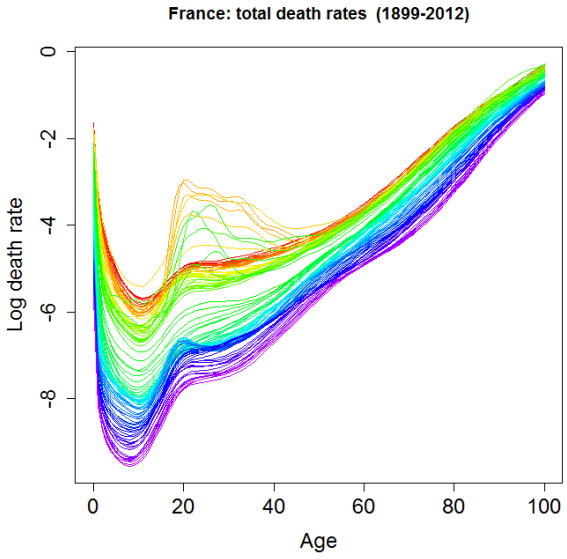


Figure 4.16: Smoothed French log total mortality rates (1899-2012).

We now apply FPCA and decompose the curves into a mean function and a set of functional principal components $\phi_k(x)$ with their corresponding principal component scores $\beta_{s,k}$. The first and second principal components has explained 97.9% and 1.2% of the total variation respectively, which sum up to 99.1% of the total variation. Hence, we consider it to be strongly adequate to take the scores of the first two principal components to represent the

features of the smoothed mortality curves. The components of FPCA decomposition on the French total mortality data are displayed in Figure 4.17.

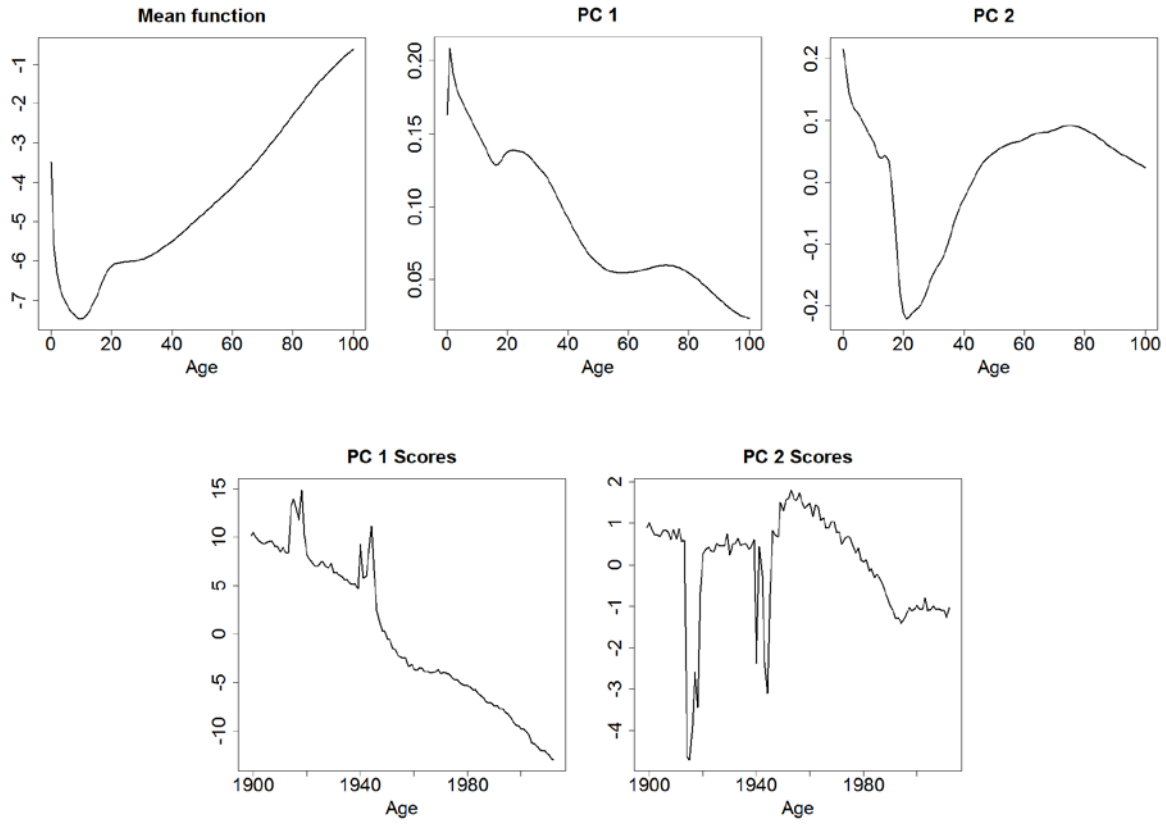


Figure 4.17: Components from FPCA decomposition on the French total mortality data (1899-2012). The mean function, first two functional principal components and their associated scores are displayed.

The bivariate plot of the scores $\beta_s = (\beta_{s,1}\ ,\ \beta_{s,2})$ is shown in Figure 4.18, with the potential noise identified by the Robust Covariance Estimation (NNVE) method.

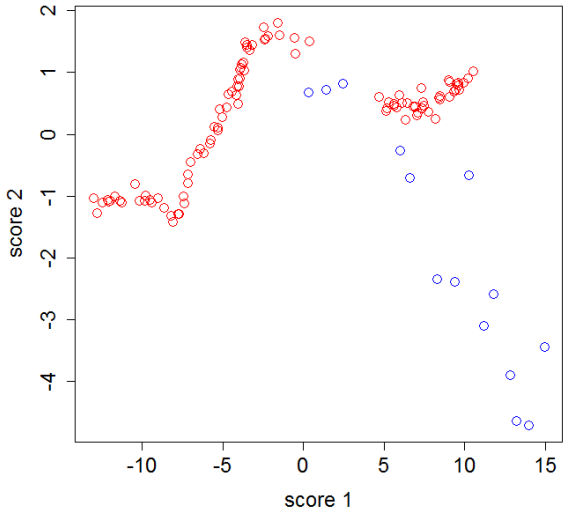


Figure 4.18: Bivariate plot of the first two principal component scores. Red circles indicate feature points while blue circles indicate potential noise.

The initial clustering of the feature points by the model-based clustering mclust suggests that these feature points can be divided into four clusters, as shown in Figure 4.19 below.

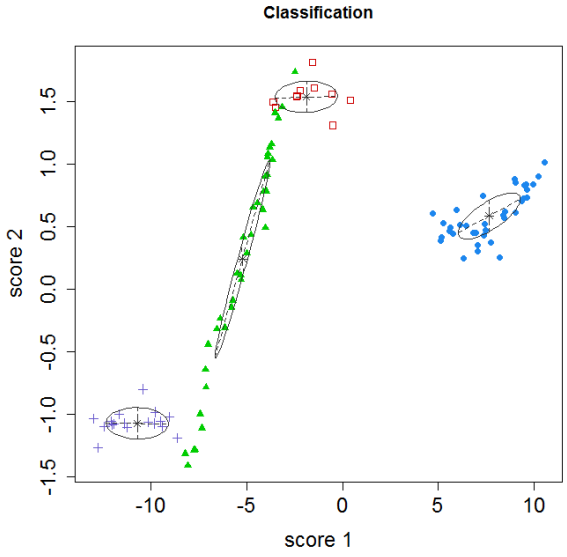


Figure 4.19: Initial clustering of the feature points after removal of potential noise. Four initial clusters are recognized.

Finally, we implement our clustering algorithm on the initial clusters and look for all possible mergences of clusters whenever the merging criterion is met for each iteration. The final

result indicates that the left three clusters can be merged as one, while leaving rightmost cluster as an individual one. It is obvious that the rightmost cluster of scores should be identified as a single cluster. As for the leftmost cluster, we notice that there is a sharp elbow in between it and its neighbouring cluster. Despite the elbow, the maximum BIC value of one-feature model still exceeds that of two-feature model by more than two, within the selected range of DF. However, this is a very marginal situation in deciding whether two clusters should be merged since the difference in BIC values by the two models is very small, and our algorithm can quite sensitively detect it. Hence we are inclined to identify the leftmost cluster as the continuation of the previous feature, rather than considering it as the start of a new feature. The final result generated from the principal curve clustering method is shown in Figure 4.20.

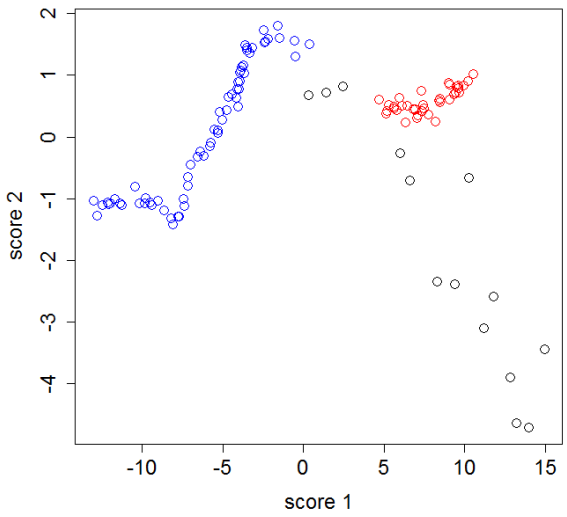


Figure 4.20: Final clustering of the scores. Red and blue circles represent two clusters of feature points. Black circles stand for the noise.

Based on the clustering of the scores $\beta_s = (\beta_{s,1}, \beta_{s,2})$, we are able to convert it to the clustering of the original mortality curves which is shown in Figure 4.21.

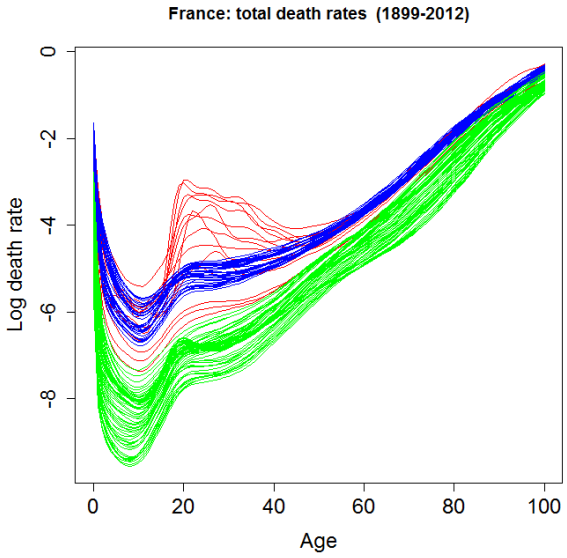


Figure 4.21: Final clustering result of the French total mortality curves from 1899 to 2012. Blue and green curves belong to cluster 1 & 2. Red curves belong to the noise cluster.

In the above figure, the blue curves, referring to the French total mortality curves from 1899 to 1913, 1920 to 1939 and 1941, are classified as cluster one. The green curves, corresponding to those from 1949 to 2012, are classified as cluster two. Looking deeper into the principal components of the mortality curves, there is something interesting to remark. The first principal component actually models the degree of variation of mortality among different age groups. The elder the age group is, the milder the variation in mortality against time it tends to have (see Figure 8). And the decreasing trend of first principal component scores indicates that the mortality rates are evolving lower against time. Meanwhile, the second principal component scores models the differences of mortality between middle age groups and young & old age groups. Apart from these two clusters, the red curves are recognized as anomalies. From observation, we can further distinguish the anomaly curves into two subgroups: the upper group whose mortality rates are extremely high for the middle aged population corresponds to mortality curves of 1914-1919, 1940 and 1942-1945; the lower group situated between cluster one and cluster two refers to mortality curves of 1946-1948. As we know, 1914-1919 is the period of WW1 while 1940-1945 is the period of WW2. And France's participation in both wars explains the extremely high mortality rates of its middle-aged population during those two periods. And the years 1946-1948 is the period right after the ending of WW2 and can be considered as a transition to the after-war booming

period. From historical point of view, the clustering result generated by our functional data clustering method is quite sensible.

4.4.2 Australian fertility

The second example is related to the age-specific Australian fertility data from 1921 to 2002. The data are defined as the number of live births during a calendar year, according to the age of the mother, per 1,000 of the female resident population of the same age at 30 June. The observed Australian fertility rates are smoothed and the result is shown in Figure 4.22.

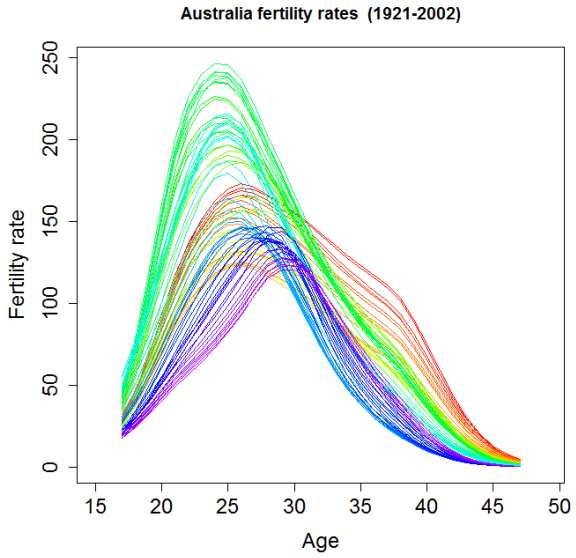


Figure 4.22: Smoothed Australian fertility rates (1921-2002).

The decomposition of the curves by FPCA is shown in Figure 4.23, where the first and second principal components have explained 67.9% and 28.5% of the total variation respectively.

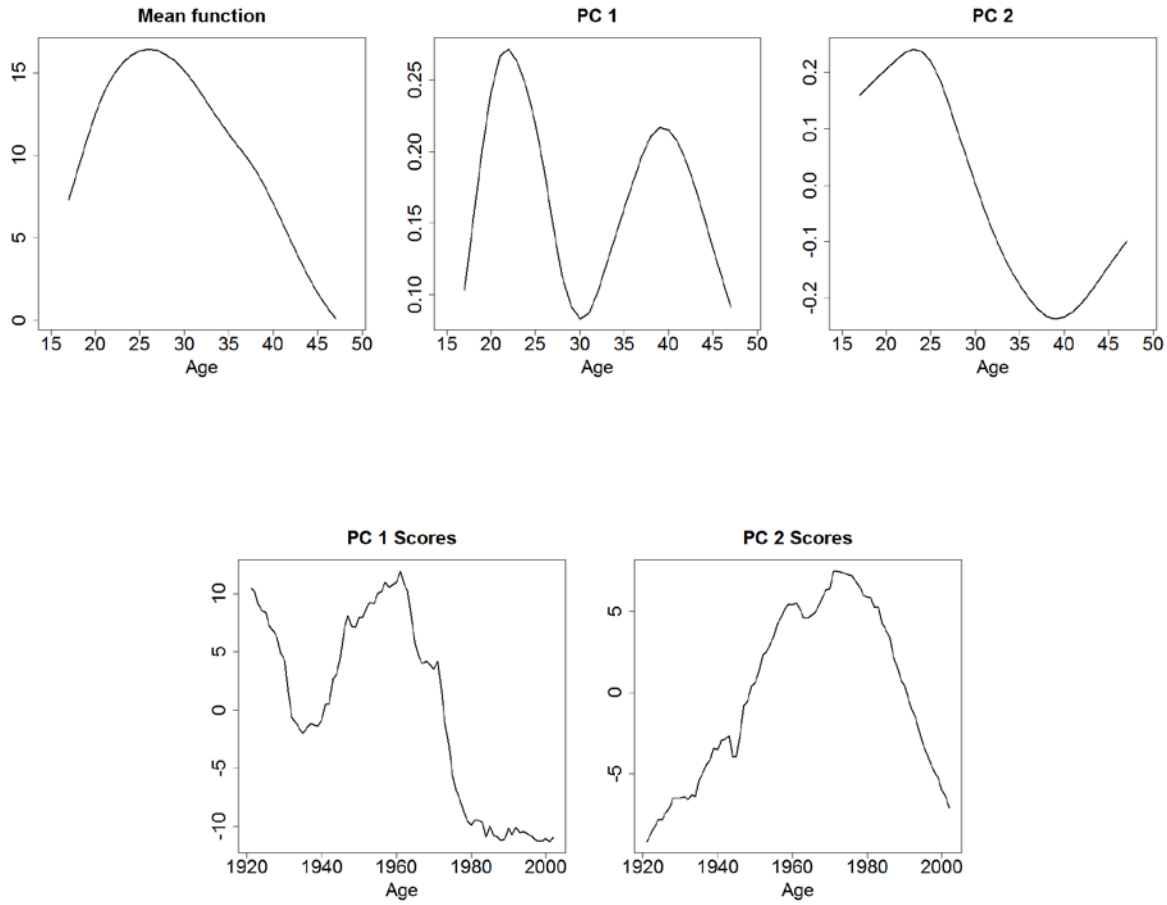


Figure 4.23: Components from FPCA decomposition on the Australian fertility data (1921-2002). The mean function, first two functional principal components and their associated scores are displayed.

The bivariate plot of the associated scores $\beta_s = (\beta_{s,1}\ ,\ \beta_{s,2})$ is displayed in Figure 4.24.

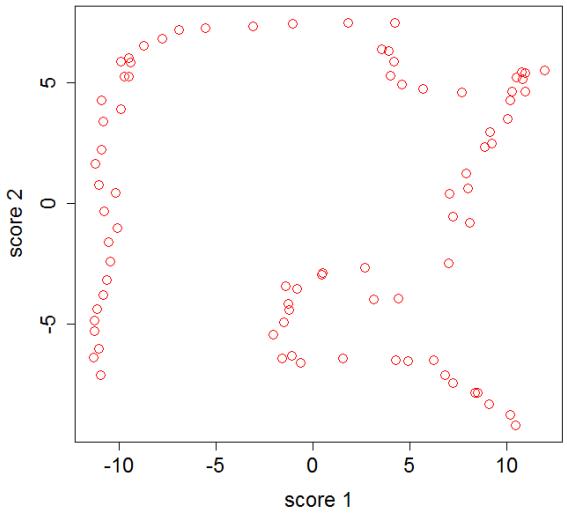


Figure 4.24: Bivariate plot of the first two principal component scores. Red circles indicate feature points. No noise has been detected.

In this example, no noise is detected and all of these 2D points are recognized as feature points. The initial clustering of these feature points is shown in Figure 4.25 below, and the final clustering by our method is shown in Figure 4.26.

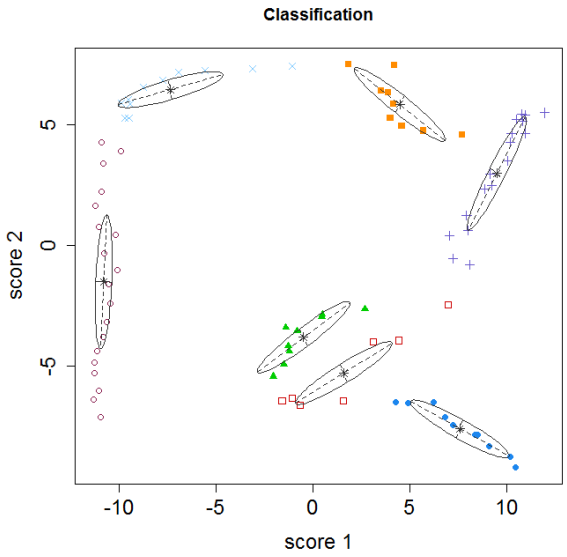


Figure 4.25: Initial clustering of the feature points after removal of potential noise. Seven initial clusters are recognized.

The algorithm suggests that the seven initial clusters can be finally merged into three clusters.

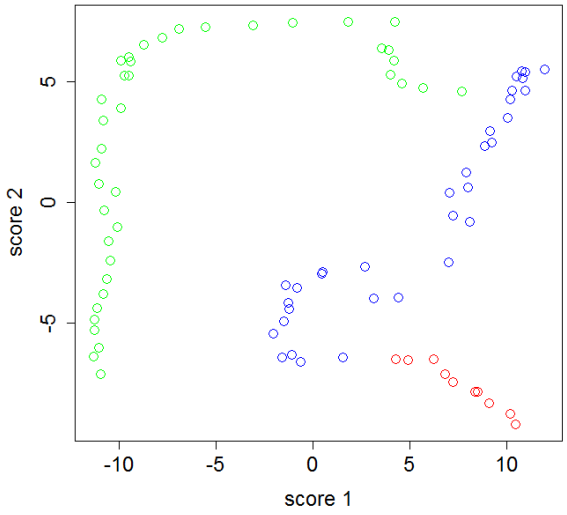


Figure 4.26: Final clustering of the scores. Red, blue and green circles represent three clusters of feature points respectively.

The clustering result of the Australian fertility curves from 1921 to 2002 is displayed in Figure 4.27.

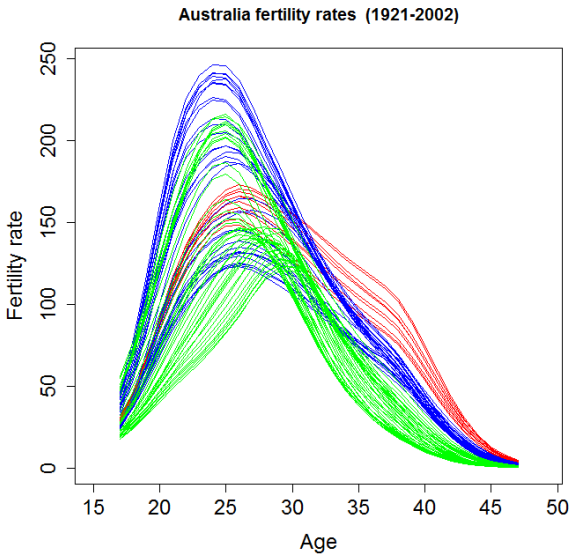


Figure 4.27: Final clustering result of the Australian fertility curves from 1921 to 2002. Red, blue and green curves belong to cluster 1, 2 & 3 respectively.

It is interesting to find that the clustering of the Australian fertility curves actually implies a chronological order: the red curves represent cluster-one Australian fertility curves from 1921 to 1930; the blue curves represent cluster-two curves from 1931 to 1963; the green curves represent the remaining cluster-three curves from 1964 to 2002. The result shows some interesting trends of how the Australian fertility was evolving during the past century. Cluster-one (red) curves demonstrate a generally decreasing trend of fertility over almost all age groups of women from 1921 to 1930. During that period, the peak value of fertility rate is relatively low while the fertility rate at higher ages is quite high. From 1931 to 1963, cluster-two (blue) curves show a rapid increasing trend of fertility over almost all age groups of women. Meanwhile, the age groups that give highest birth rate become younger. From 1964 on to 2002, the overall fertility rate suddenly began to drop drastically with the peak value of fertility rate shifting to higher age groups. And these features are clearly reflected from the cluster-three curves (green). Overall, our principal curve clustering method is able to sensitively detect different patterns of Australian fertility curves and then categorize them into proper clusters.

4.5 Conclusion

In this chapter, we introduced a new method for clustering functional data by principal curves. Our clustering algorithm starts with smoothing the observed discrete data and applying functional principal component analysis to the smoothed functional data for dimension reduction. The scores of the corresponding functional principal components which account for over 95% of the total variation are then collected. If the dimension of the scores exceeds two, the multidimensional scaling is applied to project the high dimensional scores onto 2D plane and the projected scores are used as input for principal curve clustering in the next stage. The algorithm of principal curve clustering consists of three steps: based on the set of 2D scores obtained, potential noise is first removed. Then, an initial clustering is carried out by model-based clustering (mclust). Afterwards, a modified hierarchical principal curve clustering (HPCC) algorithm is applied to construct a principal curve going through each cluster, in order to look for all possible mergences of the initial clusters. We improved

Stanford and Raftery's algorithm by modifying the probability model and designing an iterative method to compare the maximum BIC values of one-feature and two-feature models for all possible mergences of clusters. Finally, the optimal number of clusters is automatically determined and the clustering result of the scores can be converted to the clustering result of the corresponding functional data. In the simulation study, we demonstrated that our principal curve clustering method is capable of identifying the features of functional curves whose associated 2D scores are either semicircle-shaped or sine-shaped. In the empirical study, we found that our principal curve clustering algorithm generated very sensible results in clustering both the French total mortality data and the Australian fertility data.

Chapter 5
Discussions and future work

5.1 Weight chosen for historical data of mean function in GPR model

In chapter 2, we have introduced a new Gaussian process regression (GPR) method and apply it to the modelling and forecasting of age-specific human mortality rates for a single population. The proposed method incorporates a weighted mean function to accurately capture the long-term trend, as well as the spectral mixture covariance function to automatically discover potential patterns, of the mortality rates for specific age groups over time. After forecasts are made for some selected age groups, the mortality rates at other ages for a particular future year can be obtained by interpolating the forecasted mortality rates to all age groups. Compared with Lee-Miller model and the functional data model, our method have presented a more stable and accurate performance in the context of forecasting the French total mortality rates.

One important issue to be discussed is the weight chosen for the historical data in the mean function. When forecasts are made for long horizons, the correlations between the future points and the historical points become very low and the forecast by Gaussian process will converge to the mean function in long term. The weight for the historical data determines the mean function and therefore can also impact the accuracy of forecast. In chapter 2, we used the inverse of the time distance as the common weight for all the age groups assigned to historical data. It is of course possible to use other weights, and if the weights involve tuning parameters, they can be determined by cross validation. An alternative is the geometrically decaying weight raised by Hyndman and Shang (2009) (also mentioned in chapter 3), which is defined as $w_t = \kappa(1-\kappa)^{n-t}$ with $0 < \kappa < 1$. We can regard κ as a tuning parameter and use this weight to calculate the mean function of GPR model. The optimized value of κ for a particular age group can be determined by cross validation. Compared with a fixed common weight for all age groups, this weight with a tuning parameter κ allows flexibility among different age groups, which may lead to an improved performance of the GPR model in forecasting. And this can be the future direction of research on this topic.

5.2 Multivariate time series for modelling level-two scores in the weighted MFPCA model

In chapter 3, we have proposed a new model for coherent forecasting of mortality rates among multiple subpopulations. The model is developed on the basis of multilevel functional principal component analysis (MFPCA) framework and is further modified by incorporating weights to allow more recent data to affect forecasting result more. The mortality curves of different subpopulations are treated as a set of multilevel functional data, and the principal components and their corresponding scores are obtained at two levels and forecasts are made by extrapolating the scores using time series models. Fitting this model to the sex-specific mortality data of nine developed countries, we have demonstrated the effectiveness of the proposed model and the forecasting results suggest that the model outperforms the independent model and is comparable to the Product-Ratio model, in terms of minimizing forecasting errors.

We have discussed several advantages of the weighted MFPCA model in the conclusion part of chapter 3, including a simple and explicit form of the model, flexibility in modelling the age pattern of change, no need to pre-processing the data and explicit calculation of the percentage of variance explained by each principal component. It is also worth noting that, since the level-two scores of different subpopulations all share the same basis functions (principal components), these scores between subpopulations may not be independent. Currently we model the level-two scores using univariate time series in order to reduce model and computation complexity. However, analysing and modelling these level-two scores as joint series over time may better reflect their dynamic relationships and improve the accuracy of forecasts. There have been developed theories and representations for the class of multivariate (vector) time series. It can be more suitable for us to consider applying multivariate time series models such as vector autoregressive (VAR) model or vector autoregressive moving average (VARMA) model with stationary restriction to forecast the level-two scores. And this is left for our future work.

5.3 Improving the functional clustering method by reclassifying the noise and introducing new probability models

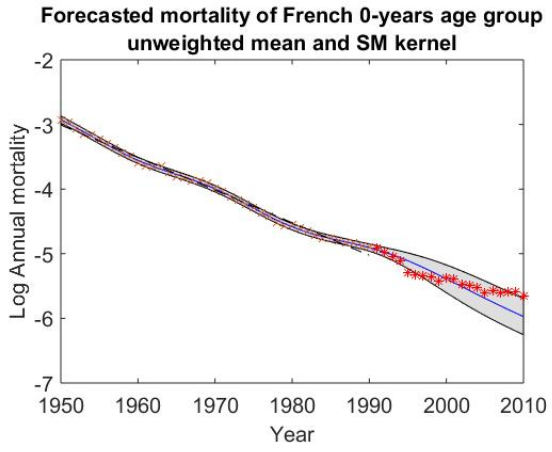
In chapter 4, we have developed an innovative clustering method for functional data, based on the principal curves. Our method makes use of nonparametric principal curves to model the curvilinear features of the two-dimensional scores extracted from the original functional data for clustering purpose. Incorporated in this clustering method is a probability model with Bayesian Information Criterion (BIC) for open principal curves which can automatically detect the appropriate number of features and the optimal degree of smoothing. We have also applied our method to the age-specific French mortality and Australian fertility as functional data for clustering analysis and the results have proved the effectiveness of our method.

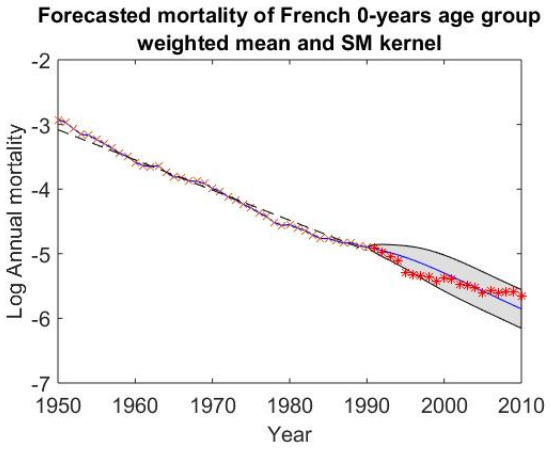
It is noted that in our principal curve clustering algorithm, we conduct denoising step at the beginning to remove potential noises from the true underlying features. And the algorithm will continue without considering these noises any more. However, it should be more appropriate to double-check if these noises can possibly belong to any one of the identified feature clusters. That is to add one more step, at the end of the clustering algorithm, to reclassify these noises and then refine the final clustering results.

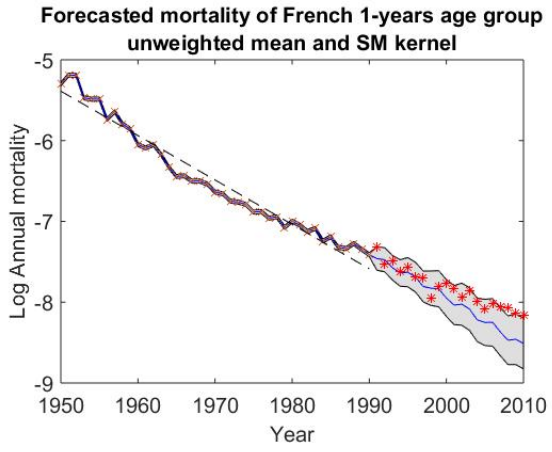
Moreover, our principal curve clustering algorithm actually utilizes open principal curves to identify curvilinear features from two-dimensional principal component scores in order to cluster functional data. However, after decomposing functional data and using multidimensional scaling to obtain two-dimensional scores, the scores may not always present curvilinear features on a 2D plane. Sometimes, these scores may scatter on the plane and display circular features, which open principal curves are unable to capture. In that sense, closed principal curves can be more capable of finding two-dimensional circular features. New probability models and algorithms are needed to tackle such issue.

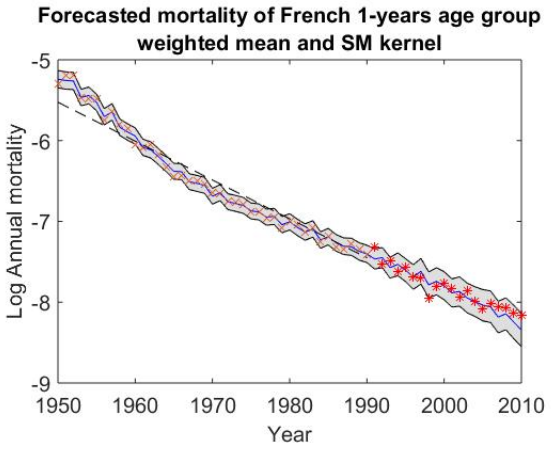
Appendix

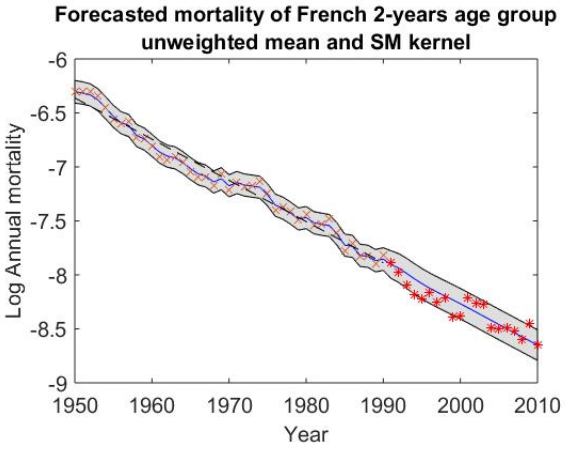
Figure A.1: Figures showing the forecasting results of the 20 age groups using GPR model with and without weighted mean function. All figures displayed on the left are from GPR model with unweighted mean function while those displayed on the right are from GPR model with weighted mean function. The training data are displayed in blue X-mark while the testing data are displayed in red stars. The blue solid line is the predictive mean, with 95% confidence interval indicated by grey shade. The black dashed line represents the mean function of the GP models.

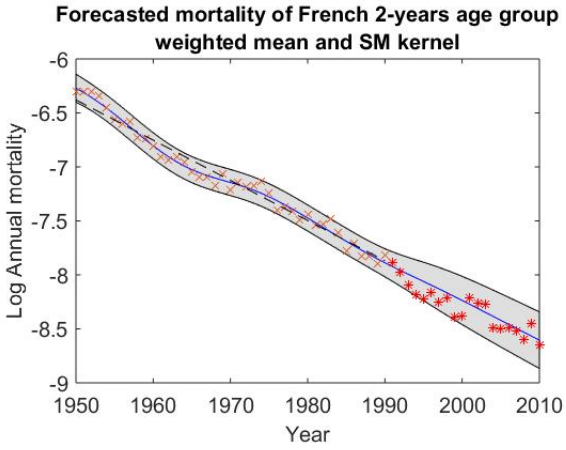


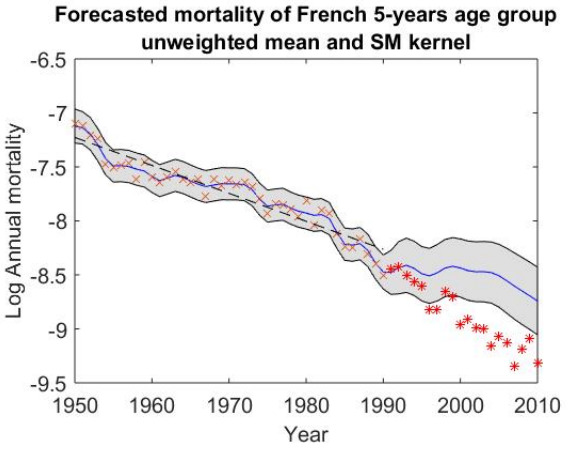


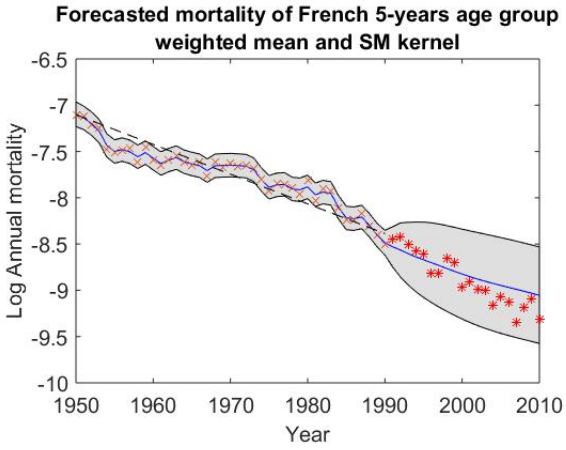


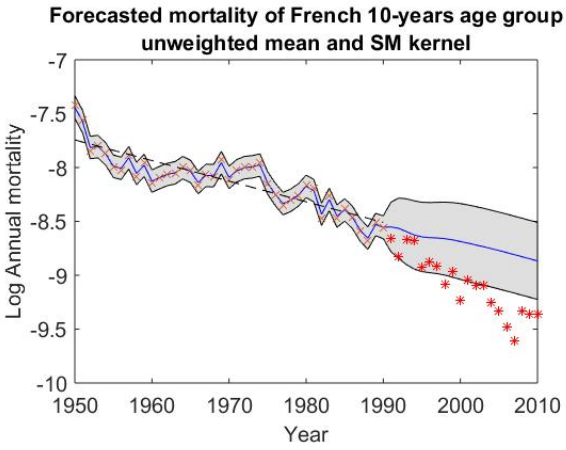


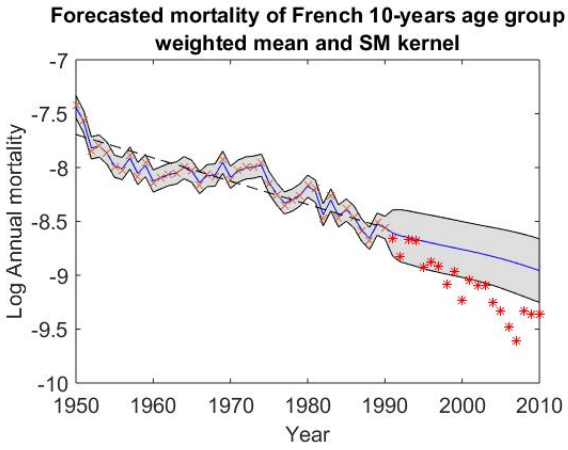

Forecasted mortality of French 2-years age group
unweighted mean and SM kernel
Log Annual mortality
Year
-6
-6.5
-7
-7.5
-8
-8.5
-9
1950
1960
1970
1980
1990
2000
2010

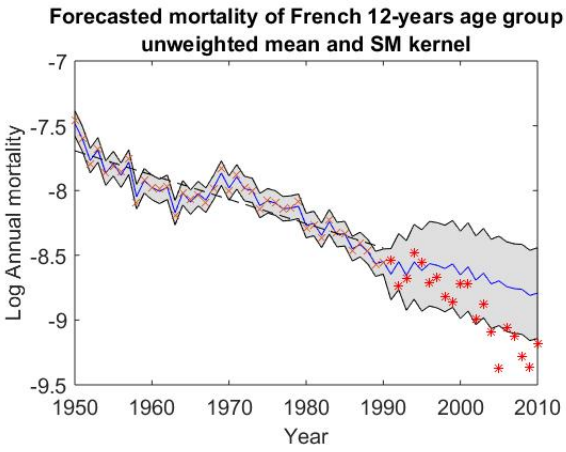

Forecasted mortality of French 2-years age group
weighted mean and SM kernel
Log Annual mortality
Year
-6
-6.5
-7
-7.5
-8
-8.5
-9
1950
1960
1970
1980
1990
2000
2010

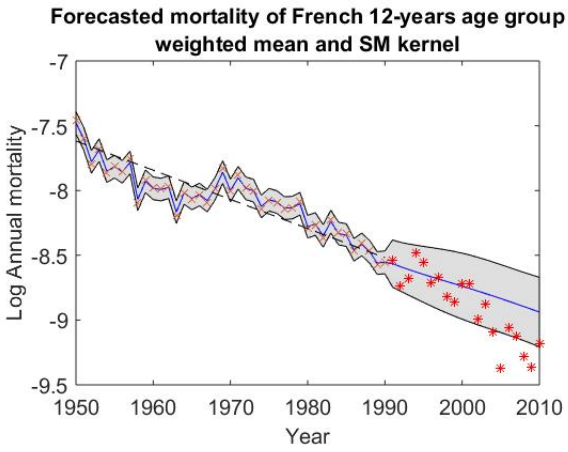

Forecasted mortality of French 5-years age group
unweighted mean and SM kernel
Log Annual mortality
Year
-6.5
-7
-7.5
-8
-8.5
-9
-9.5
1950
1960
1970
1980
1990
2000
2010

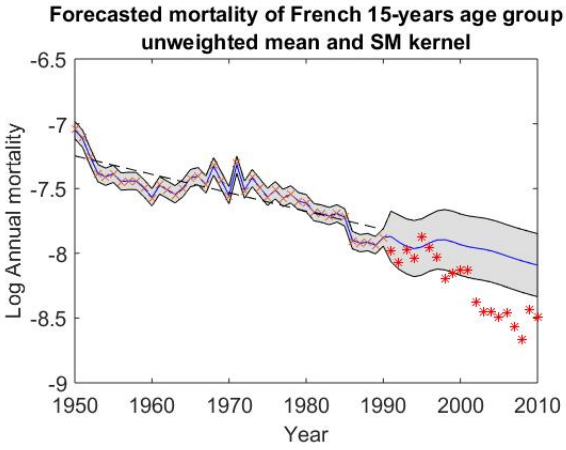

Forecasted mortality of French 5-years age group
weighted mean and SM kernel
Log Annual mortality
Year
-6.5
-7
-7.5
-8
-8.5
-9
-9.5
-10
1950
1960
1970
1980
1990
2000
2010

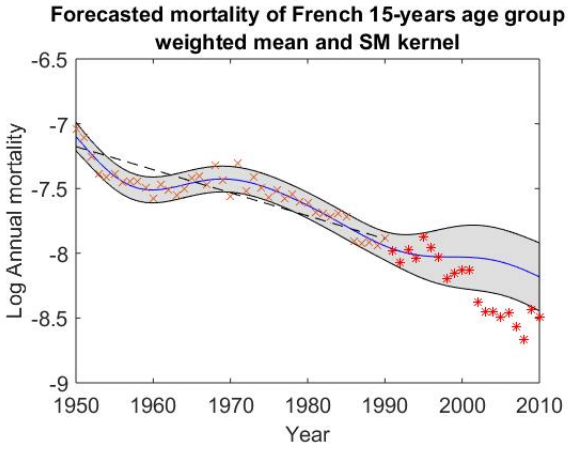

Forecasted mortality of French 10-years age group
unweighted mean and SM kernel
Log Annual mortality
Year
-7
-7.5
-8
-8.5
-9
-9.5
-10
1950
1960
1970
1980
1990
2000
2010

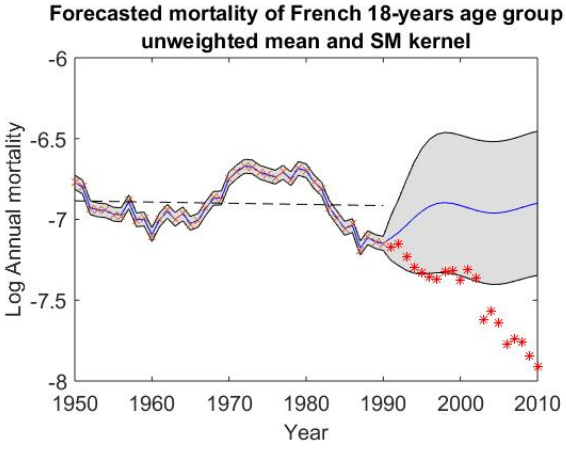

Forecasted mortality of French 10-years age group
weighted mean and SM kernel
Log Annual mortality
Year
-7
-7.5
-8
-8.5
-9
-9.5
-10
1950
1960
1970
1980
1990
2000
2010

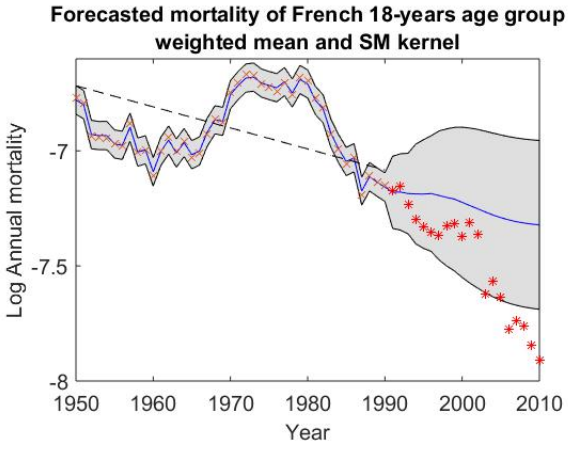

Forecasted mortality of French 12-years age group
unweighted mean and SM kernel
Log Annual mortality
Year
-7
-7.5
-8
-8.5
-9
-9.5
1950
1960
1970
1980
1990
2000
2010

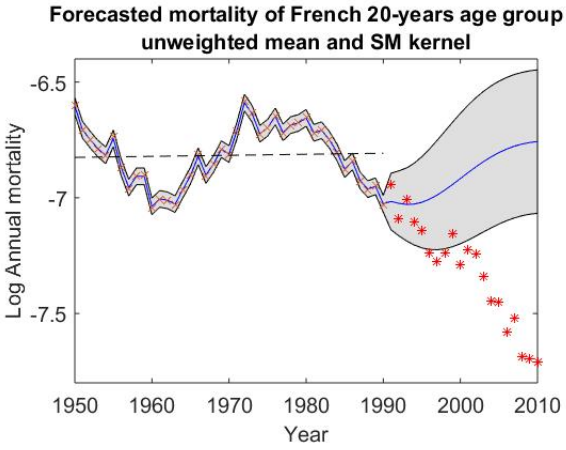

Forecasted mortality of French 12-years age group
weighted mean and SM kernel
Log Annual mortality
Year
-7
-7.5
-8
-8.5
-9
-9.5
1950
1960
1970
1980
1990
2000
2010

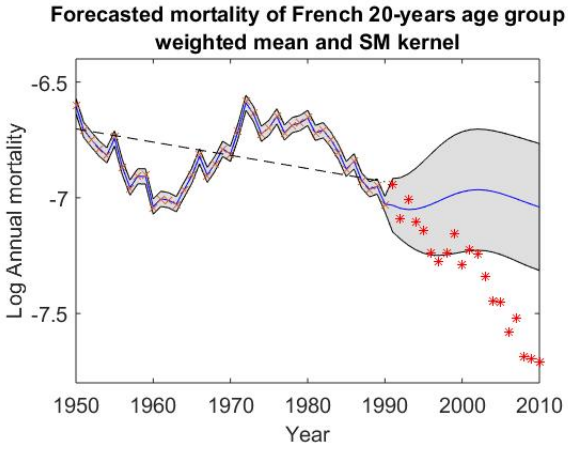

Forecasted mortality of French 15-years age group
unweighted mean and SM kernel
Log Annual mortality
Year
-6.5
-7
-7.5
-8
-8.5
-9
1950
1960
1970
1980
1990
2000
2010

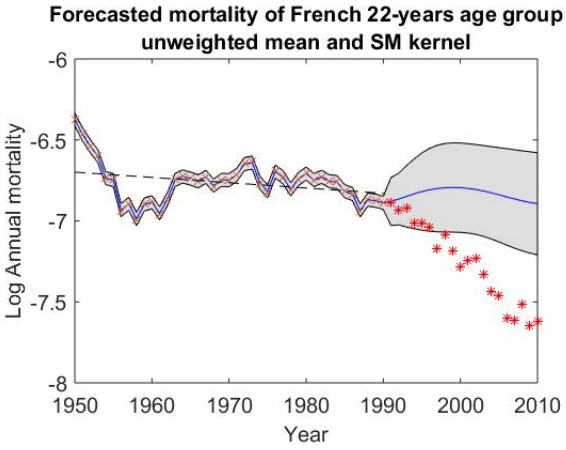

Forecasted mortality of French 15-years age group
weighted mean and SM kernel
Log Annual mortality
Year
-6.5
-7
-7.5
-8
-8.5
-9
1950
1960
1970
1980
1990
2000
2010

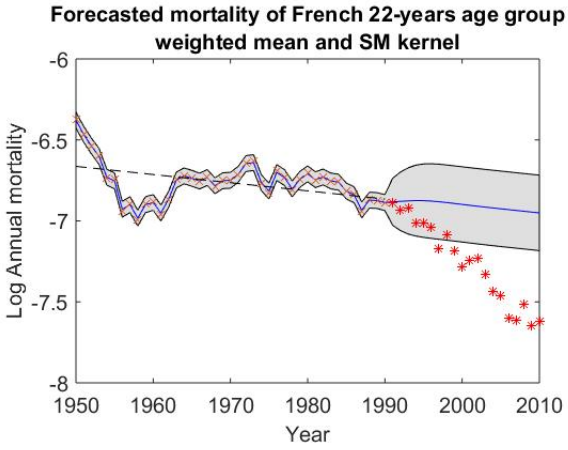

Forecasted mortality of French 18-years age group
unweighted mean and SM kernel
Log Annual mortality
Year
-6
-6.5
-7
-7.5
-8
1950
1960
1970
1980
1990
2000
2010

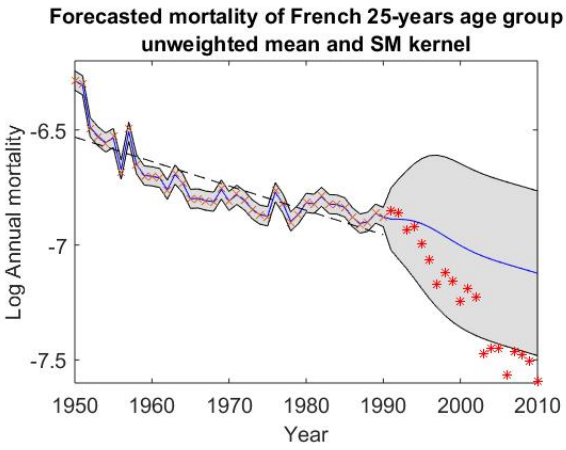

Forecasted mortality of French 18-years age group
weighted mean and SM kernel
Log Annual mortality
Year
-7
-7.5
-8
1950
1960
1970
1980
1990
2000
2010

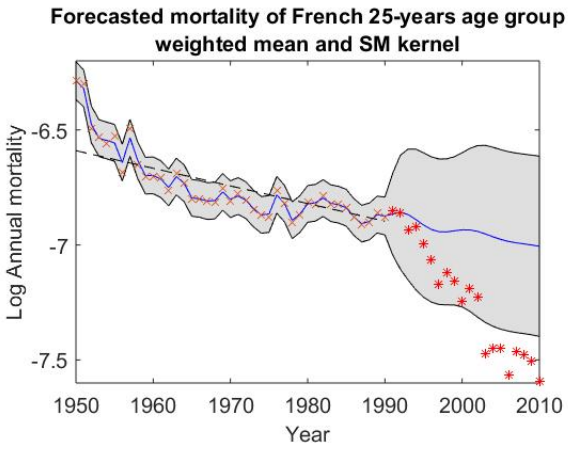

Forecasted mortality of French 20-years age group
unweighted mean and SM kernel
Log Annual mortality
-6.5
-7
-7.5
1950
1960
1970
1980
1990
2000
2010
Year

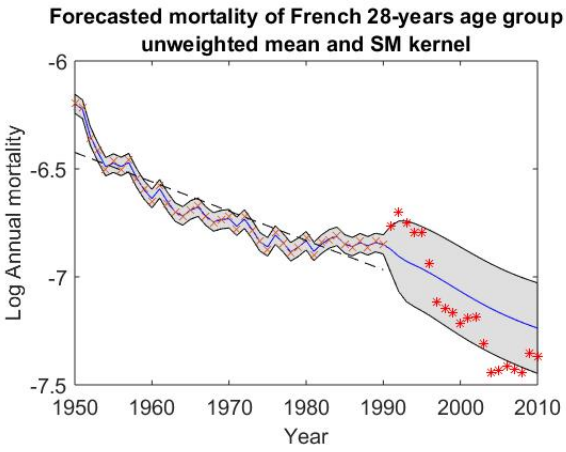

Forecasted mortality of French 20-years age group
weighted mean and SM kernel
Log Annual mortality
-6.5
-7
-7.5
1950
1960
1970
1980
1990
2000
2010
Year

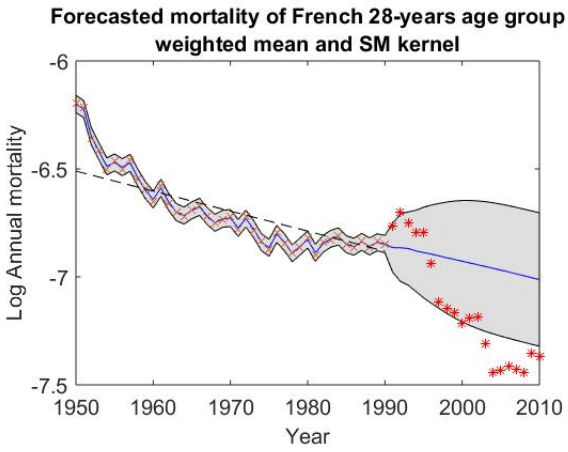

Forecasted mortality of French 22-years age group
unweighted mean and SM kernel
Log Annual mortality
-6
-6.5
-7
-7.5
-8
1950
1960
1970
1980
1990
2000
2010
Year

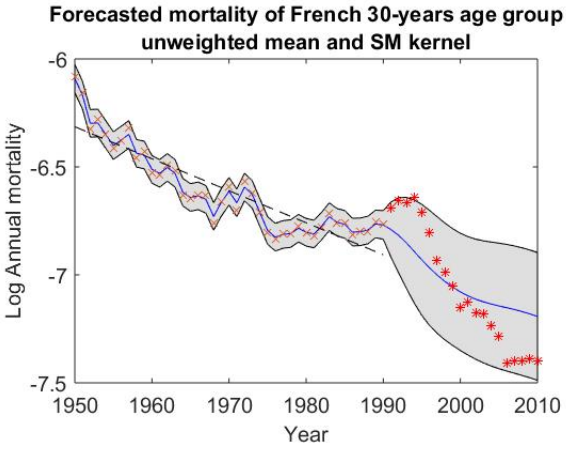

Forecasted mortality of French 22-years age group
weighted mean and SM kernel
Log Annual mortality
-6
-6.5
-7
-7.5
-8
1950
1960
1970
1980
1990
2000
2010
Year

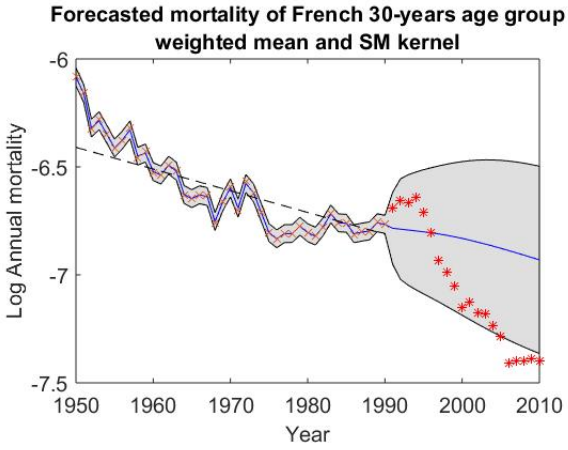

Forecasted mortality of French 25-years age group
unweighted mean and SM kernel
Log Annual mortality
-6.5
-7
-7.5
1950
1960
1970
1980
1990
2000
2010
Year

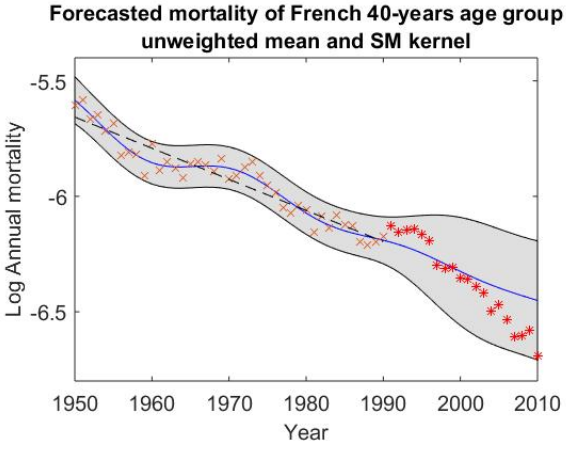

Forecasted mortality of French 25-years age group
weighted mean and SM kernel
Log Annual mortality
-6.5
-7
-7.5
1950
1960
1970
1980
1990
2000
2010
Year

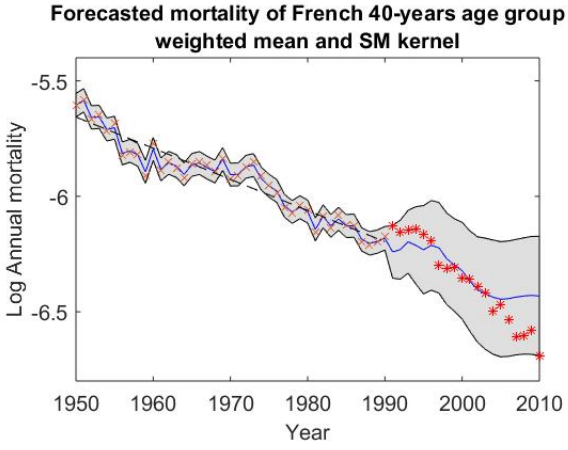

Forecasted mortality of French 28-years age group
unweighted mean and SM kernel
Log Annual mortality
-6
-6.5
-7
-7.5
1950
1960
1970
1980
1990
2000
2010
Year

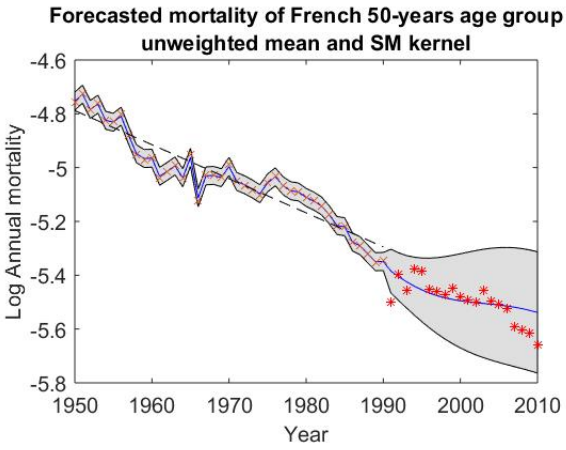

Forecasted mortality of French 28-years age group
weighted mean and SM kernel
Log Annual mortality
-6
-6.5
-7
-7.5
1950
1960
1970
1980
1990
2000
2010
Year

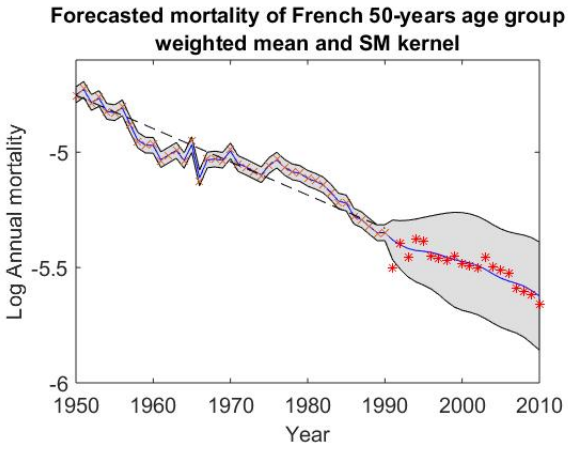

Forecasted mortality of French 30-years age group
unweighted mean and SM kernel
Log Annual mortality
-6
-6.5
-7
-7.5
1950
1960
1970
1980
1990
2000
2010
Year

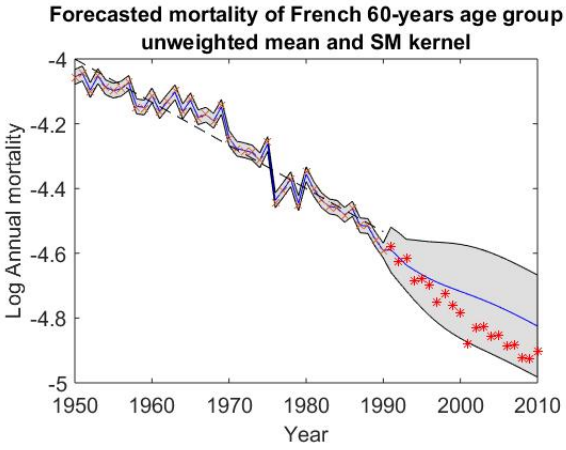

Forecasted mortality of French 30-years age group
weighted mean and SM kernel
Log Annual mortality
-6
-6.5
-7
-7.5
1950
1960
1970
1980
1990
2000
2010
Year

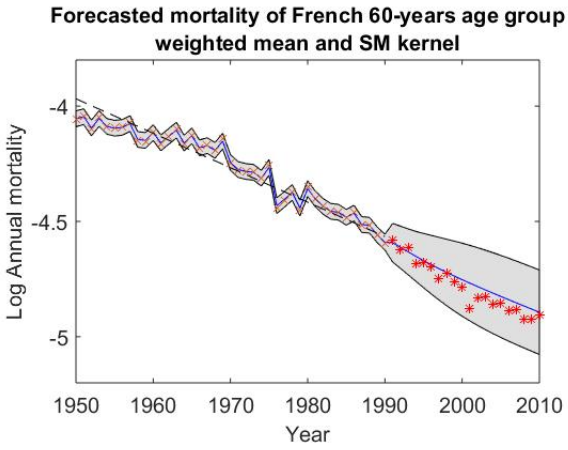

Forecasted mortality of French 40-years age group
unweighted mean and SM kernel
Log Annual mortality
-5.5
-6
-6.5
1950
1960
1970
1980
1990
2000
2010
Year

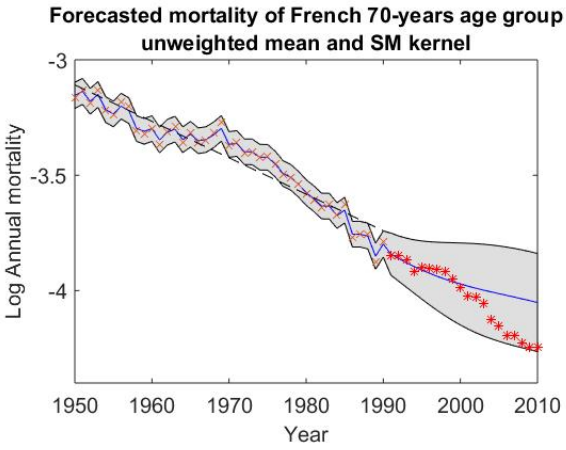

Forecasted mortality of French 40-years age group
weighted mean and SM kernel
Log Annual mortality
-5.5
-6
-6.5
1950
1960
1970
1980
1990
2000
2010
Year

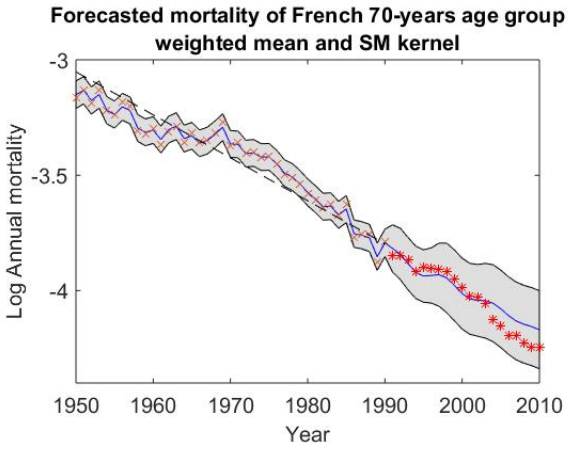

Forecasted mortality of French 50-years age group
unweighted mean and SM kernel
Log Annual mortality
Year
-4.6
-4.8
-5
-5.2
-5.4
-5.6
-5.8
1950
1960
1970
1980
1990
2000
2010

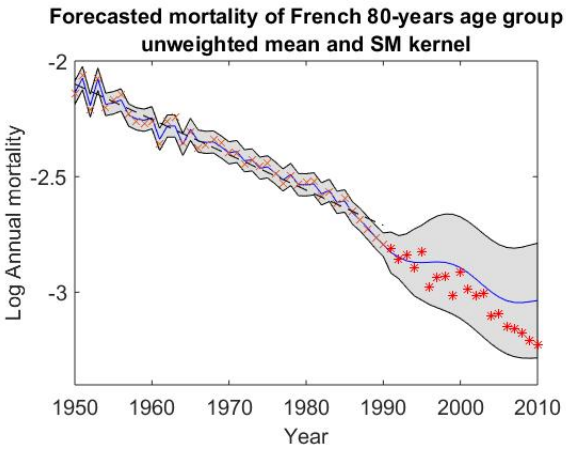

Forecasted mortality of French 50-years age group
weighted mean and SM kernel
Log Annual mortality
Year
-5
-5.5
-6
1950
1960
1970
1980
1990
2000
2010

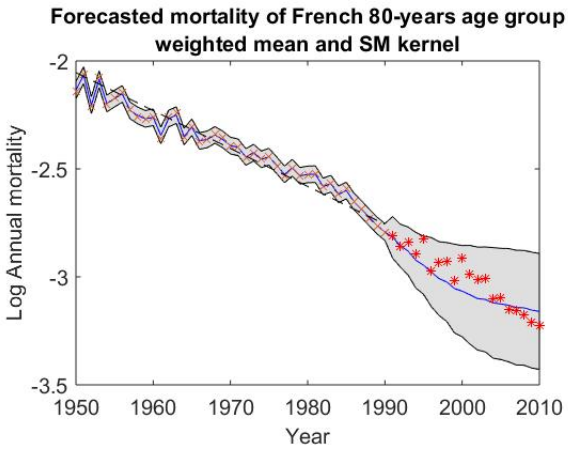

Forecasted mortality of French 60-years age group
unweighted mean and SM kernel
Log Annual mortality
Year
-4
-4.2
-4.4
-4.6
-4.8
-5
1950
1960
1970
1980
1990
2000
2010

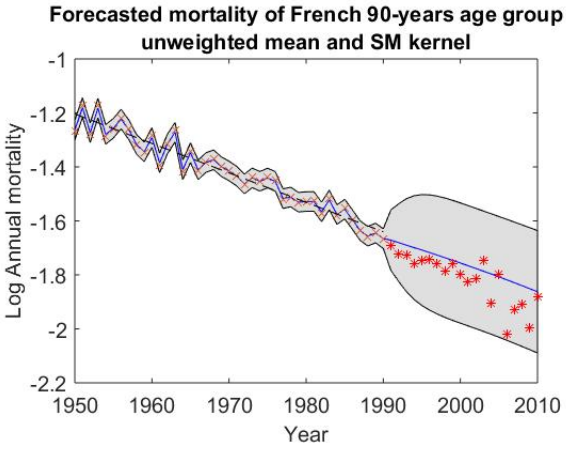

Forecasted mortality of French 60-years age group
weighted mean and SM kernel
Log Annual mortality
Year
-4
-4.5
-5
1950
1960
1970
1980
1990
2000
2010

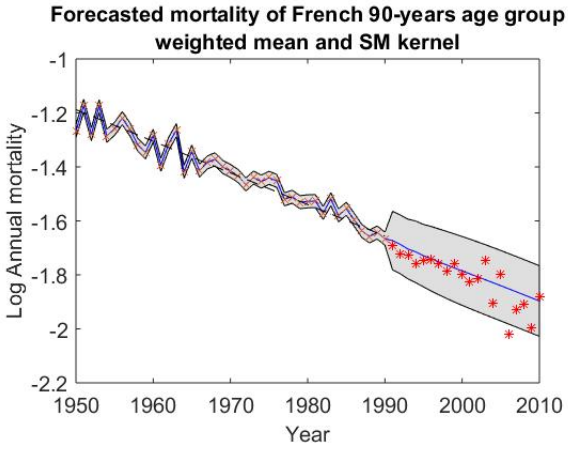

Forecasted mortality of French 70-years age group
unweighted mean and SM kernel
Log Annual mortality
Year
-3
-3.5
-4
1950
1960
1970
1980
1990
2000
2010

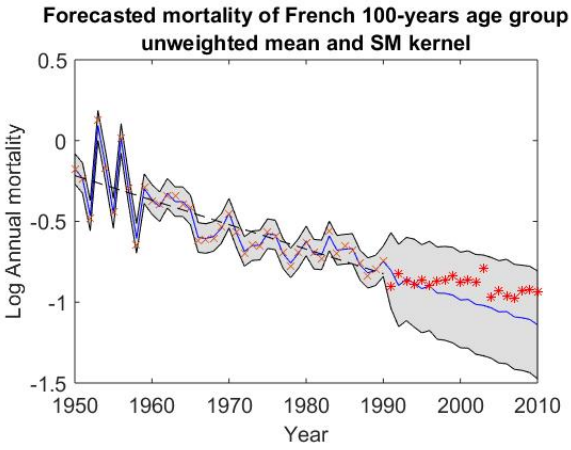

Forecasted mortality of French 70-years age group
weighted mean and SM kernel
Log Annual mortality
Year
-3
-3.5
-4
1950
1960
1970
1980
1990
2000
2010


Forecasted mortality of French 80-years age group
unweighted mean and SM kernel
Log Annual mortality
Year
-2
-2.5
-3
1950
1960
1970
1980
1990
2000
2010


Forecasted mortality of French 80-years age group
weighted mean and SM kernel
Log Annual mortality
Year
-2
-2.5
-3
-3.5
1950
1960
1970
1980
1990
2000
2010


Forecasted mortality of French 90-years age group
unweighted mean and SM kernel
Log Annual mortality
Year
-1
-1.2
-1.4
-1.6
-1.8
-2
-2.2
1950
1960
1970
1980
1990
2000
2010


Forecasted mortality of French 90-years age group
weighted mean and SM kernel
Log Annual mortality
Year
-1
-1.2
-1.4
-1.6
-1.8
-2
-2.2
1950
1960
1970
1980
1990
2000
2010


Forecasted mortality of French 100-years age group
unweighted mean and SM kernel
Log Annual mortality
Year
0.5
0
-0.5
-1
-1.5
1950
1960
1970
1980
1990
2000
2010

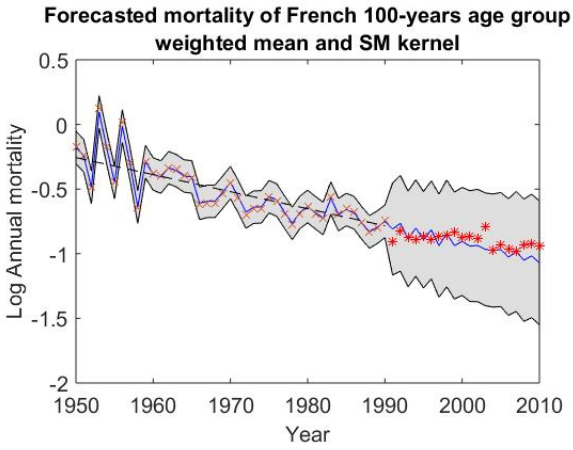

Forecasted mortality of French 100-years age group
weighted mean and SM kernel
Log Annual mortality
Year
0.5
0
-0.5
-1
-1.5
-2
1950
1960
1970
1980
1990
2000
2010

Table A.1: Record of RMSEs using GPR model with SM kernel and unweighted mean function, and GPR model with SM kernel and weighted mean function

Age group	GPR with SM kernel and unweighted mean function	GPR with SM kernel and weighted mean function
0	0.1558	0.1324
1	0.2014	0.0927
2	0.0909	0.0977
5	0.4385	0.1366
10	0.4301	0.3648
12	0.3118	0.2264
15	0.3385	0.2754
18	0.5818	0.2942
20	0.5261	0.3878
22	0.4872	0.4213
25	0.2839	0.3558
28	0.1753	0.3046
30	0.1610	0.3217
40	0.1059	0.1003
50	0.0559	0.0386
60	0.0794	0.0418
70	0.1044	0.0497
80	0.0882	0.0680
90	0.0789	0.0575
100	0.1200	0.0808

Bibliography

Abraham, C., Cornillon, P., Matzner-Løber, E. and Molinari, N. (2003). Unsupervised curve clustering using B-splines. *Scand J Stat Theory Appl 30(3):*581-595.

Banfield, J.D., Raftery, A.E. (1992). Ice Floe identification in satellite images using mathematical morphology and clustering about principal curves. *J. Amer. Statist. Assoc. 87*:7-16.

Bell, W. R. (1997). Comparing and assessing time series methods for forecasting age-specific fertility and mortality rates. *Journal of Official Statistics 13*: 279-303.

Besse, P. (1992). PCA Stability and choice of dimensionality. *Stat. Probab. Lette. 13(5),* 405-410.

Bochner, Salomon. (1959). Lectures on Fourier Integrals. *(AM-42),* volume 42. Princeton University Press.

Booth, J.G., Casella, G., and Hobert, J.P. (2008). Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society, Series B 70,* 119-140.

Booth, H., Maindonald, J., & Smith, L. (2002). Applying Lee-Carter under conditions of variable mortality decline. *Population Studies 56*:325-336.

Bosq, D. (2000). Linear processes in function spaces: Theory and applications. Springer, New York.

Bouveyron, C, Jacques J (2011). Model-based clustering of time series in group-specific functional subspaces. Adv Data Anal Classif 5(4):281-300.

Chiou, J. & Müller, H.G. (2009). Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *Journal of American Statistical Association 104*: 572-585.

Dauxois, J., Pousse, A., and Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *J. Multivariate Anal. 12(1),* 136-154.

Delicado, P. (2001). Another Look at Principal Curves and Surfaces. Journal of Multivariate Analysis 77:84-116.

Di, C. Z., Crainiceanu, C. M., Caffo, B. S., & Punjabi, N. M. (2009). Multilevel functional principal component analysis. *The Annals of Applied Statistics 3(1),* 458-488.

Einbeck, J., Tutz, G., and Evers L. (2005). Local Principal Curves. Statistics and Computing 15: 301-313.

Fraley, C., Raftery, A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association 97,* 611-631.

Ferraty, F., Vieu, P. (2006). Nonparametric functional data analysis: Theory and practice. Springer, New York.

Gaffney, S (2004). Probabilistic curve-aligned clustering and prediction with mixture models. PhD thesis, Department of Computer Science, University of California, Irvine, USA.

Gorban, A.N., Zinovyev, A. (2010). Principal manifolds and graphs in practice: from molecular biology to dynamical systems. *International Journal of Neural System 20 (3),* 219-232.

Hastie, T., Stuetzle, W. (1996). Principal curves. *J. Amer. Statist. Assoc. 84*:502-516.

Hall, P., Müller, H.G., and Wang, J.L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Stat. 34(3),* 1493-1517.

Hall, P., Müller, H.G., & Yao, F. (2008). Modeling sparse generalized longitudinal observations with latent Gaussian processes. *J. R. Stat. Soc. Ser. B Statist. Methodol. 70,* 703-723.

Human Mortality Database. (2010). Univeristy of California, Berkeley (USA), and Max Plank Institute for Demographic Research (Germany). http://www.mortality.org/.

Hyndman, R.J., Booth H., & Yasmeen, F. (2013). Coherent mortality forecasting: The product-ratio method with functional time series models. *Demography, 50,* 261-283.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software, 27(3),* 1-22.

Hyndman, R. J., & Shang, H. L. (2009). Forecasting functional time series. *Journal of the Korean Statistical Society 38,* 199-211.

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

James, G., Sugar, C. (2003). Clustering for sparsely sampled functional data. J. *Am. Stat Assoc., 98(462):*397-408.

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

Kass, R.E., Raftery, A.E. (1995). Bayes Factors. J. Am. Statistical Assoc., vol. 90, pp. 773-795.

Kégl, B., Krzyzak, A., Linder, T., and Zeger, K. (2000). Learning and Design of Principal Curves. IEEE Transactions on Pattern Analysis and Machine Intelligence 24:59-74.

Lee, R. & Carter, L. (1992). Modeling and forecasting the time series of U.S. mortality. *Journal of the American Statistical Association 87*: 659-671.

Lee, R. & Miller, T. (2001). Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography 38*: 537-549.

Li, N., & Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography, 42*, 575-594.

Liu, X. & Yu, H. (2011). Assessing and extending the Lee-Carter model for long-term mortality prediction. Living to 100 Symposium.

Loève, M. (1946). Fonctions aléatoires a decomposition orthognale exponentielle. *La Revue Scientifique 84*, 159-162.

Mas, A. (2002). Weak convergence for the covariance operators of a Hilbertian linear process. *Stoch. Process. Appl. 99(1),* 117-135.

Mas, A. (2008). Local functional principal component analysis. *Complex Anal. Oper. Theory 2(1),* 135-167.

Oeppen, J. (2008). Coherent forecasting of multiple-decrement life tables: A test using Japanese cause of death data (Technical report). Rostock, Germany: Max Planck Institute for Demographic Research.

Peng, J., Müller, H.G. (2008). Distance-based clustering of sparsely observed stochastic processed, with applications to online auctions. *Ann Appl Stat 2(3):*1056-1077.

Ramsay, J.O. & Silverman, B.W. (2005). Functional Data Analysis. Springer.

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

Rasmussen, C. & Williams, C. (2006). Gaussian process for machine learning. MIT Press.

Renshaw, A. & Haberman, S. (2003). Lee-Carter mortality forecasting: A parallel generalized linear modelling approach for England and Wales mortality projections. *Applied Statistics 52(1),* 119-137.

Rossi, F., Conan-Guez, B., and El Golli, A. (2004). Clustering functional data with the SOM algorithm. In: Proceedings of ESANN 2004. Bruges, Belgium, pp 305-312.

Shang, H.L. (2014). A survey of functional data principal component analysis. *AstA Adv Stat Anal 2014 98*:121-142.

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

Shi, J. & Choi, T. (2010). Gaussian process regression analysis for functional data. CRC Press.

Shi, J.Q. & Wang, B. (2008). Curve Prediction and Clustering with Mixtures of Gaussian Process Functional Regression Models. *Statistics and Computing 18,* 267-283.

Stanford, D.C., Raftery, A.E. (2000). Finding curvilinear features in spatial point patterns: Principal curve clustering with noise. *IEEE Transaction on Pattern Analysis and Machine Intelligence, vol.22, No.6.*

Stein, M.L. (1999). Interpolation of spatial data: Some theory for Kriging. Springer Verlag.

Tarpey, T., Kinateder, K. (2003). Clustering functional data. *Journal of Classification 20(1):* 93-114.

Tibshirani, R. (1992). Principal Curves Revisited. Statistics and Computing 2: 183-190.

Tucker, L.R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika 23(1)*, 19-23.

United Nations. (1998). World population prospects: The 1996 revision. New York: Population Division, United Nations.

Wakefield, J., Zhou, C., and Self, S. (2003). Modelling gene expression data over time: Curve clustering with informative prior distributions. *Bayesian Statistics 7*, 721-732.

White, K.M. (2002). Longevity advances in high-income countries, 1955-96. *Population and Development Review 28(1),* 59-76.

Wilson, A.G. & Adams, R.P. (2013). Gaussian process kernels for pattern discovery and extrapolation. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013.

Yao, F., Müller, H.G., and Wang, J.L. (2005). Functional data analysis for sparse longitudinal data. *J. Am. Stat. Assoc. 100(470)*, 577-590.