Contribute 4
Common functional principal component models for mortality forecasting

Rob J Hyndman, Farah Yasmeen

Abstract We explore models for forecasting groups of functional time series data that exploit common features in the data. Our models involve fitting common (or partially common) functional principal component models and forecasting the coefficients using univariate time series methods. We illustrate our approach by forecasting age-specific mortality rates for males and females in Australia.

4.1 Functional time series models

We are interested in forecasting groups of functional time series data. For example, the annual age-specific mortality rates for a country can be considered a functional time series, $\{f_1(x), f_2(x), \dots, f_T(x)\}$, observed over years $1, \dots, T$, where x denotes age. When we observe mortality rates for subsets of the population (e.g., split by sex, or by race, or by geographical region), we have grouped functional time series.

Methods for forecasting functional time series data have been developed by [9, 8, 3, 13, 6]. A related but different problem involves forecasting a segmented continuous time series [1]. While our methods have application to this second problem, we will not consider it in this paper.

The most common approach to the problem of forecasting a functional time series is to use a principal components decomposition [12] and then to use univariate time series models to forecast each of the principal component scores.

Rob J Hyndman
Monash University, Australia, e-mail: rob.hyndman@monash.edu

Farah Yasmeen
University of Karachi, Pakistan, e-mail: riazfarah@yahoo.com

That is, we use the model

$$f_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_t(x) \tag{4.1}$$

where $\mu(x)$ is estimated as the mean of the observed data for each x, the first K functional principal components are given by $\phi_1(x), \dots, \phi_K(x)$, and $\beta_{t,1}, \dots, \beta_{t,K}$ are the corresponding principal component scores. The error term, $e_t(x)$, is often assumed to be normally distributed for each x. A robust version of this model was proposed by [9] and versions involving partial least squares and weighted functional principal components were discussed by [8].

Our interest in this paper is where we have two or more groups of functional time series data and we wish to model and forecast them, taking account of any shared patterns and constraining the forecasts so they do not diverge. Our partial common functional principal component (PCFPC) model for J groups can be written as

$$f_{t,j}(x) = \mu_j(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + \sum_{\ell=1}^{L} \gamma_{t,j,\ell}\psi_{j,\ell}(x) + \varepsilon_{t,j}(x), \tag{4.2}$$

where $j = 1, \dots, J$. Thus we allow a different mean μ_j for each group, and a set of common principal components $\phi_1(x), \dots, \phi_K(x)$. Optionally, we also allow for some uncommon principal components for each group, $\psi_{j,1}(x), \dots, \psi_{j,L}(x)$. We call this a PCFPC($K, L$) model.

The product-ratio model proposed by [6] is a special case of (4.2), where the estimation procedure used requires each time series $\{\gamma_{t,j,\ell}\}$ to be stationary. In this paper, we allow greater flexibility.

There are several advantages with using (4.2) rather than applying (4.1) for each group independently. First, if there are common features in the groups of data, these can be captured with the common principal components. Second, we can prevent the forecasts of the groups from diverging by requiring $\gamma_{t,j,\ell} - \gamma_{t,i,\ell}$ to be stationary for each combination of i, j and ℓ so that

$$\lim_{t\to\infty} \mathrm{E}\|f_{t,j} - f_{t,i}\| < \infty \qquad \text{for all } i \text{ and } j.$$

In demography, this is known as having "coherent" forecasts [11, 6]. Thus we can impose coherence where appropriate by either requiring cointegrated scores, or stationary scores.

We consider several variations on (4.2) when there are $J = 2$ groups:

- Model 1: PCFPC($K, 0$). That is, there are no idiosyncratic principal components in the model.
- Model 2: PCFPC(K, L) with a coherence constraint. For each ℓ, the time series given by $\{\gamma_{t,1,\ell} - \gamma_{t,2,\ell}\}$ is stationary.
- Model 3: PCFPC($0, L$). That is, all principal components and scores are idiosyncratic. This is equivalent to applying the model (4.1) to each group independently.

Estimation for each of these models is described in [14], using methods discussed in [2].

4.2 All-cause mortality data of Australia

We will apply these models to the all-cause mortality data of males and females in Australia starting from 1950 to avoid the outliers due to the World Wars. The data for 1950–2009 were obtained from [4]. All data were smoothed (independently for each year) using penalized regression splines as described in [9].

A few of the curves are shown in Figure 4.1. While some features of the curves are clearly similar between sexes, there are differences. For example, the "accident hump" for males around age 20 is much less pronounced for females. Also the male mortality rates are generally higher than those for females.

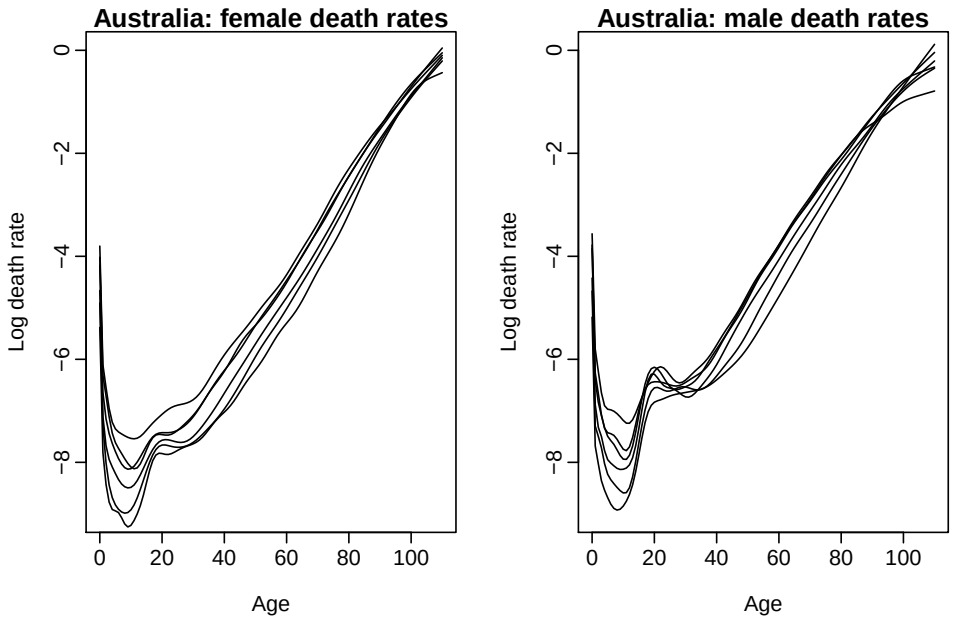

Figure 4.1: Some of the annual curves showing Australian mortality rates for each sex. The years 1950, 1960, 1970, 1980, 1990 and 2000 are shown. For most ages, the death rate decreases with time.

For all models, we set the number of principal components to be six following [5]. For common principal component scores, we use univariate ARIMA models, selected and estimated using the automatic algorithm of [7]. Where a score is assumed to be stationary, we fit a univariate stationary ARFIMA model where the differencing order is between 0 and 0.5, and the ARMA part of the model is automatically selected in the same way as for ARIMA models. Where the scores are cointegrated, we fit a VECM model using the Johansen procedure [10].

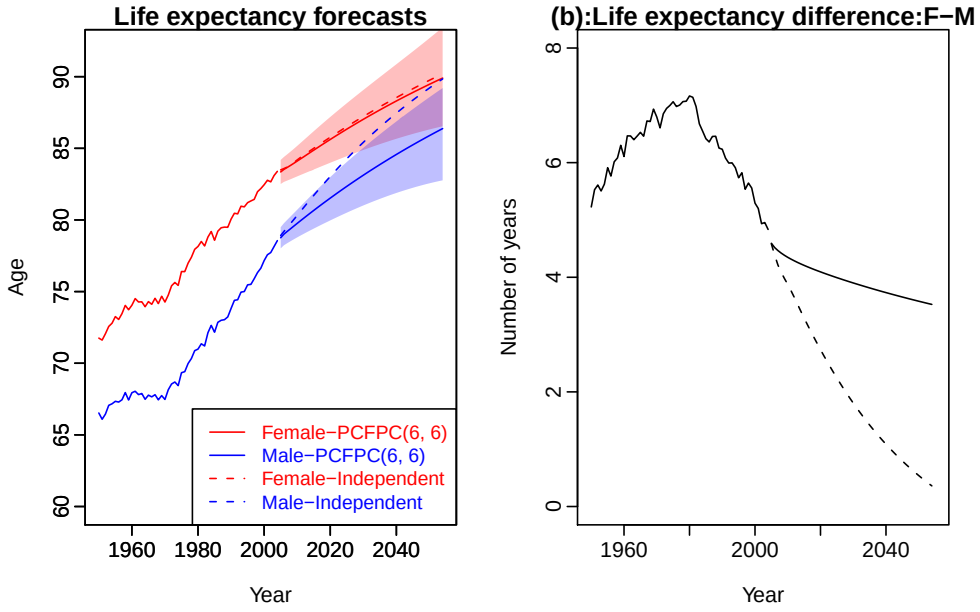

Figure 4.2: Twenty-year life-expectancy forecasts for males and females in Australia using the PCFPC(6,6) and independent models (with $L = 6$). 80% prediction intervals for the PCFPC models are also plotted for the two groups.

Figure 4.2 shows twenty-year forecasts of the life expectancy at birth of males and females using both Models 2 and 3. The independent forecasts have converged unrealistically fast, and if extrapolated further the male life expectancy would exceed the female life expectancy. On the other hand, the PCFPC model has taken account of common features in the two series, and has constrained the forecasts so they cannot diverge. Instead, the difference between the two life expectancy forecasts is approaching a constant.

To test the forecast accuracy of the models, we apply the rolling forecast origin approach of [6] where we fit each model to the first t observations, and then forecast the mortality rates for the observations in years $t+1, t+2, \ldots$. The value of t varies from 1969 to 2008 and up to twenty forecasts are generated from each model and compared against the actual mortality rates for the corresponding years.

In Table 4.1, the mean squared forecast errors at forecast horizons $h = 5, 10, 15$ and 20 are given for the three models described above. The bold entries represent the minimum value at each forecast horizon. While Model 3 (independent models for each group) gives good forecast accuracy over the horizons considered, there is no coherence constraint applied to these models, and so they are unsuitable for longer-term forecasting. Of the two constrained models, Model 1 (where the groups differ only in mean) provides the best results over the forecast horizons considered. The more complicated PCFPC model gives slightly less accurate

Forecast horizon	Groups	Model 1 PCFPC(6,0)	Model 2 PCFPC(6,6)	Model 3 PCFPC(0,6)
$h = 5$	Combined (F & M)	2.59	2.60	**2.52**
	Female (F)	2.81	2.75	**2.63**
	Male (M)	**2.38**	2.45	2.42
$h = 10$	Combined (F & M)	**4.57**	4.66	4.65
	Female(F)	4.67	4.43	**4.23**
	Male (M)	**4.48**	4.89	5.06
$h = 15$	Combined (F & M)	**7.72**	8.00	8.15
	Female (F)	7.31	6.64	**6.47**
	Male(M)	**8.14**	9.36	9.82
$h = 20$	Combined (F & M)	**12.97**	13.56	14.10
	Female (F)	12.26	10.41	**10.35**
	Male (M)	**13.69**	16.70	17.86

Table 4.1: Forecast accuracy of the independent and coherent models, showing the out-of-sample MSFE of log mortality. The bold entries denote the minimum value in each row.

forecasts over these horizons than the model with no idiosyncratic principal components or scores.

The independent models work better for female data. One possible reason for this might be the hump in the male mortality rates around age 20, which may be being captured in the common principal components.

4.3 Conclusions and further work

Our proposed PCFPC model has provided encouraging results in the application described here, and in other applications detailed in [14]. However, it has not routinely out-performed some simpler models. We regard this as a problem of model selection rather than a problem with the model itself, as our PCFPC model is more general than the competitor models.

For example, the full PCFPC model has a large number of coefficients, and so more degrees of freedom. So far, we have simply set $K = L = 6$ in all applications. We expect an appropriate order selection procedure would lead to better forecast accuracy. Also, our automated VECM approach has not been widely tested, unlike the automated ARIMA models used for the simpler models. We expect a better VECM selection algorithm will also lead to better forecast accuracy.

Bibliography

[1] G. Aneiros-Pérez and P. Vieu (2008) Nonparametric time series prediction: A semi-functional partial linear modeling. *Journal of Multivariate Analysis*, 99(5), 834 – 857.

[2] B. Flury (1988) *Common principal components and related multivariate models*. John Wiley and Sons.

[3] S. Hays, H. Shen, and J. Z. Huang (2012) Functional dynamic factor models with application to yield curve forecasting. *Annals of Applied Statisitcs*, 6(3), 870–894.

[4] Human Mortality Database (2014) University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Accessed 15 January 2014.

[5] R. J. Hyndman and H. Booth (2008) Stochastic population forecasts using functional data models for mortality, fertility and migration. *International Journal of Forecasting*, 24(3), 323–342.

[6] R. J. Hyndman, H. Booth, and F. Yasmeen (2013) Coherent mortality forecasting: the product-ratio method with functional time series models. *Demography*, 50(1), 261–283.

[7] R. J. Hyndman and Y. Khandakar (2008) Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3).

[8] R. J. Hyndman and H. L. Shang (2009) Forecasting functional time series (with discussion). *Journal of the Korean Statistical Society*, 38(3), 199–221.

[9] R. J. Hyndman and Md.S. Ullah (2007) Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics and Data Analysis*, 51(10), 4942–4956.

[10] S.J. Johansen (1998) Statistical analysis of cointegration vectors. *Journal of Economics Dynamics and Control*, 12, 231–254.

[11] N. Li and R. D. Lee (2005) Current mortality forecasts for a group of populations: an extension of the Lee-Carter method. *Demography*, 42(3), 575–594.

[12] J. O. Ramsay and B. W. Silverman (2005) *Functional data analysis*. 2nd edition. Springer: New York

[13] H. Shen (2008) On modeling and forecasting time series of smooth curves. *Technometrics*, 51(3), 1–32.

[14] F. Yasmeen. *Functional linear models for mortality forecasting*. PhD, Monash University, 2010.