

Article

# Multivariate Functional Time Series Forecasting: Application to Age-Specific Mortality Rates

Yuan Gao <sup>†</sup> and Han Lin Shang <sup>\*</sup>

Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, ACT 2601, Australia; u5758483@anu.edu.au

<sup>\*</sup> Correspondence: hanlin.shang@anu.edu.au; Tel.: +61-2-6125-0535

<sup>†</sup> Current address: Research School of Finance, Actuarial Studies and Statistics, Level 4, Building 26C, Australian National University, Kingsley Street, Canberra, ACT 2601, Australia.

Academic Editor: Pavel Shevchenko

Received: 26 October 2016; Accepted: 21 March 2017; Published: 25 March 2017

**Abstract:** This study considers the forecasting of mortality rates in multiple populations. We propose a model that combines mortality forecasting and functional data analysis (FDA). Under the FDA framework, the mortality curve of each year is assumed to be a smooth function of age. As with most of the functional time series forecasting models, we rely on functional principal component analysis (FPCA) for dimension reduction and further choose a vector error correction model (VECM) to jointly forecast mortality rates in multiple populations. This model incorporates the merits of existing models in that it excludes some of the inherent randomness with the nonparametric smoothing from FDA, and also utilizes the correlation structures between the populations with the use of VECM in mortality models. A nonparametric bootstrap method is also introduced to construct interval forecasts. The usefulness of this model is demonstrated through a series of simulation studies and applications to the age-and sex-specific mortality rates in Switzerland and the Czech Republic. The point forecast errors of several forecasting methods are compared and interval scores are used to evaluate and compare the interval forecasts. Our model provides improved forecast accuracy in most cases.

**Keywords:** age-and sex-specific mortality rate; bootstrapping prediction interval; vector autoregressive model; vector error correction model; interval score

---

## 1. Introduction

Most countries around the world have seen steady decreases in mortality rates in recent years, which also come with aging populations. Policy makers from both insurance companies and government departments seek more accurate modeling and forecasting of the mortality rates. The renowned Lee–Carter model [1] is a benchmark in mortality modeling. Their model was the first to decompose mortality rates into one component, age, and the other component, time, using singular value decomposition. Since then, many extensions have been made based on the Lee–Carter model. For instance, Booth et al. [2] address the non-linearity problem in the time component. Koissi et al. [3] propose a bootstrapped confidence interval for forecasts. Renshaw and Haberman [4] introduce the age-period-cohort model that incorporates the cohort effect in mortality modeling. Other than the Lee–Carter model, Cairns et al. [5] propose the Cairns–Blake–Dowd (CBD) model that satisfies the new-data-invariant property. Chan et al. [6] use a vector autoregressive integrated moving average (VARIMA) model for the joint forecast of CBD model parameters.

Mortality trends in two or more populations may be correlated, especially between sub-populations in a given population, such as females and males. This calls for a model that makes predictions in several populations simultaneously. We would also expect that the forecasts

of similar populations do not diverge over the long run, so coherence between forecasts is a desired property. Carter and Lee [7] examine how mortality rates of female and male populations can be forecast together using only one time-varying component. Li and Lee [8] propose a model with a common factor and a population-specific factor to achieve coherence. Yang and Wang [9] use a vector error correction model (VECM) to model the time-varying factors in multi-populations. Zhou et al. [10] argue that the VECM performs better than the original Lee–Carter and vector autoregressive (VAR) models, and that the assumption of a dominant population is not needed. Danesi et al. [11] compare several multi-population forecasting models and show that the preferred models are those providing a balance between model parsimony and flexibility. These mentioned approaches model mortality rates using raw data without smoothing techniques. In this paper, we propose a model under the functional data analysis (FDA) framework.

In functional data analysis settings (see Ramsay and Silverman [12] for a comprehensive Introduction to FDA), it is assumed that there is an underlying smooth function of age as the mortality rate in each year. Since mortality rates are collected sequentially over time, we use the term functional time series for the data. Let  $y_t(x)$  denote the log of the observed mortality rate of age  $x$  at year  $t$ . Suppose  $f_t(x)$  is a underlying smooth function, where  $x \in \mathcal{I}$  represents the age continuum defined on a finite interval. In practice, we can only observe functional data on a set of grid points and the data are often contaminated by random noise:

$$y_t(x_j) = f_t(x_j) + u_{t,j}, \quad t = 1, \dots, n, \quad j = 1, \dots, p,$$

where  $n$  denotes the number of years and  $p$  denotes the number of discrete data points of age observed for each function. The errors  $\{u_{t,j}\}$  are independent and identically distributed (iid) random variables with mean zero and variances  $\sigma_t^2(x_j)$ . Smoothing techniques are thus needed to obtain each function  $f_t(x)$  from a set of realizations. Among many others, localized least squares and spline-based smoothing are two of the approaches frequently used (see, for example, [13,14]). We are not the first to use the functional data approach to model mortality rates. Hyndman and Ullah [15] propose a model under the FDA framework, which is robust to outlying years. Chiou and Müller [16] introduce a time-varying eigenfunction to address the cohort effect. Hyndman et al. [17] propose a product–ratio model to achieve coherency in the forecasts of multiple populations.

Our proposed method is illustrated in Section 2 and the Appendices. It can be summarized in four steps:

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

Yang and Wang [9] and Zhou et al. [10] also use VECM to model the time-varying factor, namely, the first set of principal component scores. Our model is different in the following three ways. First, the studied object is in an FDA setting. Nonparametric smoothing techniques are used to eliminate extraneous variations or noise in the observed data. Second, as with other Lee–Carter based models, only the first set of principal component scores are used for prediction in [9,10]. For most countries, the fraction of variance explained is not high enough for one time-varying factor to adequately explain the mortality change. Our approach uses more than one set of principal component scores, and we review some of the ways to choose the optimal number of principal component scores. Third, in their previous papers, only point forecasts are calculated, while we use a bootstrap algorithm for constructing interval forecasts. Point and interval forecast accuracies are both considered.

The article is organized as follows: in Section 2, we revisit the existing functional time series models and put forward a new functional time series method using a VECM. In Section 3, we illustrate how the forecast results are evaluated. Simulation experiments are shown in Section 4. In Section 5, real data analyses are conducted using age-and sex-specific mortality rates in Switzerland and the Czech Republic. Concluding remarks are given in Section 6, along with reflections on how the methods presented here can be further extended.

## 2. Forecasting Models

Let us consider the simultaneous prediction of multivariate functional time series. Consider two populations as an example:  $f_t^{(\omega)}(x)$ ,  $\omega = 1, 2$  are the smoothed log mortality rates of each population. According to (A1) in the Appendices, for a sequence of functional time series  $\{f_t^{(\omega)}(x)\}$ , each element can be decomposed as:

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

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

The functional time series curves are characterized by the corresponding principal component scores that form a time series of vectors with the dimension  $K^{(\omega)}$ :  $\zeta_t^{(\omega)} = (\zeta_{t,1}^{(\omega)}, \dots, \zeta_{t,K^{(\omega)}}^{(\omega)})^\top$ . To construct  $h$ -step-ahead predictions  $\hat{f}_{n+h|n}^{(\omega)}$  of the curve, we need to construct predictions for the  $K^{(\omega)}$ -dimension vectors of the principal component scores; namely,  $\hat{\zeta}_{n+h|n}^{(\omega)} = (\hat{\zeta}_{(n+h|n),1}^{(\omega)}, \dots, \hat{\zeta}_{(n+h|n),K^{(\omega)}}^{(\omega)})^\top$ , with techniques from multivariate time series using covariance structures between multiple populations (see also [18]). The  $h$ -step-ahead prediction for  $f_{n+h|n}^{(\omega)}$  can then be constructed by forward projection

$$\begin{aligned} \hat{f}_{n+h|n}^{(\omega)} &= E \left[ f_{n+h}^{(\omega)} | f_1^{(\omega)}(x), \dots, f_n^{(\omega)}(x) \right] \\ &= \hat{\mu}^{(\omega)}(x) + \hat{\zeta}_{(n+h|n),1}^{(\omega)} \hat{\phi}_1^{(\omega)}(x) + \dots + \hat{\zeta}_{(n+h|n),K^{(\omega)}}^{(\omega)} \hat{\phi}_{K^{(\omega)}}^{(\omega)}(x), \quad \omega = 1, 2. \end{aligned}$$

In the following material, we consider four methods for modeling and predicting the principal component scores  $\zeta_{n+h}$ , where  $h$  denotes a forecast horizon.

### 2.1. Univariate Autoregressive Integrated Moving Average Model

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

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

where  $B$  denotes the lag operator, and  $w_{t,k}$  is the white noise.  $\Phi(B)$  denotes the autoregressive part and  $\Theta(B)$  denotes the moving average part. The orders  $p, d, q$  can be determined automatically according to either the Akaike information criterion or the Bayesian information criterion value [19]. Then, the maximum likelihood method can be used to estimate the parameters.

This prediction model is efficient in some cases. However, Aue et al. [18] argue that, although the FPC scores have no instantaneous correlation, there may be autocovariance at lags greater than zero.

The following model addresses this problem by using a vector time series model for the prediction of each series of FPC scores.

### 2.2. Vector Autoregressive Model

#### 2.2.1. Model Structure

Now that each function  $f_t^{(\omega)}(x)$  is characterized by a  $K^{(\omega)}$ -dimension vector  $\zeta_t^{(\omega)}$ , we can model the  $\zeta_t^{(\omega)}$ s using a VAR( $p$ ) model:

$$\zeta_t^{(\omega)} = \mathbf{v}^{(\omega)} + \mathbf{A}_1^{(\omega)}\zeta_{t-1}^{(\omega)} + \dots + \mathbf{A}_p^{(\omega)}\zeta_{t-p}^{(\omega)} + \epsilon_t,$$

where  $\mathbf{A}^{(\omega)} = \{\mathbf{A}_1^{(\omega)}, \dots, \mathbf{A}_p^{(\omega)}\}$  are fixed  $K^{(\omega)} \times K^{(\omega)}$  coefficient matrices and  $\{\epsilon_t\}$  form a sequence of iid random  $K^{(\omega)}$ -vectors with a zero mean vector. There are many approaches to estimating the VAR model parameters in [20] including multivariate least squares estimation, Yule–Walker estimation and maximum likelihood estimation.

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

#### 2.2.2. Relationship between the Functional Autoregressive and Vector Autoregressive Models

As mentioned in the Introduction, Bosq [21] proposes functional autoregressive (FAR) models for functional time series data. Although the computations for FAR( $p$ ) models are challenging, if not unfeasible, one exception is FAR(1), which takes the form of:

$$f_t = \Psi(f_{t-1}) + \epsilon_t, \tag{1}$$

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

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

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

With matrix notation, we get  $\zeta_t = \mathbf{B}\zeta_{t-1} + \delta_t$ , for  $t = 2, \dots, n$  where  $\mathbf{B} \in \mathbb{R}^{d \times d}$ . This is a VAR(1) model for the estimated principal component scores. In fact, it can be proved that the two models make asymptotically equivalent predictions [18].

### 2.3. Vector Error Correction Model

The VAR model relies on the assumption of stationarity; however, in many cases, that assumption does not stand. For instance, age-and sex-specific mortality rates over a number of years show persistently varying mean functions. The extension we suggest here uses the VECMs to fit pairs of principal component scores of the two populations. In a VECM, each variable in the vector is non-stationary, but there is some linear combination between the variables that is stationary in the long run. Integrated variables with this property are called co-integrated variables, and the process

involving co-integrated variables is called a co-integration process. For more details on VECMs, consult [20].

### 2.3.1. Fitting a Vector Error Correction Model to Principal Component Scores

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

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

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

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

where  $k = 1, \dots, K$ , and  $\alpha_1, \alpha_2, \gamma_{1,1}, \gamma_{1,2}, \gamma_{2,1}, \gamma_{2,2}$  are the coefficients,  $\epsilon_{t,k}^{(1)}$  and  $\epsilon_{t,k}^{(2)}$  are innovations. Note that further lags of  $\Delta \zeta_{t,k}$ ’s may also be included.

### 2.3.2. Estimation

Let us consider the VECM( $p$ ) without the deterministic term written in a more compact matrix form:

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

where

$$\begin{aligned} \Delta \zeta_k &= [\Delta \zeta_{1,k}, \dots, \Delta \zeta_{t,k}], \\ \zeta_{-1,k} &= [\zeta_{0,k}, \dots, \zeta_{n-1,k}], \\ \Gamma_k &= [\Gamma_{1,k}, \dots, \Gamma_{p-1,k}], \\ \Delta \Psi_k &= [\Delta \Psi_{0,k}, \dots, \Delta \Psi_{n-1,k}] \quad \text{with} \quad \Delta \Psi_{t-1,k} = \begin{bmatrix} \Delta \zeta_{t-1,k} \\ \vdots \\ \Delta \zeta_{t-p+1,k} \end{bmatrix}, \\ \epsilon_k &= [\epsilon_{1,k}, \dots, \epsilon_{t,k}]. \end{aligned}$$

With this simple form, least squares, generalized least squares and maximum likelihood estimation approaches can be applied. The computation of the model with deterministic terms is equally easy, requiring only minor modifications. Moreover, the asymptotic properties of the parameter estimators are essentially unchanged. For further details, refer to [20]. There is a sequence of tests to determine the lag order, such as the likelihood ratio test. Since our purpose is to make predictions, a selection scheme based on minimizing the forecast mean squared error can be considered.

### 2.3.3. Expressing a Vector Error Correction Model in a Vector Autoregressive Form

In a matrix notation, the model in Equation (2) can be written as:

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

or

$$\xi_{t,k} - \xi_{t-1,k} = \alpha \beta^\top \xi_{t-1,k} + \Gamma_1 (\xi_{t-1,k} - \xi_{t-2,k}) + \epsilon_{t,k}, \quad (3)$$

where

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

Rearranging the terms in Equation (3) gives the VAR(2) representation:

$$\xi_{t,k} = (\mathbf{I}_K + \Gamma_1 + \alpha \beta^\top) \xi_{t-1,k} - \Gamma_1 \xi_{t-2,k} + \epsilon_{t,k}.$$

Thus, a VECM(1) can be written in a VAR(2) form. When forecasting the scores, it is quite convenient to write the VECM process in the VAR form. The optimal  $h$ -step-ahead forecast with a minimal mean squared error is given by the conditional expectation.

#### 2.4. Product–Ratio Model

Coherent forecasting refers to non-divergent forecasting for related populations [8]. It aims to maintain certain structural relationships between the forecasts of related populations. When we model two or more populations, joint modeling plays a very important role in terms of achieving coherency. When modeled separately, forecast functions tend to diverge in the long run. The product–ratio model forecasts the population functions by modeling and forecasting the ratio and product of the populations. Coherence is imposed by constraining the forecast ratio function to stationary time series models. Suppose  $f^{(1)}(x)$  and  $f^{(2)}(x)$  are the smoothed functions from the two populations to be modeled together, we compute the products and ratios by:

$$p_t(x) = \sqrt{f_t^{(1)}(x)f_t^{(2)}(x)},$$

$$r_t(x) = \sqrt{f_t^{(1)}(x)/f_t^{(2)}(x)}.$$

The product  $\{p_t(x)\}$  and ratio  $\{r_t(x)\}$  functions are then decomposed using FPCA and the scores can be modeled separately with a stationary autoregressive moving average (ARMA)( $p, q$ ) [22] in the product functions or an autoregressive fractionally integrated moving average (ARFIMA)( $p, d, q$ ) process [23,24] in the ratio functions, respectively. With the  $h$ -step-ahead forecast values for  $\hat{p}_{n+h|n}(x)$  and  $\hat{r}_{n+h|n}(x)$ , the  $h$ -step-ahead forecast values for  $\hat{f}_{n+h|n}^{(1)}(x)$  and  $\hat{f}_{n+h|n}^{(2)}(x)$  can be derived by

$$\hat{f}_{n+h|n}^{(1)}(x) = \hat{p}_{n+h|n}(x)\hat{r}_{n+h|n}(x),$$

$$\hat{f}_{n+h|n}^{(2)}(x) = \hat{p}_{n+h|n}(x)/\hat{r}_{n+h|n}(x).$$

#### 2.5. Bootstrap Prediction Interval

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

The univariate model proposed by [15], discussed in Section 2.1, computes the variance of the predicted function by adding up the variance of each component as well as the estimated error variance. The  $(1 - \alpha) \times 100\%$  prediction interval is then constructed under the assumption of normality, where  $\alpha$  denotes the level of significance. The same approach is used in the product–ratio model; however, when the normality assumption is violated, alternative approaches may be used.

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

- 1) Smooth the functions with  $y_t^{(\omega)}(x_j) = f_t^{(\omega)}(x_j) + u_t^{(\omega)}(x_j)$ ,  $\omega = 1, 2$ , where  $u_t^{(\omega)}$  is the smoothing error with mean zero and estimated variance  $\hat{\sigma}_t^2(x_j)^{(\omega)}$ ,  $j = 1, \dots, p$ .
- 2) Perform FPCA on the smoothed functions  $f_t^{(1)}$  and  $f_t^{(2)}$  separately, and obtain  $K$  pairs of principal component scores  $\xi_{t,k} = (\xi_{t,k}^{(1)}, \xi_{t,k}^{(2)})^\top$ .
- 3) Fit  $K$  VECM models to the principal component scores. From the fitted scores  $\hat{\xi}_{t,k}$ , for  $t = 1, \dots, n$  and  $k = 1, \dots, K$ , obtain the fitted functions  $\hat{f}_t = (\hat{f}_t^{(1)}, \hat{f}_t^{(2)})^\top$ .
- 4) Obtain residuals  $e_t$  from  $e_t = f_t - \hat{f}_t$ .
- 5) Express the estimated VECM from step 3 in its VAR form:  $\xi_{t,k} = \hat{A}_1 \xi_{t-1,k} + \hat{A}_2 \xi_{t-2,k} + \epsilon_{t,k}$ ,  $t = 1, \dots, n$  and  $k = 1, \dots, K$ . Construct  $K$  sets of bootstrap principal component scores time series  $\xi_{t,k}^* = \hat{A}_1 \xi_{t-1,k}^* + \hat{A}_2 \xi_{t-2,k}^* + \epsilon_{t,k}^*$ , where the error term  $\epsilon_{t,k}^*$  is re-sampled with replacement from  $\epsilon_{t,k}$ .
- 6) Refit a VECM with  $\xi_{t,k}^*$  and make  $h$ -step-ahead predictions  $\hat{\xi}_{n+h|n}^*$  and hence a predicted function  $\hat{f}_{n+h|n}^*$ .
- 7) Construct a bootstrapped  $h$ -step-ahead prediction for the function by

$$\hat{f}_{n+h|n}^{**}(x_j) = \hat{f}_{n+h|n}^*(x_j) + e_t^* + u_t^*(x_j),$$

where  $e_t^*$  is a re-sampled version of  $e_t$  from step 4 and  $u_t^*(x_j)$  are generated from a normal distribution with mean 0 and variance  $\sigma_{t,j}^2$ , where  $\sigma_{t,j}^2$  is re-sampled from  $\{\hat{\sigma}_{1,j}^2, \dots, \hat{\sigma}_{n,j}^2\}$  from step 1).

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

Koissi et al. [3] extend the Lee–Carter model with a bootstrap prediction interval. The prediction interval we suggest in this paper is different from their method. First, we work under a functional framework. This means that there is extra uncertainty from the smoothing step. Second, in both approaches, errors caused by dimension reduction are bootstrapped. Third, after dimension reduction, their paper uses an ARIMA(0, 1, 0) model to fit the time-varying component. There is no need to consider forecast uncertainty since the parameters of the time series are fixed. In our approach, parameters are estimated using the data. We adopt similar ideas from the early work of Masarotto [25] for the bootstrap of the autoregression process. This step can also be further extended to a bootstrap-after-bootstrap prediction interval [26]. To summarize, we incorporate three sources of uncertainties in our prediction interval, whereas Koissi et al. [3] only considers one due to the simplicity of the Lee–Carter model.

### 3. Forecast Evaluation

We split the data set into a training set and a testing set. The four models are fitted to the data in the training set and predictions are made. The data in the testing set is then used for forecast evaluation. Following the early work by [27], we allocate the first two-thirds of the observations into the training set and the last one-third into the testing set.

We use an expanding window approach. Suppose the size of the full data set is 60. The first 40 functions are modeled and one to 20-step-ahead forecasts are produced. Then, the first 41 functions are used to make one to 19-step-ahead forecasts. The process is iterated by increasing the sample size by one until reaching the end of the data. This produces 20 one-step-ahead forecasts, 19 two-step-ahead forecasts, . . . and, finally, one 20-step-ahead forecast. The forecast values are compared with the true values of the last 20 functions. Mean absolute prediction errors (MAPE) and mean squared prediction errors (MSPE) are used as measures of point forecast accuracy [11]. For each population, MAPE and MSPE can be calculated as:

$$\begin{aligned} \text{MAPE}(h) &= \frac{1}{(21-h) \times p} \sum_{\eta=h}^{20} \sum_{j=1}^p \left| y_{n+\eta}(x_j) - \widehat{f}_{n+\eta|n+\eta-h}(x_j) \right|, \\ \text{MSPE}(h) &= \frac{1}{(21-h) \times p} \sum_{\eta=h}^{20} \sum_{j=1}^p \left[ y_{n+\eta}(x_j) - \widehat{f}_{n+\eta|n+\eta-h}(x_j) \right]^2, \end{aligned} \tag{4}$$

where  $\widehat{f}_{n+\eta|n+\eta-h}$  represents the  $h$ -step-ahead prediction using the first  $n + \eta - h$  years fitted in the model, and  $y_{n+\eta}(x_j)$  denotes the true value.

For the interval forecast, coverage rate is a commonly used evaluation standard. However, coverage rate alone does not take into account the width of the prediction interval. Instead, the interval score is an appealing method that combines both a measure of the coverage rate and the width of the prediction interval [28]. If  $\widehat{f}_{n+h|n}^u$  and  $\widehat{f}_{n+h|n}^l$  are the upper and lower  $(1 - \alpha) \times 100\%$  prediction bounds, and  $y_{n+h}$  is the realized value, the interval score at point  $x_j$  is:

$$\begin{aligned} S_\alpha(x_j) &= \left[ \widehat{f}_{n+h|n}^u(x_j) - \widehat{f}_{n+h|n}^l(x_j) \right] \\ &+ \frac{2}{\alpha} \left[ \widehat{f}_{n+h|n}^l(x_j) - y_{n+h}(x_j) \right] \mathbb{1} \left\{ y_{n+h}(x_j) < \widehat{f}_{n+h|n}^l(x_j) \right\} \\ &+ \frac{2}{\alpha} \left[ y_{n+h}(x_j) - \widehat{f}_{n+h|n}^u(x_j) \right] \mathbb{1} \left\{ y_{n+h}(x_j) > \widehat{f}_{n+h|n}^u(x_j) \right\}, \end{aligned} \tag{5}$$

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

$$\bar{S}_\alpha(h) = \frac{1}{(21-h) \times p} \sum_{\eta=h}^{20} \sum_{j=1}^p S_\alpha \left[ \widehat{f}_{n+\eta|n+\eta-h}^u(x_j), \widehat{f}_{n+\eta|n+\eta-h}^l(x_j); y_{n+\eta}(x_j) \right], \tag{6}$$

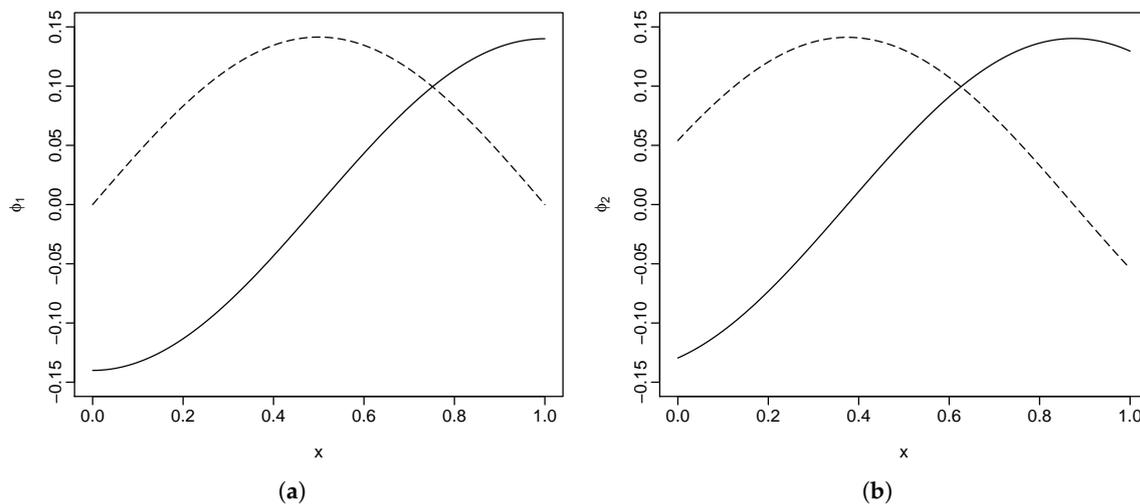
where  $p$  denotes the number of age groups and  $h$  denotes the forecast horizons.

### 4. Simulation Studies

In this section, we report the results from the prediction of simulated non-stationary functional time series using the models discussed in Section 2. We generated two series of correlated populations, each with two orthogonal basis functions. The simulated functions are constructed by

$$f_t^{(\omega)}(x) = \zeta_{t,1}^{(\omega)} \phi_1^{(\omega)}(x) + \zeta_{t,2}^{(\omega)} \phi_2^{(\omega)}(x), \quad \omega = 1, 2. \tag{7}$$

The construction of the basis functions is arbitrary, with the only restriction being that of orthogonality. The two basis functions for the first population we used are  $\phi_1^{(1)}(x) = -\cos(\pi x)$  and  $\phi_2^{(1)} = \sin(\pi x)$ , and, for the second population, these are  $\phi_1^{(2)}(x) = -\cos(\pi x + \pi/8)$  and  $\phi_2^{(2)}(x) = \sin(\pi x + \pi/8)$ , where  $x \in [0, 1]$ . Here, we are using  $n = 100$  discrete data points for each function. As shown in Figure 1, the basis functions are scaled so that they have an  $L_2$  norm of 1.



**Figure 1.** Simulated basis functions for the first and second populations. (a) basis functions for population 1; (b) basis functions for population 2.

The principal component scores, or coefficients  $\zeta_{t,k}$ , are generated with non-stationary time series models and centered to have a mean of zero. In Section 4.1, we consider the case with co-integration, and, in Section 4.2, we consider the case without co-integration.

4.1. With Co-Integration

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

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

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

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

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

4.2. Without Co-Integration

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

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

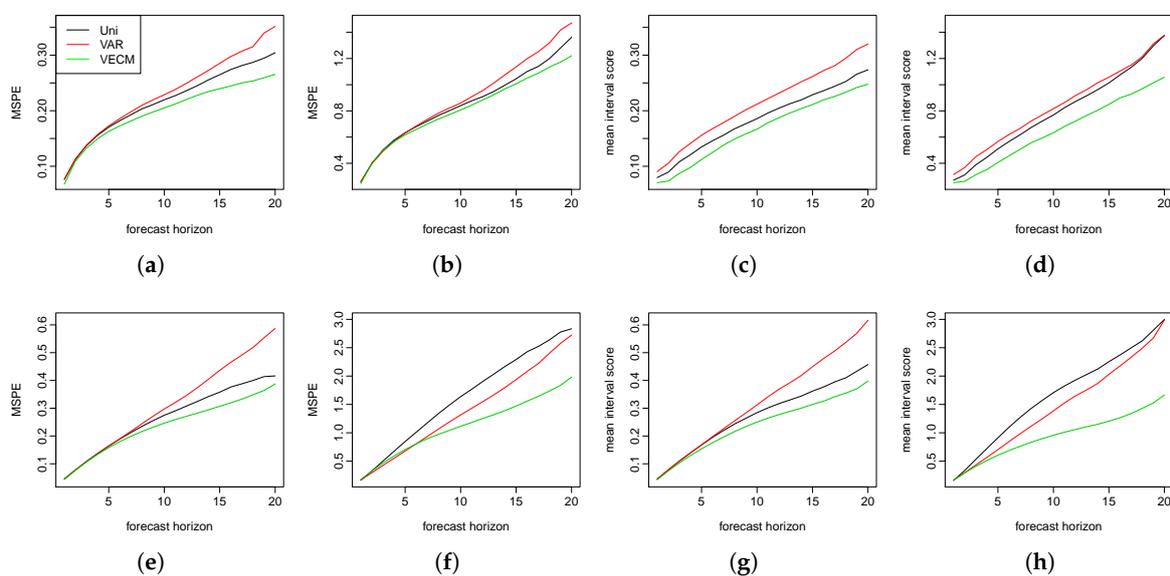
$$\begin{bmatrix} \Delta \tilde{\zeta}_{t,2}^{(1)} \\ \Delta \tilde{\zeta}_{t,2}^{(2)} \end{bmatrix} = \begin{bmatrix} 0.3 & 0.1 \\ 0.2 & 0.5 \end{bmatrix} \begin{bmatrix} \Delta \tilde{\zeta}_{t-1,2}^{(1)} \\ \Delta \tilde{\zeta}_{t-1,2}^{(2)} \end{bmatrix} + \begin{bmatrix} \epsilon_{t,2}^{(1)} \\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

For a VAR(1) model to be stable, it is required that  $\det(I_p - A_1 z) = 0$  should have all roots outside the unit circle.

4.3. Results

The principal component scores are generated using the aforementioned two models for observations  $t = 1, \dots, 60$ . Two sets of simulated functions are generated using Equation (7). We performed an FPCA on the two populations separately. The estimated principal component scores are then modeled using the univariate model, the VAR model and the VECM.

We repeated the simulation procedures 150 times. In each simulation, 500 bootstrap samples are generated to calculate the prediction intervals. We show the MSPE and the mean interval scores at each forecast horizon in Figure 2. The three models performed almost equally well in the short-term forecasts. In the long run, however, the functional VECM produced better predictions than the other two models. This advantage grew bigger as the forecast horizons increased.



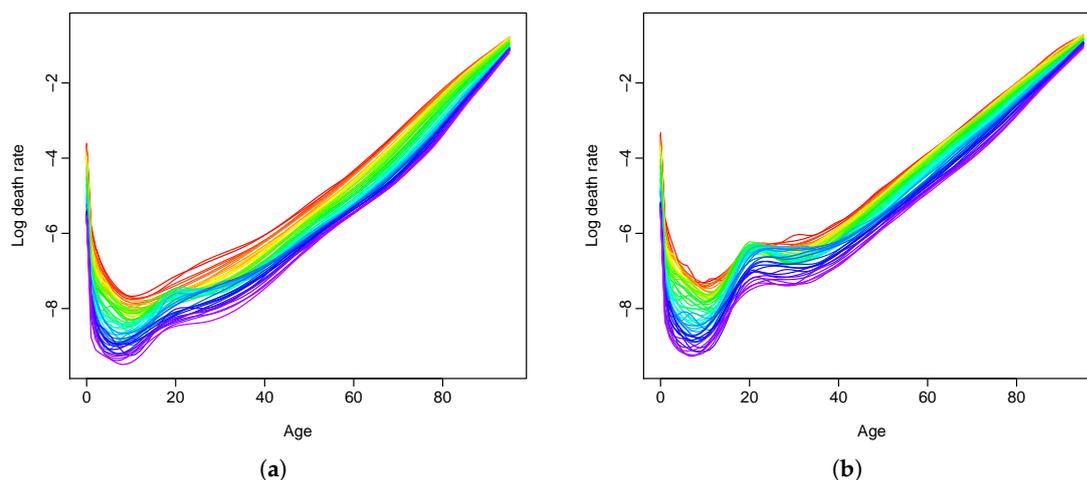
**Figure 2.** The first row presents the mean squared prediction error (MSPE) and the mean interval scores for the two populations in a co-integration setting. The second row presents the MSPE and the mean interval scores for the two populations without the co-integration. (a) 1st population; (b) 2nd population; (c) 1st population; (d) 2nd population; (e) 1st population; (f) 2nd population; (g) 1st population; and (h) 2nd population.

## 5. Empirical Studies

To show that the proposed model outperformed the existing ones using real data, we applied the four models illustrated in Section 2 to the sex-and age-specific mortality rates in Switzerland and the Czech Republic. The observations are yearly mortality curves from ages 0 to 110 years, where the age is treated as the continuum in the rate function. Female and male curves are available from 1908 to 2014 in [29]. We only used data from 1950 to 2014 for our analysis to avoid the possibly abnormal rates before 1950 due to war deaths. With the aim of forecasting, we considered the data before 1950 to be too distant to provide useful information. The data at ages 95 and older are grouped together, in order to avoid problems associated with erratic rates at these ages.

### 5.1. Swiss Age-Specific Mortality Rates

Figure 3 shows the smoothed log mortality rates for females and males from 1950 to 2014. We use a rainbow plot [30], where the red color represents the curves for more distant years and the purple color represents the curves for more recent years. The curves are smoothed using penalized regression splines with a monotonically increasing constraint after the age of 65 (see [15,31]). Over a span of 65 years, the mortality rates in general have decreased over all ages, with exceptions in the male population at around age 20. Female rates have been slightly lower than male rates over the years.



**Figure 3.** Smoothed log mortality rates in Switzerland from 1950 to 2014. (a) female population; (b) male population.

First, we tested the stationarity of our data set. The Monte Carlo test, in which the null hypothesis is stationarity, was applied to both the male and female populations. We used data from all 65 of the years in our range and performed 5000 Monte Carlo replications [32]. The  $p$ -values for the male and female populations were 0.0256 and 0.0276, respectively. These small  $p$ -values indicated a strong deviation from stationary functional time series.

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

$$K = \underset{m}{\operatorname{argmin}} \left\{ \frac{1}{15} \sum_{h=1}^{15} \sum_{j=0}^{95} \left[ \widehat{f}_{n'+h|n'}(x_j; m) - y_{n'+h}(x_j) \right]^2 \right\},$$

where  $\hat{f}_{n'+h|n'}(x_j; m)$  denotes the  $h$ -step-ahead forecast based on the first  $n' = 30$  years of data, with  $m$  dimensions retained.  $y_{n'+h}$  denotes the true rate at year  $n' + h$ . This selection scheme led to both the VAR and VECM models with  $K = 3$  basis functions in this case, which explained 91.20%, 4.37% and 1.56% of the variation in the training set, respectively. These add up to 97.13% of the total variances in the training data being explained. In the univariate and the product-ratio models, order  $K = 6$  is used as in [17,33], where they found that six components would suffice and that having more than six made no difference to the forecasts. With chosen  $K$  values, the four models were fitted using an expanding window approach (as explained in Section 3). This produced 20 one-step-ahead forecasts, 19 two-step-ahead forecasts... and, finally, one 20-step-ahead forecast. These forecasts are compared with the holdout data from the years 1995 to 2014. We calculated MAPE and MSPE as point forecast errors using Equation (4).

Table 1 presents the MSPE of the log mortality rates. The smallest errors at each forecast horizon are highlighted in bold face. For the prediction of the female rates, the proposed functional VECM has proved to make more accurate point forecasts for all forecast horizons except for the 20-step-ahead prediction. It should be noted that there is only one error estimate for the 20-step-ahead forecast, so the error estimate may be quite volatile. The other three approaches are somewhat competitive for the 11-step-ahead forecasts or less. For the longer forecast horizons, the errors of the product-ratio method increase quickly. For the forecasting of male mortality rates, although the VAR model produces slightly smaller values of the forecast errors, there is hardly any difference between the four models in the short term. For long-term predictions, the product-ratio approach performs much better than the univariate and the VAR models, but the VECM still dominates. In fact, the product-ratio model usually outperforms the existing models for the male mortality forecasts, while, for the female mortality forecasts, it is not as accurate. MAPEs of the models followed a similar pattern to the MSPE values and are not shown here.

**Table 1.** Mean squared prediction error (MSPE) for Swiss female and male rates (the smallest values are highlighted in bold).

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

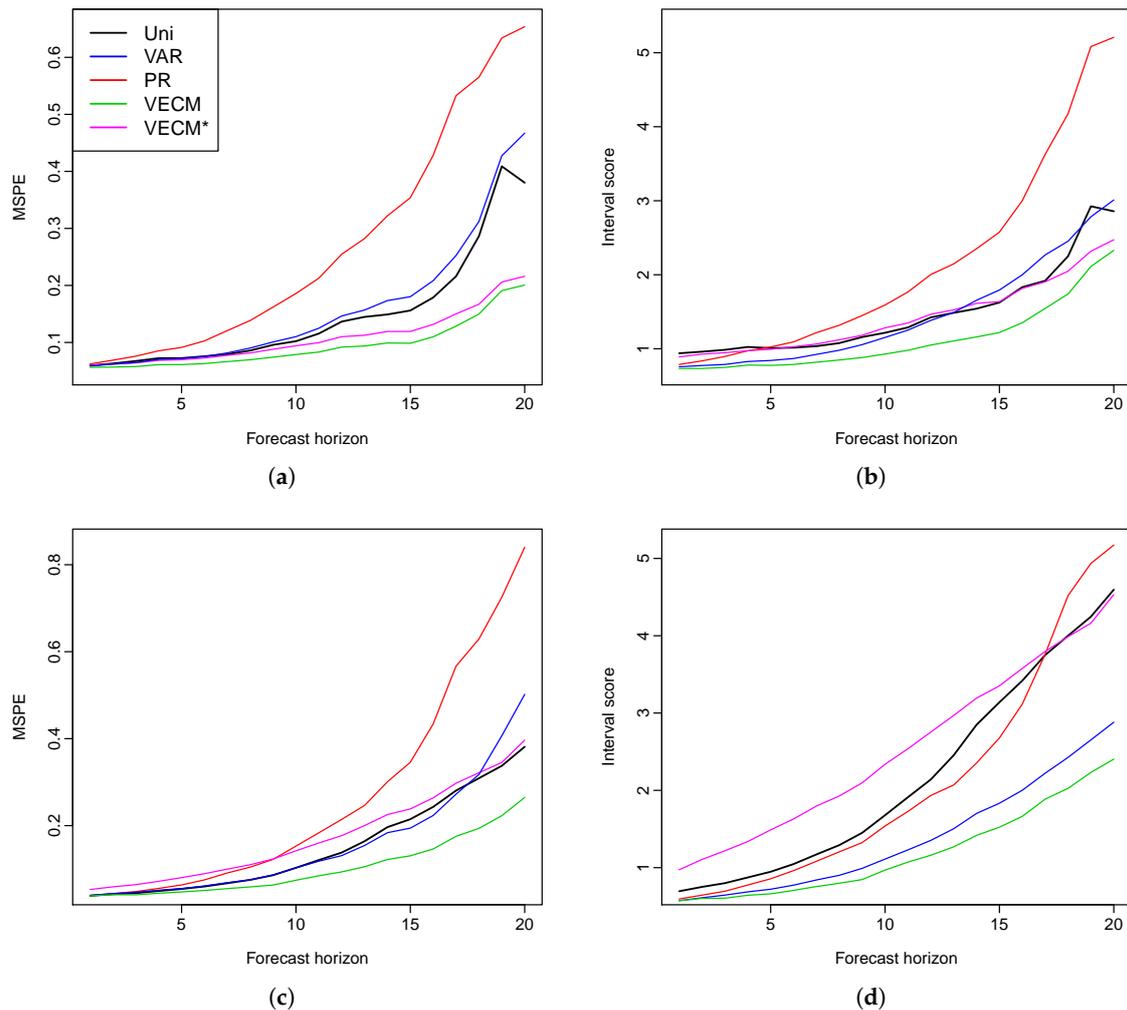
To examine how the models perform in interval forecasts, Equations (5) and (6) are used to calculate the mean interval scores. We generate 1,000 bootstrap samples in the functional VECM and VAR. Table 2 shows the mean interval scores. The 80% prediction intervals are produced using the four different approaches. As explained earlier, smaller mean interval score values indicate better interval predictions. For the female forecasts, functional VECM makes superior interval predictions at all forecast steps, while, for the male forecasts, the product–ratio model and VECM are very competitive, with the latter having a minor advantage for the mean value.

**Table 2.** Mean interval score (80%) for Swiss female and male rates (the smallest values are highlighted in bold).

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

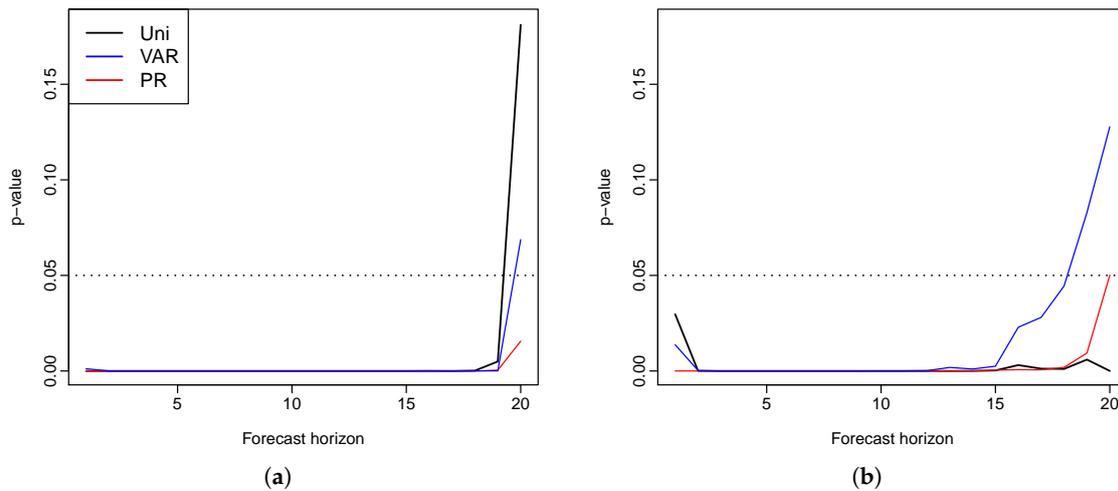
### 5.2. Czech Republic Age-Specific Mortality Rates

We have also applied the four models to other countries, such as the Czech Republic, to show that the proposed functional VECM does not only work in the case of the Swiss mortality rates. The raw data are grouped and smoothed as was done for the Swiss data.  $K = 5$  is chosen in the VAR and the VECM, and the proportions of the explained variance are 93.04%, 1.99%, 1.55%, 1.18%, and 0.79% respectively, which add up to 98.55% of the total variance explained. Figure 4 shows the MSPE and mean interval scores for the point and interval forecast evaluations. In order to compare with the VECM model in the literature, we also try fitting only the first set of principal component scores, shown in the figure by VECM\*. Among all five models, functional VECM produces better predictions in both the point and interval forecasts. Compared to our model that uses five principal component scores, VECM\* produces larger errors, especially in the male forecasts. We consider that an important fraction of information is lost if only the first set of principal component scores is used.



**Figure 4.** Czech Republic: forecast errors for female and male mortality rates (MSPE and interval scores are presented). (a) MSPE for female data; (b) mean interval score for female data; (c) MSPE for male data; (d) mean interval score for male data.

To examine whether or not the differences in the forecast errors are significant, we conduct the Diebold–Mariano test [34]. We use a null hypothesis where the two prediction methods have the same forecast accuracy at each forecast horizon, while the three alternative hypotheses used are that the functional VECM method produces more accurate forecasts than the three other methods. Thus, a small  $p$ -value is expected in favor of the alternatives. A squared error loss function is used and the  $p$ -values for one-sided tests are calculated at each forecast horizon, as shown in Figure 5. The  $p$ -values are hardly greater than zero at most forecast horizons. Almost all are below  $\alpha = 0.05$ , denoted by the horizontal line, with the exception of the 19- and 20-step-ahead forecasts. We conclude that there is strong evidence that the functional VECM method produces more accurate forecasts than the other three methods for most of the forecast horizons.



**Figure 5.** Czech Republic:  $p$ -values for the three tests comparing a functional VECM to the univariate, VAR, and product-ratio models, respectively (the horizontal line is the default level of significance  $\alpha = 0.05$ ). (a) female population; (b) male population.

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

## 6. Conclusions

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

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

While we use two populations for the illustration of the model and in the empirical analysis, functional VECM can easily be applied to populations with more than two groups. A higher rank of co-integration order may need to be considered and the Johansen test can then be used to determine the rank [35].

In this paper, we have focused on comparing our model with others within functional time series frameworks. There are numerous other mortality models in the literature, and many of them try to deal with multiple populations. Further research is needed to evaluate our model against the performance of these models.

**Acknowledgments:** The authors would like to thank three reviewers for insightful comments and suggestions, which led to a much improved manuscript. The authors thank Professor Michael Martin for his helpful comments and suggestions. Thanks also go to the participants of a school seminar at the Australian National University and Australian Statistical Conference held in 2016 for their comments and suggestions. The first author would also like to acknowledge the financial support of a PhD scholarship from the Australian National University.

**Author Contributions:** The authors contributed equally to the paper. Yuan Gao analyzed the data and wrote the paper. Han Lin Shang initiated the project and contributed analysis and a review of the literature.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Functional Principal Component Analysis

Let  $\{f_t(x), t \in Z\}$  be a set of functional time series in  $L_2(\mathcal{I})$  from a separable Hilbert space  $\mathcal{H}$ .  $\mathcal{H}$  is characterized by the inner product  $\langle \cdot, \cdot \rangle$ , where  $\langle f_1, f_2 \rangle = \int_{\mathcal{I}} f_1(x)f_2(x)dx$ . We assume that  $f(x)$  has a continuous mean function  $\mu(x)$  and covariance function  $G(w, x)$ :

$$\begin{aligned}\mu(x) &= E[f(x)], \\ G(w, x) &= \text{Cov}[f(w), f(x)] = E\{[f(w) - \mu(w)][f(x) - \mu(x)]\},\end{aligned}$$

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

$$C(w)(f) = \int_{\mathcal{I}} G(w, x)f(x)dx.$$

The eigenequation  $C(w)(f) = \rho f$  has solutions with orthonormal eigenfunctions  $\phi_k(x)$ , and associated eigenvalues  $\lambda_k$  for  $k = 1, 2, \dots$  such that  $\lambda_1 \geq \lambda_2 \geq \dots$  and  $\sum_k \lambda_k < \infty$ .

According to the Karhunen–Loève theorem, the function  $f(x)$  can be expanded by:

$$f(x) = \mu(x) + \sum_{k=1}^{\infty} \xi_k \phi_k(x), \quad (\text{A1})$$

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

$$\xi_k = \int_{\mathcal{I}} [f(x) - \mu(x)]\phi_k(x)dx.$$

The principal component scores also satisfy:

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

## Appendix B. Functional Principal Component Regression

According to Equation (A1), for a sequence of functional time series  $\{f_t(x)\}$ , each element can be decomposed as:

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

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

Assorted approaches for selecting the number of principal components,  $K$ , include: (a) ensuring that a certain fraction of the data variation is explained [36]; (b) cross-validation [14]; (c) bootstrapping [37]; and (d) information criteria [38].

With the smoothed functions  $\{f_1(x), \dots, f_n(x)\}$ , the mean function  $\mu(x)$  is estimated by

$$\hat{\mu}(x) = \frac{1}{n} \sum_{t=1}^n f_t(x).$$

The covariance operator for a function  $g$  is estimated by

$$\hat{C}(g) = \frac{1}{n} \sum_{t=1}^n \langle f_t - \hat{\mu}, g \rangle (f_t - \hat{\mu}),$$

where  $n$  is the number of observed curves. Sample eigenvalue and eigenfunction pairs  $\hat{\lambda}_k$  and  $\hat{\phi}_k(x)$  can be calculated from the estimated covariance operator using singular value decomposition. Empirical principal component scores  $\xi_{t,k}$  are obtained by  $\xi_{t,k} = \langle f_t, \hat{\phi}_k \rangle$  with numerical integration  $\int_{\mathcal{I}} [f_t(x) - \hat{\mu}(x)] \hat{\phi}_k(x) dx$ . These simple estimators are proved to be consistent under weak dependence when the functions collected are dense and regularly spaced [39,40]. In sparse data settings, other methods should be applied. For instance, Ref. [38] proposes principal component conditional expectation using pooled information between the functions to undertake estimations.

## References

1. Lee, R.D.; Carter, L.R. Modeling and Forecasting U. S. Mortality. *J. Am. Stat. Assoc.* **1992**, *87*, 659–671.
2. Booth, H.; Maindonald, J.; Smith, L. Applying Lee–Carter under conditions of variable mortality decline. *Popul. Stud.* **2002**, *56*, 325–336.
3. Koissi, M.C.; Shapiro, A.F.; Högnäs, G. Evaluating and extending the Lee–Carter model for mortality forecasting: Bootstrap confidence interval. *Insur. Math. Econ.* **2006**, *38*, 1–20.
4. Renshaw, A.E.; Haberman, S. A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insur. Math. Econ.* **2006**, *38*, 556–570.
5. Cairns, A.J.G.; Blake, D.; Dowd, K. A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *J. Risk Insur.* **2006**, *73*, 687–718.
6. Chan, W.; Li, J.S.; Li, J. The CBD Mortality Indexes: Modeling and Applications. *N. Am. Actuarial J.* **2014**, *18*, 38–58.
7. Carter, L.R.; Lee, R.D. Modelling and Forecasting US sex differentials in Modeling. *Int. J. Forecast.* **1992**, *8*, 393–411.
8. Li, N.; Lee, R. Coherent mortality forecasts for a group of populations: An extension of the Lee–Carter method. *Demography* **2005**, *42*, 575–594.
9. Yang, S.S.; Wang, C. Pricing and securitization of multi-country longevity risk with mortality dependence. *Insur. Math. Econ.* **2013**, *52*, 157–169.
10. Zhou, R.; Wang, Y.; Kaufhold, K.; Li, J.S.H.; Tan, K.S. Modeling Mortality of Multiple Populations with Vector Error Correction Models: Application to Solvency II. *N. Am. Actuarial J.* **2014**, *18*, 150–167.
11. Danesi, I.L.; Haberman, S.; Millosovich, P. Forecasting mortality in subpopulations using Lee–Carter type models: A comparison. *Insur. Math. Econ.* **2015**, *62*, 151–161.
12. Ramsay, J.O.; Silverman, J.W. *Functional Data Analysis*; Springer: New York, NY, USA, 2005.
13. Wahba, G. Smoothing noisy data with spline function. *Numer. Math.* **1975**, *24*, 383–393.
14. Rice, J.; Silverman, B. Estimating the Mean and Covariance Structure Nonparametrically When the Data Are Curves. *J. R. Stat. Soc. Ser. B (Methodol.)* **1991**, *53*, 233–243.
15. Hyndman, R.J.; Ullah, M.S. Robust forecasting of mortality and fertility rates: A functional data approach. *Comput. Stat. Data Anal.* **2007**, *51*, 4942–4956.
16. Chiou, J.M.; Müller, H.G. Linear manifold modelling of multivariate functional data. *J. R. Soc. Stat. Ser. B (Stat. Methodol.)* **2014**, *76*, 605–626.

17. Hyndman, R.J.; Booth, H.; Yasmineen, F. Coherent Mortality Forecasting: The Product-Ratio Method with Functional Time Series Models. *Demography* **2013**, *50*, 261–283.
18. Aue, A.; Norinho, D.D.; Hörmann, S. On the prediction of stationary functional time series. *J. Am. Stat. Assoc.* **2015**, *110*, 378–392.
19. Hyndman, R.J.; Khandakar, Y. Automatic Time Series Forecasting: The forecast Package for R. *J. Stat. Softw.* **2008**, *27*, doi:10.18637/jss.v027.i03.
20. Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*; Springer: New York, NY, USA, 2005.
21. Bosq, D. *Linear Processes in Function Spaces: Theory and Applications*; Springer Science & Business Media: New York, NY, USA, 2012; Volume 149.
22. Box, G.E.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*, 5th ed.; John Wiley & Sons: Hoboken, NJ, USA, 2015.
23. Granger, C.W.; Joyeux, R. An introduction to long-memory time series models and fractional differencing. *J. Time Ser. Anal.* **1980**, *1*, 15–29.
24. Hosking, J.R. Fractional differencing. *Biometrika* **1981**, *68*, 165–176.
25. Masarotto, G. Bootstrap prediction intervals for autoregressions. *Int. J. Forecast.* **1990**, *6*, 229–239.
26. Kim, J. Bootstrap-after-bootstrap prediction intervals for autoregressive models. *J. Bus. Econ. Stat.* **2001**, *19*, 117–128.
27. Faraway, J.J. Does data splitting improve prediction? *Stat. Comput.* **2016**, *26*, 49–60.
28. Gneiting, T.; Raftery, A.E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378.
29. Human Mortality Database. *University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany)*. 2016. Available online: <http://www.mortality.org> (accessed on 8 March 2016).
30. Hyndman, R.J.; Shang, H.L. Rainbow plots, bagplots, and boxplots for functional data. *J. Comput. Graph. Stat.* **2010**, *19*, 29–45.
31. Wood, S.N. Monotonic smoothing splines fitted by cross validation. *SIAM J. Sci. Comput.* **1994**, *15*, 1126–1133.
32. Horvath, L.; Kokoszka, P.; Rice, G. Testing stationarity of functional time series. *J. Econ.* **2014**, *179*, 66–82.
33. Hyndman, R.J.; Booth, H. Stochastic population forecasts using functional data models for mortality, fertility and migration. *Int. J. Forecast.* **2008**, *24*, 323–342.
34. Diebold, F.X.; Mariano, R.S. Comparing predictive accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263.
35. Johansen, S. Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica* **1991**, *59*, 1551–1580.
36. Chiou, J.M. Dynamical functional prediction and classification with application to traffic flow prediction. *Ann. Appl. Stat.* **2012**, *6*, 1588–1614.
37. Hall, P.; Vial, C. Assessing the finite dimensionality of functional data. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2006**, *68*, 689–705.
38. Yao, F.; Müller, H.; Wang, J. Functional data analysis for sparse longitudinal data. *J. Am. Stat. Assoc.* **2005**, *100*, 577–590.
39. Yao, F.; Lee, T.C.M. Penalized spline models for functional principal component analysis. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2006**, *68*, 3–25.
40. Hörmann, S.; Kokoszka, P. Weakly dependent functional data. *Ann. Stat.* **2010**, *38*, 1845–1884.

**Sample Availability:** Computational code in R are available upon request from the authors.



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).