Coherent mortality forecasting: the product-ratio method with functional time series models

Rob J Hyndman

Department of Econometrics and Business Statistics,

Monash University, VIC 3800

Australia.

Email: rob.hyndman@monash.edu

Heather Booth

Australian Demographic and Social Research Institute

The Australian National University, ACT 0200

Australia.

Email: heather.booth@anu.edu

Farah Yasmeen

Department of Econometrics and Business Statistics,

Monash University, VIC 3800

Australia.

Email: riazfarah@yahoo.com

20 April 2012

Coherent mortality forecasting: the product-ratio method with functional time series models

Abstract

When independence is assumed, forecasts of mortality for sub-populations are almost always divergent in the long term. We propose a method for coherent forecasting of mortality rates for two or more sub-populations, based on functional principal components models of simple and interpretable functions of rates. The product-ratio functional forecasting method models and forecasts the geometric mean of sub-population rates and the ratio of sub-population rates to product rates. Coherence is imposed by constraining the forecast ratio function through stationary time series models. The method is applied to sex-specific data for Sweden and state-specific data for Australia. Based on out-of-sample forecasts, the coherent forecasts are at least as accurate in overall terms as comparable independent forecasts, and forecast accuracy is homogenised across sub-populations.

1 Introduction

In recent years, we have seen considerable development in the modeling and forecasting of mortality. This includes the pioneering work by Lee & Carter (1992) and a variety of extensions and modifications of the single-factor Lee-Carter model (for example, Lee & Miller 2001, Booth et al. 2002*a*, Currie et al. 2004, Renshaw & Haberman 2003, Cairns et al. 2006, 2008, Hyndman & Ullah 2007); Booth & Tickle (2008) provide a comprehensive review. Most of this work has focused on forecasting mortality for a single population. However, it is often important to model mortality for two or more populations simultaneously, as for example, when forecasting mortality by sex, by states within a country, or by country within a larger region such as Europe. In such cases, it is usually desirable that the forecasts do not diverge over time.

Non-divergent forecasts for sub-populations within a larger population have been labelled "coherent" (Li & Lee 2005). Coherent forecasting seeks to ensure that the forecasts for related populations maintain certain structural relationships based on extensive historic observation and theoretical considerations. For example, male mortality has been observed to be consistently higher than female mortality at all ages, and while available evidence supports both biological/genetic and social/cultural/environmental/behavioural hypotheses as to why this is so, Kalben (2002) concludes that the determining factor is biological. Thus sex-differences in mortality can be expected to persist in the future and to remain within observed limits. Mortality forecasts for regions within the same country, can also be expected not to diverge radically. Differences due to environmental and biological factors can be expected to remain unchanged, while differences due to social, political, behavioural and cultural factors are unlikely to produce persistent divergence in modern democracies where principles of equality apply. Similar arguments may apply to countries within a common economic and political framework such as the European Union.

The problem of obtaining coherent forecasts has previously been considered by Lee & Nault (1993), Lee (2000) and Li & Lee (2005) in the context of the Lee-Carter model. Lee (2006) presents a general framework for forecasting life expectancy as the sum of a common trend and the population-specific rate of convergence towards that trend.

A related problem occurs in the case of disaggregated forecasts where there are several additive absorbing states; for example, in forecasting mortality by cause of death. Wilmoth (1995) quantified the way in which the multiple-decrement forecast based on proportional changes in mortality rates by cause of death tends to be more pessimistic than the single-decrement

forecast of total rates. Using death densities and compositional methods, Oeppen (2008) found that the multiple-decrement compositional forecast is not always more pessimistic than the single-decrement case.

Coherence has not previously been defined more precisely than the non-divergence of forecasts for sub-populations. In this paper, we adopt a more precise definition: we label mortality forecasts as coherent when the forecast age-specific ratios of death rates for any two sub-populations converge to a set of appropriate constants. Exactly what this implies for the forecast trajectories is detailed in Section 4.2.

In Section 2, we propose a new method for coherent mortality forecasting which involves forecasting interpretable product and ratio functions of rates using the functional data paradigm introduced in Hyndman & Ullah (2007). The product-ratio functional forecasting method can be applied to two or more sub-populations, incorporates convenient calculation of prediction intervals as well as point forecasts, and is suitable for use within a larger stochastic population modelling framework such as Hyndman & Booth (2008). The new method is simple to apply, flexible in its dynamics, and produces forecasts that are at least as accurate in overall terms as the comparable independent method.

In Section 3, we illustrate the product-ratio method with functional time series models by applying it to Swedish male and female mortality data and Australian state-specific mortality data. We compare our mortality forecasts with those obtained from independent functional time series models, and evaluate relative accuracy through out-of-sample forecasts. Section 4 contains a discussion of the method and findings.

2   A coherent functional method

We initially frame the problem in terms of forecasting male and female age-specific death rates, as the two-sex application is the most common and the best understood. However, our method can be applied generally to any number of sub-populations.

Let $m_{t,F}(x)$ denote the female death rate for age x and year t, $t = 1,\ldots,n$. We will model the log death rate, $y_{t,F}(x) = \log[m_{t,F}(x)]$. Similar notation applies for males.

2.1 Functional data models

In the functional data paradigm, we assume that there is an underlying smooth function $f_{t,F}(x)$ that we are observing with error. Thus,

$$y_{t,F}(x_i) = \log[f_{t,F}(x_i)] + \sigma_{t,F}(x_i)\varepsilon_{t,F,i}, \tag{1}$$

where x_i is the centre of age-group i ($i = 1,\ldots,p$), $\varepsilon_{t,F,i}$ is an independent and identically distributed standard normal random variable and $\sigma_{t,F}(x_i)$ allows the amount of noise to vary with age x. Analogous notation is used for males. For smoothing, we use weighted penalized regression splines (Wood 1994) constrained so that each curve is monotonically increasing above age $x = 65$ (see Hyndman & Ullah 2007). The weights are to take care of the heterogeneity in death rates across ages. The observational variance $\sigma_{t,F}(x)$ is estimated by using a separate penalized regression spline of $\{y_t(x) - \log[f_{t,F}(x_i)]\}^2$ against x, for each t.

2.2 Product-ratio method for males and females

We define the square roots of the products and ratios of the smoothed rates for each sex:

$$p_t(x) = \sqrt{f_{t,M}(x)f_{t,F}(x)} \quad \text{and} \quad r_t(x) = \sqrt{f_{t,M}(x)/f_{t,F}(x)}.$$

We model these quantities rather than the original sex-specific death rates. The advantage of this approach is that the product and ratio will behave roughly independently of each other provided the sub-populations have approximately equal variances. On the log-scale, these are sums and differences which are approximately uncorrelated.[1] (If there are substantial differences in the variances of sub-populations, then the product and ratio will no longer be uncorrelated on the log scale. This does not bias our forecasts, but it makes them less efficient than they would otherwise be. In any case, our method will still be much better than treating the populations independently as we will still be imposing coherence constraints.)

---

[1] If X and Y are two random variables, then $\text{Corr}(X - Y, X + Y) = \text{Var}(X) - \text{Var}(Y)$. So if X and Y have equal variances, then $X - Y$ and $X + Y$ are uncorrelated.

We use functional time series models (Hyndman & Ullah 2007) for $p_t(x)$ and $r_t(x)$:

$$\log[p_t(x)] = \mu_p(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_t(x) \tag{2a}$$

$$\log[r_t(x)] = \mu_r(x) + \sum_{\ell=1}^{L} \gamma_{t,\ell}\psi_\ell(x) + w_t(x), \tag{2b}$$

where the functions $\{\phi_k(x)\}$ and $\{\psi_\ell(x)\}$ are the principal components obtained from decomposing $\{p_t(x)\}$ and $\{r_t(x)\}$ respectively, and $\beta_{t,k}$ and $\gamma_{t,\ell}$ are the corresponding principal component scores. The function $\mu_p(x)$ is the mean of the set of curves $\{p_t(x)\}$, and $\mu_r(x)$ is the mean of $\{r_t(x)\}$. The error terms, given by $e_t(x)$ and $w_t(x)$, have zero mean and are serially uncorrelated.

The models are estimated using the weighted principal components algorithm of Hyndman & Shang (2009) which places more weight on recent data, and so avoids the problem of the functions $\{\phi_k(x)\}$ and $\{\psi_\ell(x)\}$ changing over time (Lee & Miller 2001, p545).

The coefficients, $\{\beta_{t,1}, \ldots, \beta_{t,K}\}$ and $\{\gamma_{t,1}, \ldots, \gamma_{t,L}\}$, are forecast using time series models as detailed in Section 2.4. To ensure the forecasts are coherent, we require the coefficients $\{\gamma_{t,\ell}\}$ to be stationary processes. The forecast coefficients are then multiplied by the basis functions, resulting in forecasts of the curves $p_t(x)$ and $r_t(x)$ for future t. If $p_{n+h|n}(x)$ and $r_{n+h|n}(x)$ are h-step forecasts of the product and ratio functions respectively, then forecasts of the sex-specific death rates are obtained using $f_{n+h|n,M}(x) = p_{n+h|n}(x)r_{n+h|n}(x)$ and $f_{n+h|n,F}(x) = p_{n+h|n}(x)/r_{n+h|n}(x)$.

2.3 Product-ratio method for more than two sub-populations

More generally, suppose we wish to apply a coherent functional method for $J > 2$ subpopulations. Let

$$p_t(x) = \left[f_{t,1}(x)f_{t,2}(x) \cdots f_{t,J}(x) \right]^{1/J} \tag{3}$$

$$\text{and} \quad r_{t,j}(x) = f_{t,j}(x)/p_t(x), \tag{4}$$

where $j = 1, \ldots, J$ and $p_t(x)$ is the geometric mean of the smoothed rates and so represents the joint (non-stationary) behaviour of all subpopulations. The previous equations for males and females are a special case of these equations when $J = 2$.

We can assume approximately zero correlation between the product functions $p_t(x)$ and each set of ratio functions $r_{t,j}(x)$. The ratio functions will satisfy the constraint $r_{t,1}(x)r_{t,2}(x) \cdots r_{t,J}(x) = 1$.

Our models do not include a jump-off bias adjustment (Lee & Miller 2001) as it is unnecessary when several principal components are used. In fact, adjusting for jump-off bias may increase the forecast bias (Booth et al. 2006).

We use functional time series models for $p_t(x)$ and each $r_{t,j}(x)$ for $j = 1, 2, \ldots, J$:

$$\log[p_t(x)] = \mu_p(x) + \sum_{k=1}^{K} \beta_{t,k} \phi_k(x) + e_t(x) \tag{5a}$$

$$\log[r_{t,j}(x)] = \mu_{r,j}(x) + \sum_{l=1}^{L} \gamma_{t,l,j} \psi_{l,j}(x) + w_{t,j}(x). \tag{5b}$$

It is only necessary to fit $J - 1$ models for the ratio functions due to the constraint — recall that we only needed one ratio model for the two-sex case when $J = 2$. However, fitting all J models will simplify the variance calculation for producing prediction intervals.

Thus, the implied model for each group is given by

$$\log[f_{t,j}(x)] = \log[p_t(x)r_{t,j}(x)] = \mu_j(x) + \sum_{k=1}^{K} \beta_{t,k} \phi_k(x) + \sum_{\ell=1}^{L} \gamma_{t,\ell,j} \psi_{\ell,j}(x) + z_{t,j}(x), \tag{6}$$

where $\mu_j(x) = \mu_p(x) + \mu_{r,j}(x)$ is the group mean and $z_{t,j}(x) = e_t(x) + w_{t,j}(x)$ is the error term. Note that equation (6) is equivalent to equation (15) in Hyndman & Ullah (2007). Here, we are providing an easy and interpretable approach to estimation for this model.

2.4 Forecasts and prediction intervals

Forecasts are obtained by forecasting each coefficient $\beta_{t,1}, \ldots, \beta_{t,K}$ and $\gamma_{t,1}, \ldots, \gamma_{t,L}$ independently. There is no need to consider vector models as the $\beta_{t,k}$ coefficients are all uncorrelated by construction (see Hyndman & Ullah 2007), as are the $\gamma_{t,\ell}$ coefficients. They are also approximately uncorrelated with each other due to the use of products and ratios.

The coefficients of the product model, $\{\beta_{t,1}, \ldots, \beta_{t,K}\}$, are forecast using possibly non-stationary ARIMA models (Shumway & Stoffer 2006) without restriction. When fitting ARIMA models, we use the automatic model selection algorithm given by Hyndman & Khandakar (2008) to select the appropriate model orders.

The coefficients of the ratio model, $\{\gamma_{t,1,j}, \gamma_{t,2,j}, \ldots, \gamma_{t,L,j}\}$, $j = 1, 2, \ldots, J$, are each forecast using any stationary ARMA(p, q) (Box et al. 2008) or ARFIMA(p, d, q) process (Granger & Joyeux

1980, Hosking 1981). The stationarity requirement ensures that the forecasts are coherent. We have found ARFIMA models useful for forecasting the ratio function coefficients as they provide for longer-memory behaviour than is possible with ARMA models. In implementing ARFIMA(p, d, q) models, we estimate the fractional differencing parameter d using the method of Haslett & Raftery (1989), we use the algorithm of Hyndman & Khandakar (2008) to choose the orders p and q, then we use maximum likelihood estimation (Haslett & Raftery 1989) and the forecast equations of Peiris & Perera (1988). In the stationary ARFIMA models, $-0.5 < d < 0.5$; for $d = 0$, an ARFIMA(p, d, q) model is equivalent to an ARMA(p, q) model.

Let $\hat{\beta}_{n+h|n,k}$ denote the h-step ahead forecast of $\beta_{n+h,k}$ and let $\hat{\gamma}_{n+h|n,\ell,j}$ denote the h-step ahead forecast of $\gamma_{n+h,\ell,j}$. Then the h-step ahead forecast of $\log m_{n+h,j}(x)$ is given by

$$\log[\hat{m}_{n+h|n,j}(x)] = \hat{\mu}_j(x) + \sum_{k=1}^{K} \hat{\beta}_{n+h|n,k} \phi_k(x) + \sum_{\ell=1}^{L} \hat{\gamma}_{n+h|n,\ell,j} \psi_{\ell,j}(x). \tag{7}$$

Because all terms are uncorrelated, we can simply add the variances together so that

$$\text{Var}\left\{\log[m_{n+h,j}(x)] \mid \mathcal{I}_n\right\} = \hat{\sigma}_{\mu_j}^2(x) + \sum_{k=1}^{K} u_{n+h|n,k} \phi_k^2(x) + \sum_{\ell=1}^{L} v_{n+h|n,\ell,j} \psi_{\ell,j}^2(x) + s_e(x) + s_{w,j}(x) + \sigma_{n+h,j}^2(x), \tag{8}$$

where $\mathcal{I}_n$ denotes all observed data up to time n plus the basis functions $\{\phi_k(x)\}$ and $\{\psi_\ell(x)\}$, $u_{n+h|n,k} = \text{Var}(\beta_{n+h,k} \mid \beta_{1,k}, \ldots, \beta_{n,k})$ and $v_{n+h|n,k} = \text{Var}(\gamma_{n+h,k} \mid \gamma_{1,k}, \ldots, \gamma_{n,k})$ can be obtained from the time series models, $\hat{\sigma}_{\mu_j}^2(x)$ (the variances of the smoothed means) can be obtained from the smoothing method used, $s_e(x)$ is estimated by averaging $\hat{e}_t^2(x)$ for each x and $s_{w,j}(x)$ is estimated by averaging $\hat{w}_{t,j}^2(x)$ for each x. The observational variance $\sigma_{t,j}^2(x)$ is very stable from year to year and so it is estimated by taking the mean observational variance in the historical data. A prediction interval is then easily constructed under the assumption that the errors are normally distributed.

Life expectancy forecasts are obtained from the forecast age-specific death rates using standard life table methods (Preston et al. 2001). To obtain prediction intervals for life expectancies, we simulate a large number of future death rates as described in Hyndman & Booth (2008) and obtain the life expectancy for each. Then the prediction intervals are constructed from percentiles of these simulated life expectancies.

3 Empirical Studies

In this section, we will illustrate the product-ratio functional forecasting method using two applications: male and female data for Sweden, and state data for Australia. With each application, we compare the forecast accuracy of our method with the accuracy of independent functional forecasts for each sub-population, based on out-of-sample forecasts. This does not, of course, assess the accuracy of our forecasting applications, but it does provide two comparisons of the accuracy of the coherent and independent approaches. We fit each model to the first t observations and then forecast the death rates for the observations in years $t+1, t+2, \ldots, n$. The value of t is allowed to vary from $n_1 = 20$ to $n-1$. Correspondingly, the forecast horizon varies from $n-20$ to 1.

The number of components, K and L, in our models is fixed at $K = L = 6$. It was found in Hyndman & Booth (2008) that the forecasts are insensitive to the number of components provided there are enough components. That is, having too many components does not degrade forecast accuracy but having too few components does. Greenshtein & Ritov (2004) and Greenshtein (2006) termed this phenomenon "persistence in functional linear predictor selection". We have found that having more than six components makes almost no difference to the resulting forecasts.

For the jth group and forecast horizon h, we define the mean square forecast error (MSFE) to be

$$\text{MSFE}_j(h) = \frac{1}{(n-h-n_1+1)p} \sum_{t=n_1}^{n-h} \sum_{i=1}^{p} \left\{ \log[m_{t+h,j}(x_i)] - \log[\hat{m}_{t+h|t,j}(x_i)] \right\}^2. \tag{9}$$

3.1 Application to Swedish male and female mortality

Age-sex specific data for Sweden are available for the years 1751–2007 from Human Mortality Database (2010). The data consist of central death rates and mid-year populations by sex and single years of age to 110 years. We group the data at ages 100 and above in order to avoid problems associated with erratic rates at these ages. We restrict the data to 1950–2007, but this has minimal effect because the weighted fitting procedure gives greater weight to recent data.

Figure 1 shows log death rates for males and females for selected age groups as univariate time series. The bottom panels represent the same data as a series of smoothed curves (functional observations). We use rainbow plots (Hyndman & Shang 2010) with the time ordering of curves indicated by the colors of the rainbow, from red to violet. For females, mortality has declined
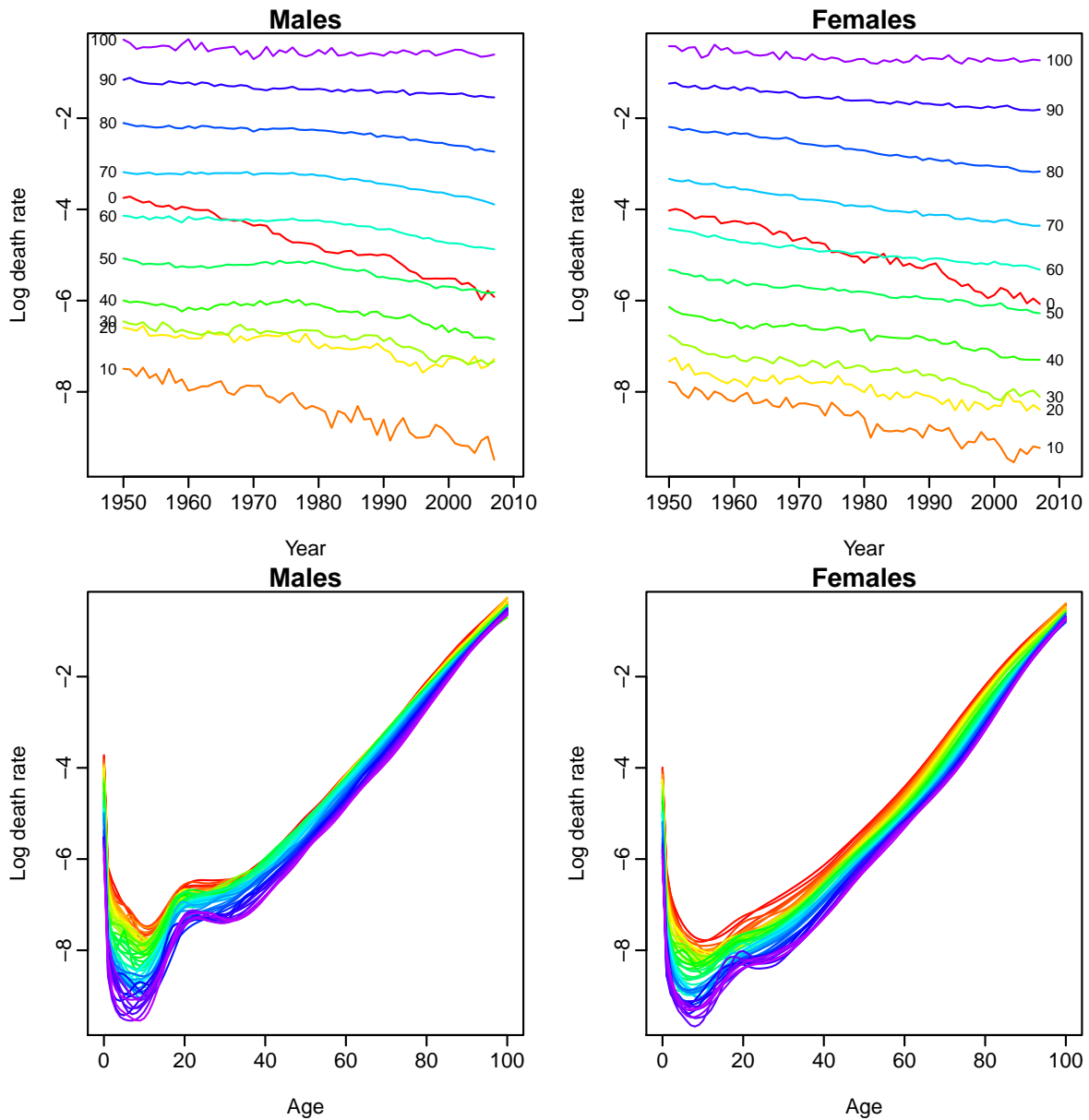
Figure 1: Top: Log death rates for males and females in Sweden for various ages as univariate time series. Bottom: Smoothed log death rates for males and females in Sweden viewed as functional time series, observed from 1950 to 2007. Penalized regression splines with a partial monotonic constraint (Hyndman & Ullah 2007) are used for smoothing.

steadily at most ages. For males, however, the decline stagnated at most adult ages in the 1960s and early 1970s. In recent years, male and female mortality at ages 20 and 30 are no longer declining, and small numbers of deaths result in relatively large fluctuations at age 10.

Figure 2 shows the product function (on the log scale) and the ratio function of observed rates. The product function is the geometric mean of smoothed male and female rates; while in general this mean declines over time, recent increases are observed at ages 15 to 30, and at ages 1 to 5. The ratio function is the square root of the sex ratio of smoothed rates. Sex ratios of less than
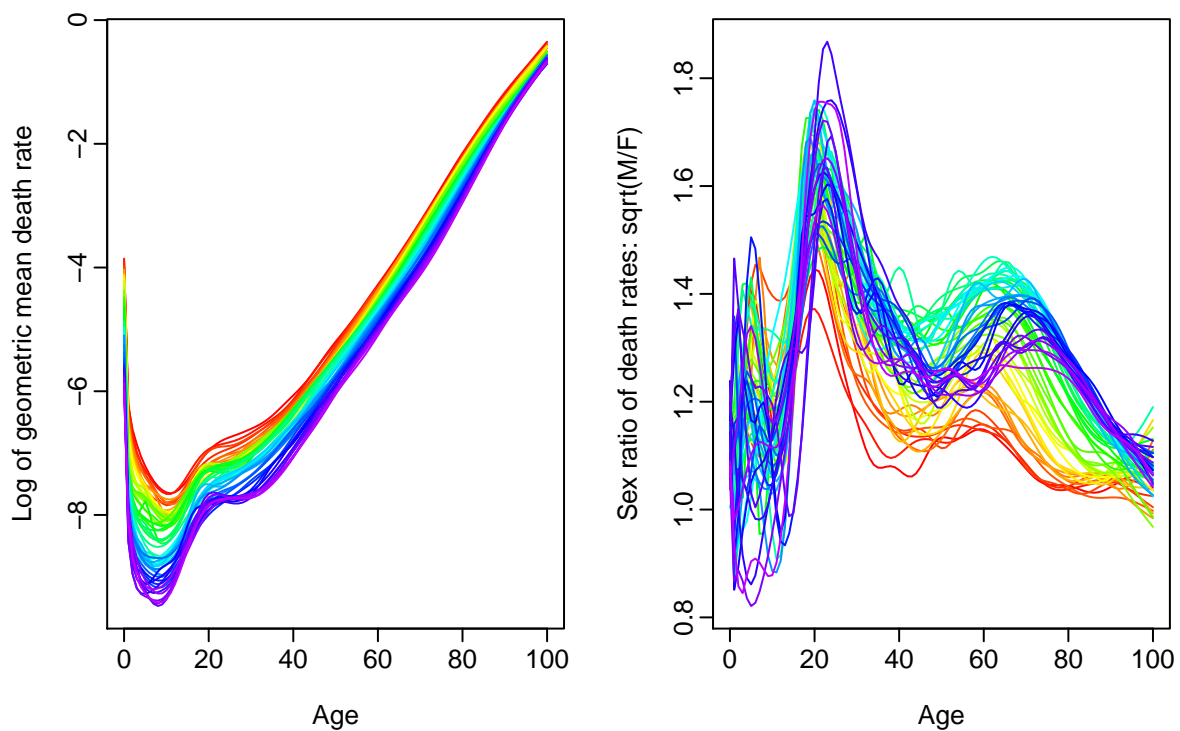
Figure 2: Product and ratio function of the coherent functional model for Swedish data (1950–2007).

one occur where small numbers of deaths result in large random variation — at very young ages in recent years and at very old ages in earlier years. Since the 1980s, the ratio function has declined at ages 40 to 80, and more recently for older ages. In recent years, large fluctuations occur at ages less than 10 and at 20 to 30.

Figure 3 show the basis functions and corresponding coefficients with thirty year forecasts, produced by applying functional time series models to the product and ratio functions. In the product model, the first principal component shows a declining trend in mortality, which is fastest at childhood ages and slowest at about age 20 and at very old ages. This term alone accounts for 97.0% of total variation. In the ratio model, the first principal component shows if anything a downward trend, but after 1990 fluctuations are substantial. This term accounts for 57.9% of total variation; the second and third terms account for 15.8% and 8.5% of total variation respectively. No cohort effects or other age patterns were observed in the residuals (i.e., one-step forecast errors) for the models.

Figure 4 shows thirty-year forecasts for sex ratios of male and female death rates using the coherent functional model and using independent functional time series models for separate male and female rates (following the approach of Hyndman & Booth 2008). The coherent
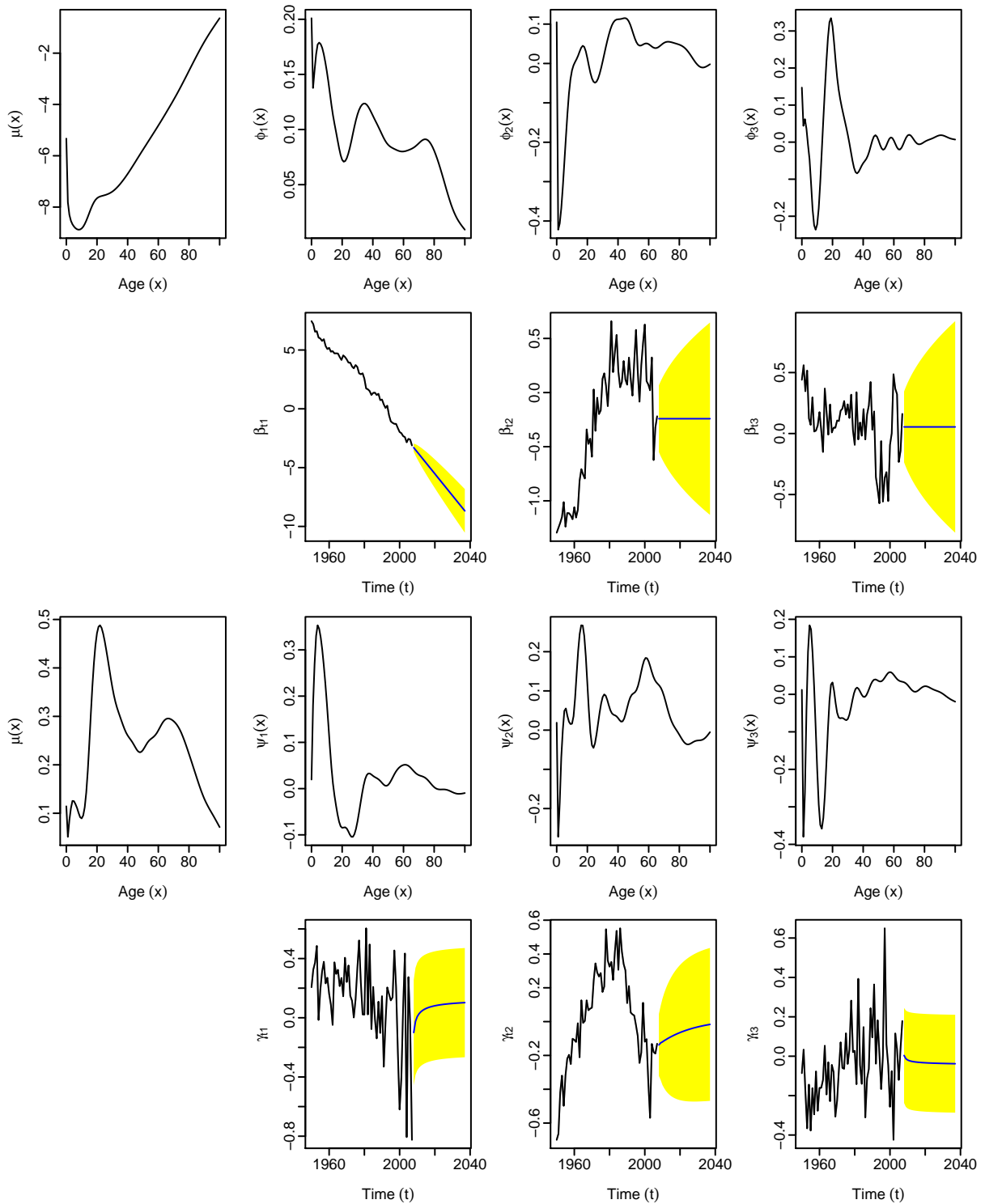
Figure 3: Top two rows: The first three basis functions and the corresponding coefficients of the product model for Swedish data along with thirty-year forecasts and 80% prediction intervals using ARIMA models. These components explain respectively 97.0%, 1.4% and 0.5% of the variation in the historical data. Bottom two rows: The first three basis functions and the corresponding coefficients of the ratio model for Swedish data along with thirty-year forecasts and 80% prediction intervals using stationary ARFIMA models. These components explain respectively 57.9%, 15.8% and 8.5% of the variation in the historical data.
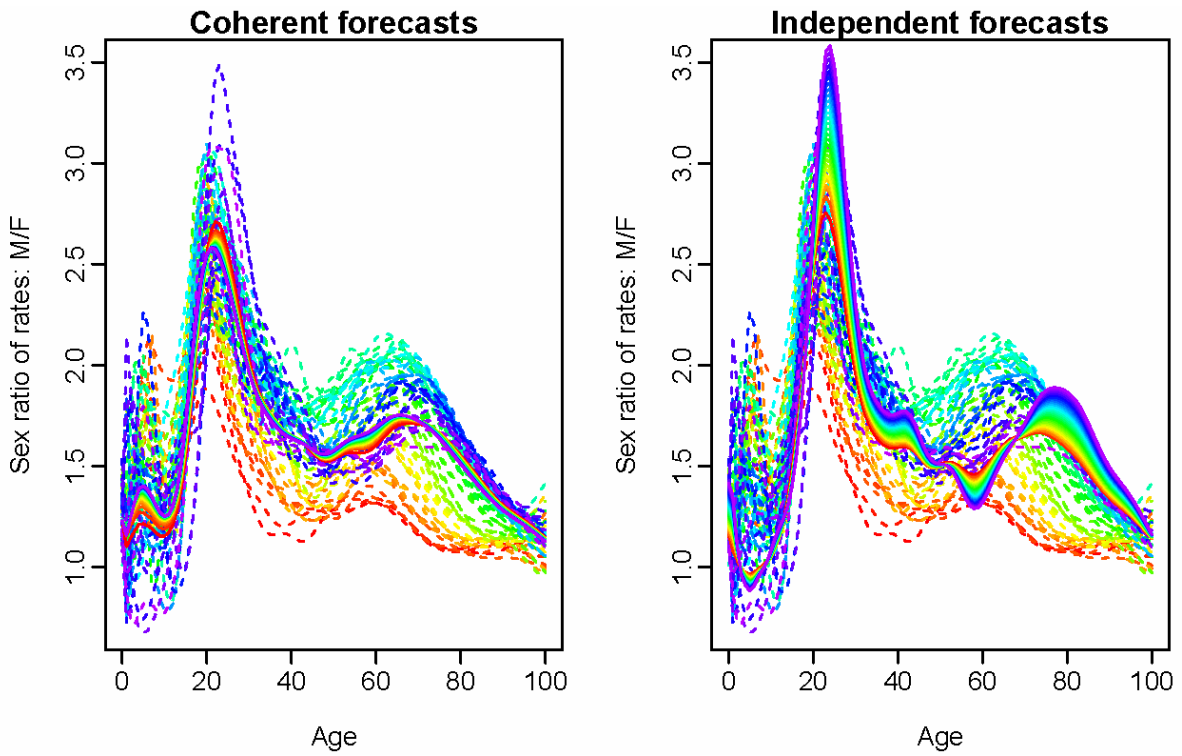
Figure 4: Thirty-year forecasts of mortality sex ratios in Sweden using coherent and independent functional models.
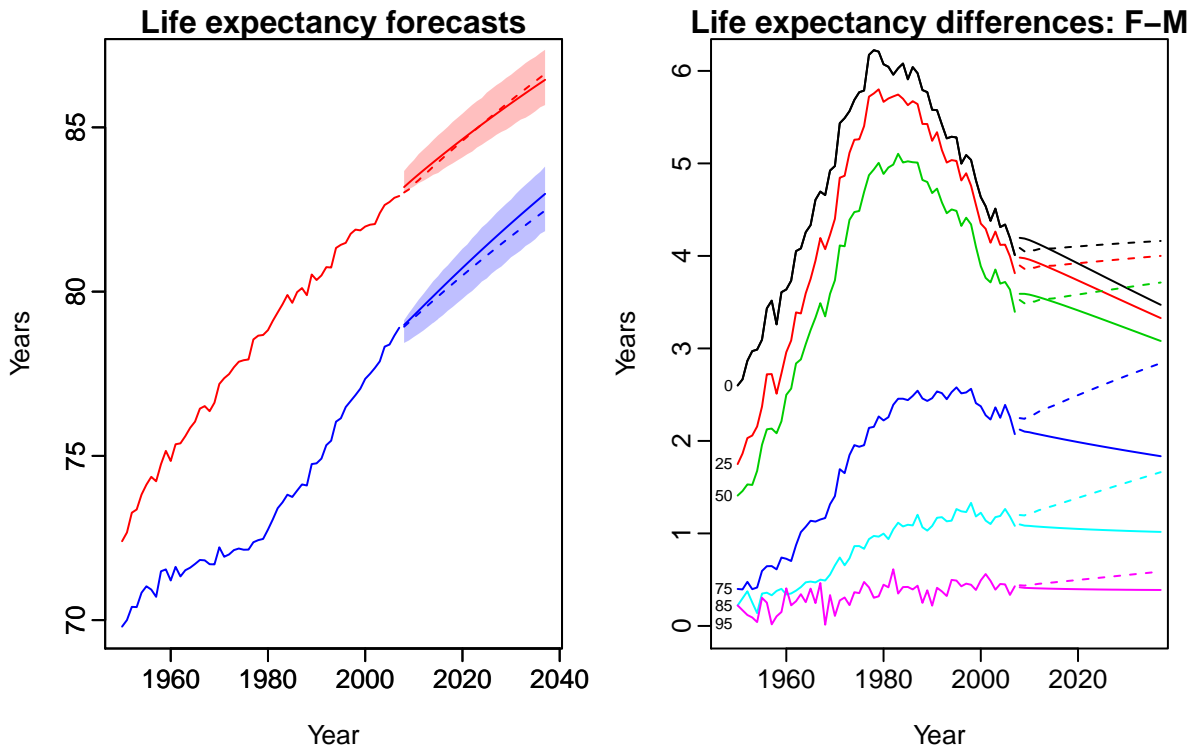

Figure 5: Thirty-year life expectancy forecasts for males and females in Sweden using coherent models (solid lines) and independent models (dotted lines). Blue is used for males and red for females.
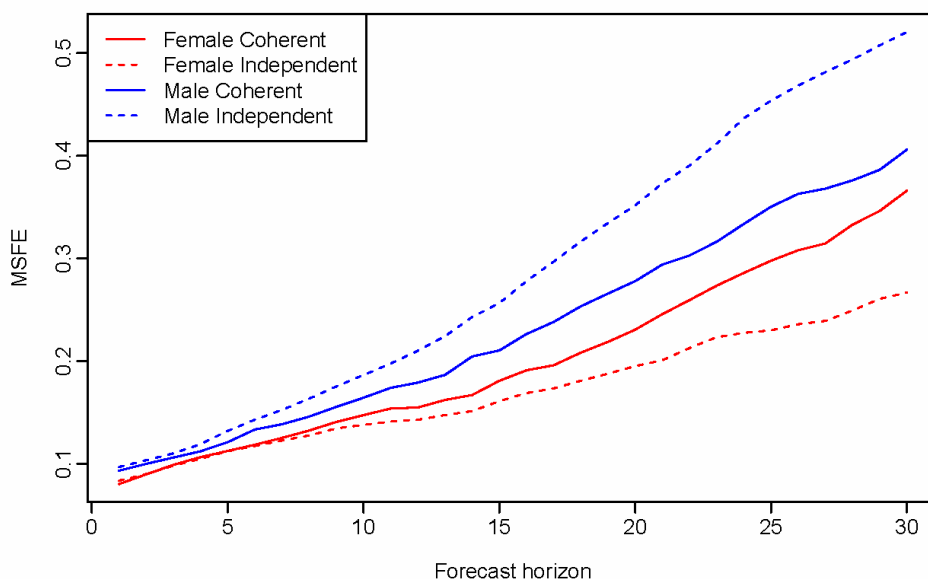
Figure 6: Forecast accuracy of coherent and independent models for different forecast horizons, showing out-of-sample MSFE of log death rates.

model produces sex ratios that are greater than one at all ages and remain within the limits of the observations on which they are based. In contrast, the independent models produce forecast sex ratios that are less than one at childhood ages and exceed the limits of observation at approximately ages 25 to 30 and 75+.

Thirty-year life expectancy at birth forecasts by sex are shown in Figure 5, along with the sex difference in forecast life expectancy at selected ages. The forecasts based on independent models (dashed lines) are diverging, a feature that goes against recent trends, and leads to sex gaps in life expectancy at older ages that exceed observation. In contrast, the sex differences embodied in the forecasts based on the coherent method continue to decrease, and all tend towards positive differentials (see Section 4.2). Thus, the coherent method produces sex differentials in both death rates and life expectancies that are in keeping with the levels and trends of the last 30 years.

The prediction intervals for life expectancies shown in Figure 5 are obtained from the percentiles of simulated life expectancy values, which in turn are obtained from simulated future death rates as described in Hyndman & Booth (2008). A feature of this method is that it takes account of the nonlinear relationship between age-specific death rates and life expectancy, thus producing asymmetric intervals.

To examine forecast accuracy, Figure 6 shows out-of-sample MSFE values of log death rates based on equation (9). It is seen that both the coherent and independent models are more

14

accurate for female than for male mortality, but that this sex difference is much smaller in the coherent case. The coherent model is more accurate than the independent model for male mortality, but less accurate for female mortality. The better performance for male mortality is attained at the expense of accuracy for female mortality, resulting in greater similarity in the degree of forecast accuracy for the two sexes. In overall terms, the coherent method performs marginally better (an average MSFE of 0.259 compared with 0.264 for the independent models). Thus, achieving coherence has not resulted in a reduction in overall accuracy.

3.2 Application to Australian state-specific mortality

In the second application, we consider the six states of Australia: Victoria (VIC), New South Wales (NSW), Queensland (QLD), South Australia (SA), Western Australia (WA) and Tasmania (TAS). The Australian Capital Territory and the Northern Territory are excluded from the analysis due to a large number of missing values in the available data. For each state, age-specific death rates and corresponding populations for 1950 to 2003 are obtained from Hyndman (2008). The ranking of life expectancy by state has changed over the fitting period; by 2003 the highest life expectancy occurred in Western Australia and Victoria and the lowest (by a margin of more than one year) in Tasmania.

Figure 7 shows the first three basis functions of the product model and their coefficients along with thirty year forecasts. The first principal component of the product model accounts for 96.5% of total variation. The first three principal components of the ratio model for Victoria are plotted in the bottom panels of Figure 7. The first term accounts for 37.0% of total variation, the second accounts for 13.5% and the third for 9.2%. For other states, the first term accounted for between 28.7% and 52.1% of the total variation. (Ratio models for other states are not shown.) No cohort effects or other age patterns were observed in the residuals for the models.

Figure 8 shows thirty-year life expectancy forecasts by state produced by the coherent and independent models. The life expectancy forecasts from the independent models are diverging, whereas the forecasts from the coherent model are more constrained. For the coherent forecasts, the ranking of the six states changes over the forecast horizon, though Western Australia and Tasmania maintain their 2003 positions as highest and lowest respectively. The overall differential between the highest and lowest forecast values and the differentials between states are consistent with observation.
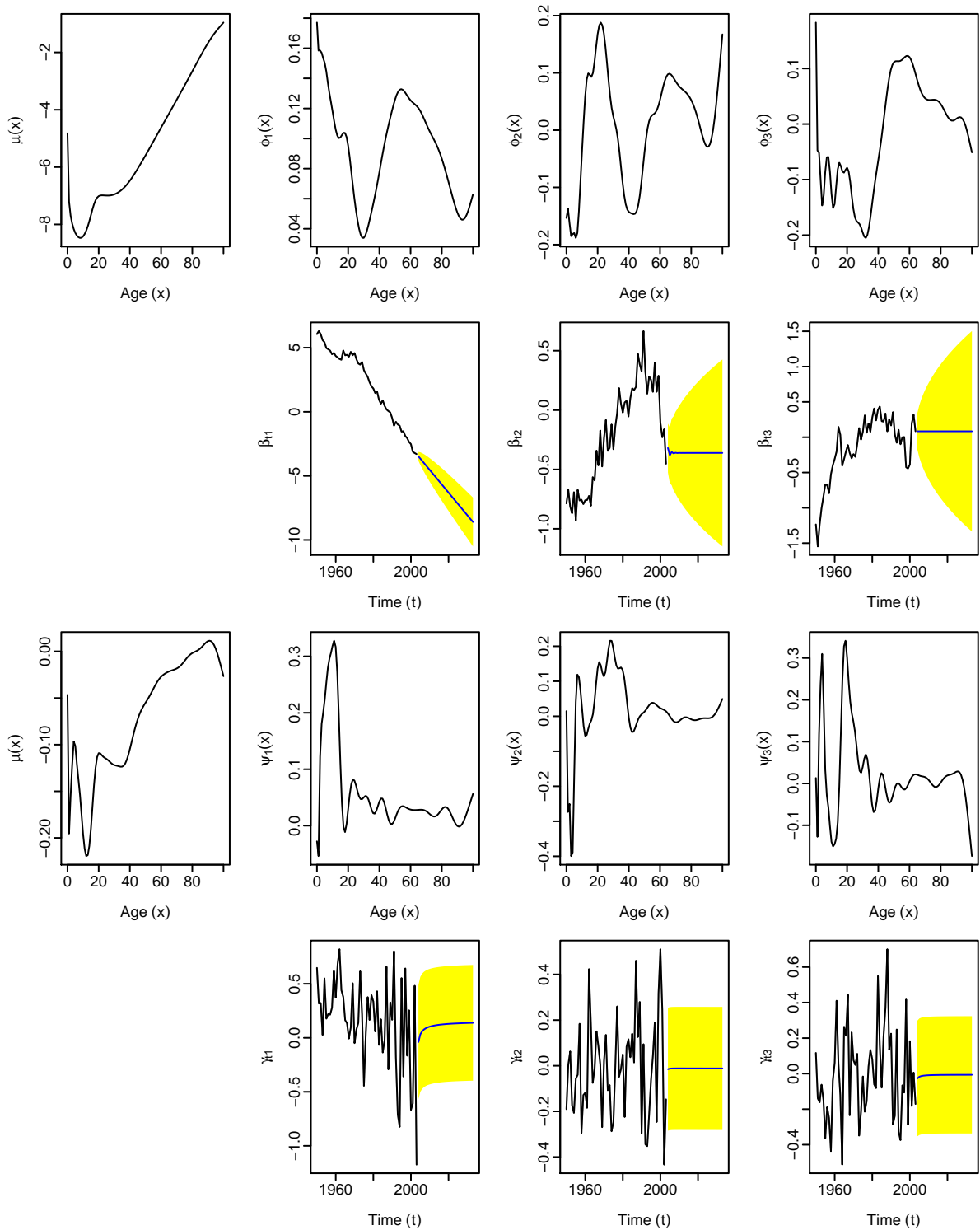
15

Figure 7: Top two rows: The first three basis functions and their coefficients for the product model given in (5a) for all states of Australia. Forecasts and 80% prediction intervals were obtained using ARIMA models. These components explain respectively 96.5%, 1.6% and 0.5% of the variation in the historical data.

Bottom two rows: The first three basis functions and their coefficients for the VIC ratio model given in (5b). Forecasts and 80% prediction intervals were obtained from ARFIMA models. These components explain respectively 37.0%, 13.5% and 9.2% of the variation in the historical data.
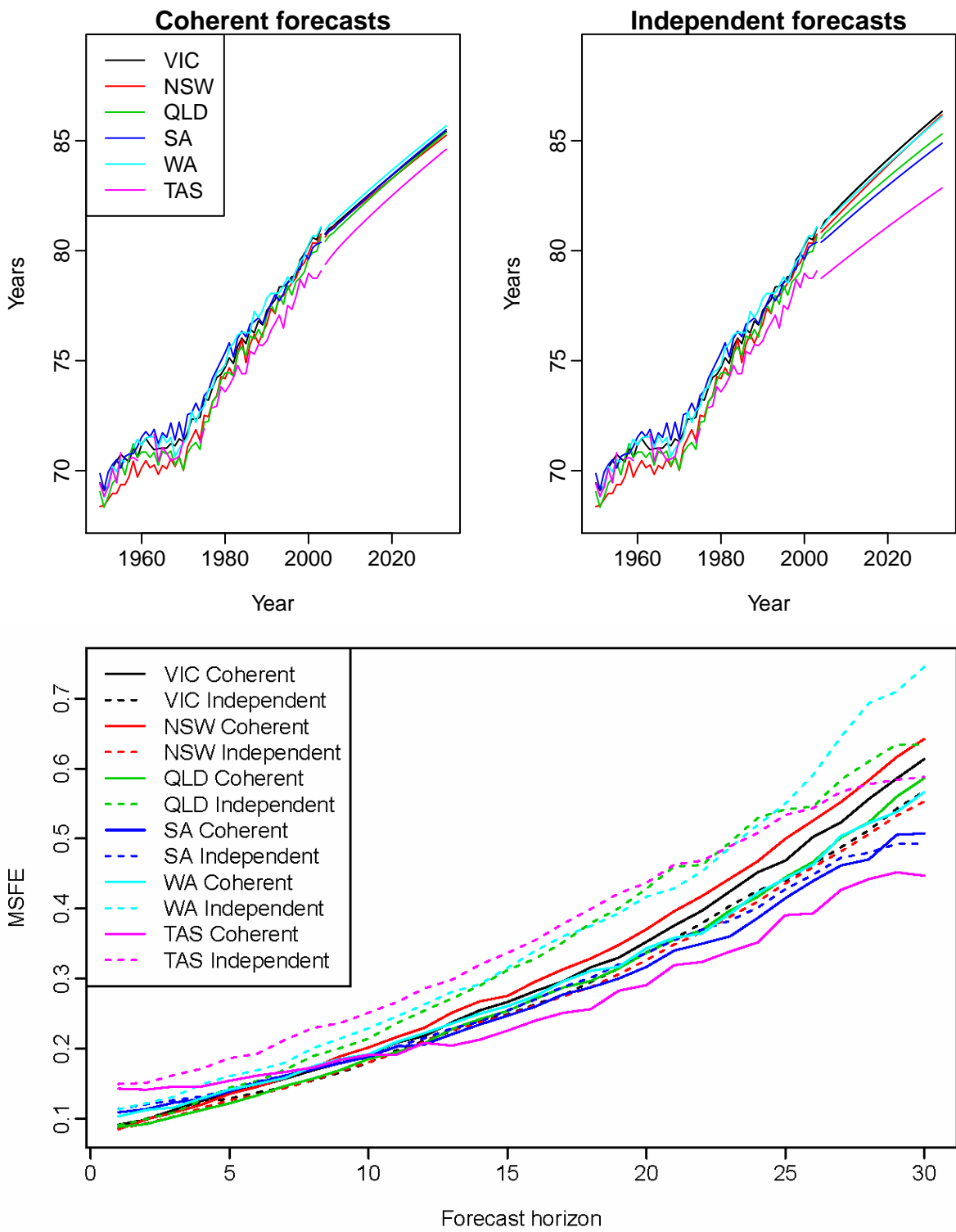
16

Figure 8: Top row: Thirty-year life expectancy forecasts for Australian states using the coherent and independent functional models.
Bottom: Forecast accuracy of coherent and independent models showing out-of-sample MSFE of log death rates for Australian states.

Figure 8 shows out-of-sample MSFE values using coherent and independent models for the Australian state data. The coherent model is more accurate at all horizons for four of the six state forecasts (the greatest loss of accuracy is for New South Wales). The increase in accuracy is greatest for Tasmania, which might be regarded as an outlier by 2003. On average, the coherent model performs slightly better (average MSFE of 0.325 compared with 0.346 for the independent models).

4 Discussion

In this paper, we have proposed a new method for coherently forecasting the age-specific death rates of two or more sub-populations. The product-ratio functional method is based on functional forecasting of simple functions of the products and ratios of the sub-population death rates, rather than the rates themselves. Using two-sex data for Sweden and six-state data for Australia, we have demonstrated the coherence of our results, while also demonstrating that the coherent method is at least as accurate in overall terms as independent forecasts.

4.1 Product-ratio method

The use of products and ratios is a simple and interpretable way of imposing coherence on the forecasts. The product function represents the geometric mean of sub-population death rates — a version of the familiar mortality schedule. The ratio function is the ratio of the death rates of a particular sub-population to the geometric mean rates; in the two-sex case, this is simply the square root of the sex ratio (and its inverse). Being a relative measure, the ratio function is easily understood and has the advantage of direct comparability over age (Figure 2).

The use of the product and ratio functions also greatly simplifies the modelling procedure compared with the direct use of rates because the two functions are uncorrelated. The resulting model is elegant in its simplicity. The new method ensures coherence in terms of long-term trends in forecast rates and produces results that remain within the observed range of variation for mortality ratios of sub-populations. The product-ratio method can handle outliers; in such cases the ratio function is larger than for other sub-populations and may have a different age pattern, but the forecast ratio function is kept within observed bounds.

A limitation of the product-ratio method is that the long-term forecasts of the ratio function will converge to the age-specific mean ratios, which are determined by the choice of fitting

18

period. We address this limitation in two ways. First the functional model that we employ incorporates weights that give greater weight to recent data (Hyndman & Shang 2009), so the long-term forecasts will converge to approximately the weighted mean ratios which are largely immune to the earliest data, increasing their relevance to the future. (In our applications we use a weight parameter of 0.05 which has the effect of giving weight 0.0500 to the most recent data, $0.05(0.95) = 0.0475$ to the year before, $0.05(0.95)^2 = 0.0451$ to the year before that, and so on.) However, we acknowledge that the weighted mean ratios may not be optimal at all ages and in all cases. Second, we use ARFIMA models for the coefficients of the ratio function; these models converge very slowly to the mean of the historical values of the coefficients. Tests were conducted to evaluate the possibility of basing the ratio function on a shorter fitting period, while maintaining the longer fitting period for the product function; this made almost no difference to average MSFE values, but destabilised the ratio forecast. The price of slow convergence using ARFIMA models is that a longer data series is needed for stable estimation. The full fitting period was therefore used for both functions. The use of the longer period for the ratio function may also be justified on conceptual grounds in that coherence implies that the subpopulation ratios are stochastically stable over time (that is, the age-specific ratios are assumed to be stationary).

An important advantage of the product-ratio method is greater flexibility in the age pattern of change in forecast mortality. When the product and ratio functions are combined to produce sub-population death rates, the age pattern of mortality embodied in the product function is modified in accordance with the age pattern of the ratio function. Where the ratio function is close to 1, there is little modification of the geometric mean death rate; where the ratio function differs from 1, the geometric mean rate is significantly modified (by multiplying by the ratio). The forecast product function is typically governed by an almost constant age pattern of change, because of a dominant first principal component. In contrast, several principal components are typically required to model the more complex ratio function, producing a variable age pattern of change over time. This variable pattern of change in the ratio function provides flexibility in the age pattern of change in mortality forecasts. The product-ratio functional approach thus allows for greater flexibility of change over time than the independent functional approach. Further, even if using only one principal component to model the product and the ratio functions — and hence fixed patterns of change for both — the product-ratio approach would allow flexibility of change in forecast mortality. The new approach allows for evolving ratios among forecast

sub-population rates (for example, sex ratios of forecast rates); this is in contrast to the Li and Lee approach where the ratios remain fixed over time (Li & Lee 2005, p582).

4.2 Coherence

The terms 'non-divergent' and 'coherent' have previously been used interchangeably (Li & Lee 2005). In developing the product-ratio method, we have more precisely defined coherence to be the convergence of the ratios of the forecast age-specific death rates from any two sub-populations to appropriate constants. In this paper we have used weighted historic mean ratios as the constants, but other values including theoretically-derived values could be used instead.

Coherence is the essential feature of the product-ratio method, and is ensured by the convergence of the ratio function to constant age-specific ratios. In the two-sex case, eventually-constant ratios are in keeping with their biological basis (Kalben 2002). In the multi-state case, the approach may be regarded as conservative in that constant ratios imply equal rates of mortality improvement in the long term.

Convergence to constant ratios does not imply that mortality differences between sub-populations tend to a set of constants. Trends in forecast differences depend on both the ratio and product functions. Only if both the forecast ratio function and forecast product mortality were constant, would sub-population differentials be constant. The forecast ratio function typically increases at some ages and decreases at others; the net effect is that differences in life expectancy may be convergent or divergent. As the ratios converge to their historic means, their ability to effect change diminishes. For constant ratios, changing product death rates produce differences between sub-populations that change at the same rate as product mortality. Thus, change in product mortality governs whether and how differentials converge or diverge in the long term. Convergence is constrained by the level of mortality and would occur only when death rates reach zero. Divergence and constant differences are similarly constrained, and for most populations are unlikely to be forecast given observed mortality trends.

For life expectancies at different ages, the age pattern of the mean ratios also influences forecast trajectories. This may result in widening differentials in life expectancy at certain ages even when mortality is declining at all ages. Thus, while forecast death rates are constrained to tend towards a defined relative differential, forecast life expectancies are less precisely constrained.

Figure 9 shows results for four additional populations (the US, Canada, UK and France), illustrating a range of outcomes. In all cases, the sex gap in forecast life expectancies from the coherent model is well-behaved over both time and age.
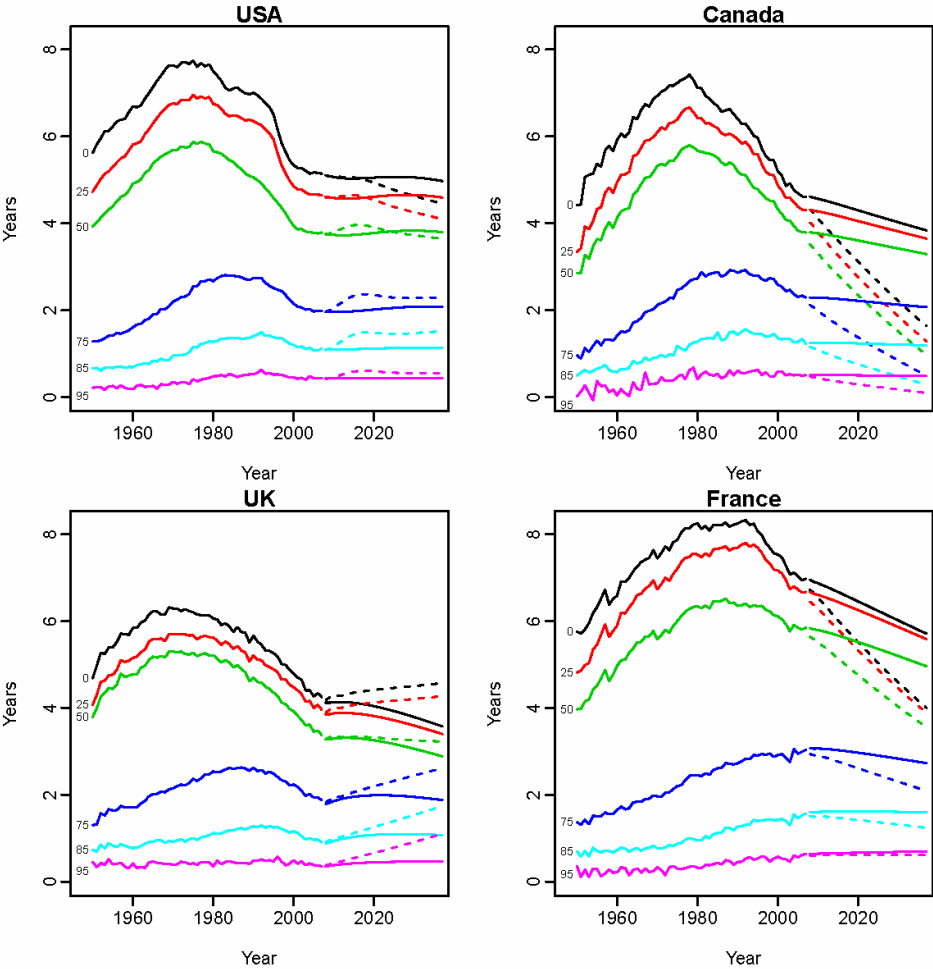

Figure 9: Differences in life expectancy forecasts for males and females (F–M) for 2008–2037 using coherent models (solid lines) and independent models (dotted lines).

4.3 Forecasts

The results for Sweden show a life expectancy at birth in 2037 of 86.4 years for females and 82.9 years for males. The forecast sex gap decreases from 4.0 years in 2007 to 3.5 years in 2037. Similarly for life expectancy at all other ages, the forecast sex gap decreases with slow convergence towards positive values.

The forecast life expectancies at birth in 2037 imply average annual increments of 0.113 years for females and 0.138 years for males. These are less than the average annual increments over the fitting period, implying deceleration in the rate of increase. For females, deceleration is

in fact observed over the fitting period (Figure 5), and is continued in the forecast. Bengtsson (2006) notes that deceleration has been a recent feature of several formerly leading countries of female life expectancy. For males, the linear trend since 1980 in observed life expectancy is not continued in the forecast.

While approximate linearity has been observed in the recent life expectancies of developed countries (White 2002), it is also the case that approximate linearity is observed in the coefficient of the first principal component of models of death rates (Lee 2006). When linearity is continued in the forecast of the coefficient, it does not produce linearity in forecast life expectancy. The main reason for this is that the principal component model assumes a fixed age pattern of change, whereas life expectancy takes into account the fact that the age pattern of change varies over time (Booth et al. 2002*b*). This is illustrated in Figure 10 for Swedish female mortality. The pessimistic forecast is based on a functional model using a single principal component with linear coefficient, and hence a fixed age pattern of change. The optimistic forecast is a random walk with drift model of life expectancy. This latter model is heavily dependent on the fitting period, and in particular on the first year of the fitting period. If we were to discount the first few years of the fitting period, when the trend is particularly steep, a more gradual increase in forecast life expectancy would result, but this would still be greater than the increase embodied in the single principal component model.

In the Swedish example, several additional interconnected factors contribute to the relatively slow and diminishing increase in forecast life expectancy. Change in the product function, which is the main determinant of the trend in forecast male and female life expectancies, is more closely linear than in either male or female mortality (as determined by the independent models). This contributes to the result that linear change (the first principal component seen in Figure 3) accounts for the very high percentage of variability after smoothing, which means that the functional forecast of the product term is effectively based on the single linear principal component, and the age pattern of change is fixed. Thus the forecast of product function mortality embodies a deceleration in the increase in life expectancy.

When there is almost total dominance of the first principal component in the product model, the forecast trend can be heavily dependent on the fitting period (especially if a random walk with drift model is used, which depends totally on the first year and last year of the fitting period). In the Swedish example, the ARIMA(0,1,1) model (with drift) used for the first coefficient somewhat reduces this dependence by giving extra weight to more recent years. Further, we
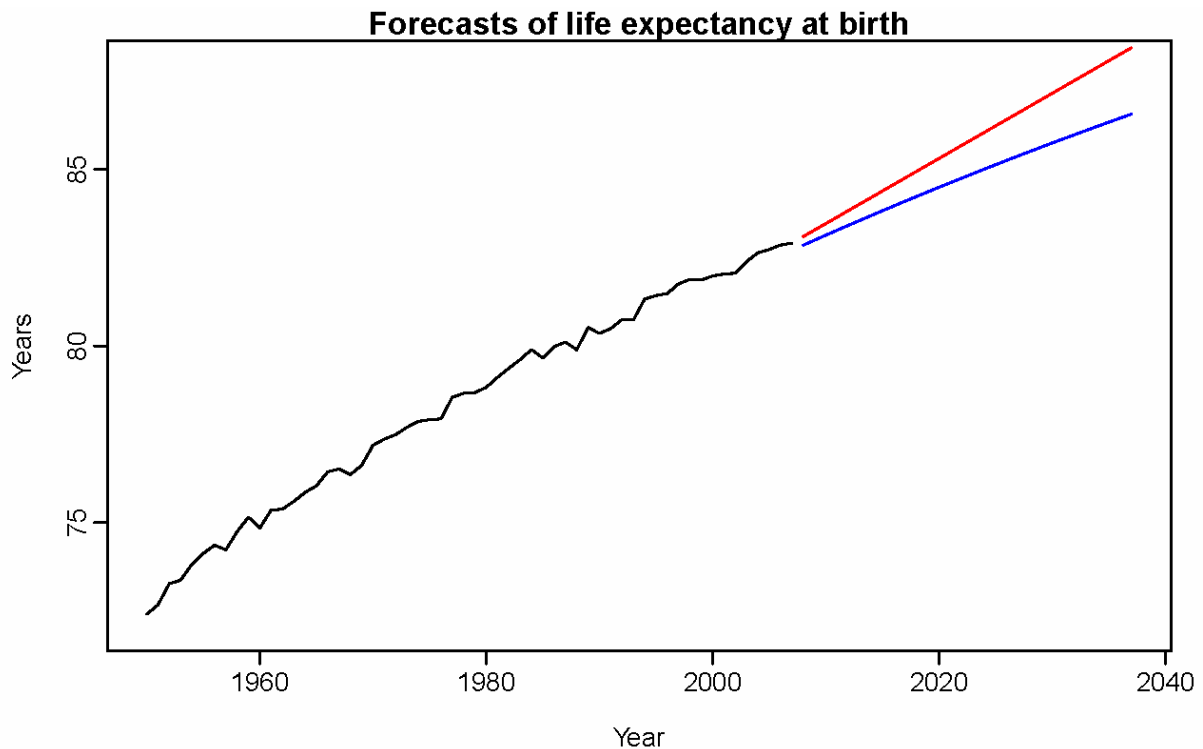
Figure 10: Forecasts of Swedish female life expectancy at birth from a random walk with drift model applied directly to life expectancy data (red line) and a single principal component with linear coefficient (blue line).

have used weighted functional methods to give greater weight to the more recent data and reduce dependence on the particular fitting period; however, this has little effect in the Swedish example because of linear change in the product function over the fitting period (Figure 3).

In other work, we have found the first principal component to account for about 94% of variation (Hyndman & Booth 2008), allowing some influence of the second and higher principal components in the forecast. The main effect of the additional principal components is to modify the age pattern of change embodied in the first principal component, so that the overall age pattern of change is not fixed. This has the effect of modifying the annual increments in forecast life expectancy.

For Sweden, the ratio model allows for greater flexibility over time in the age pattern of change. For example, one effect of the first principal component is the "shifting" of the primary mode of the ratio function to slightly older ages over the course of the fitting period, while the second component is largely responsible for the increase and then decrease in the secondary mode and also contributes to shifting it to older ages. At the oldest ages, there is little movement over time. In the forecast, the shifts are reversed because the pattern of change tends towards the weighted mean ratio function, but there is relatively little movement over the forecast horizon.

23

At childhood ages, the forecast involves an increase in the ratio — in other words, female rates are declining faster than male, and the female advantage is increased. At young adult ages (roughly 20 to 30), this trend is reversed — the female advantage is reduced. Very little change is forecast at about ages 30 to 45. At about ages 45 to 70, the female advantage is increased, while at older ages it is slowly reduced over time in line with recent trends whereby the male mortality disadvantage has diminished.

When the forecast ratio functions for male and female mortality are applied to the forecast product function, the resulting life expectancies are not equidistant from geometric mean life expectancy. Two factors contribute to this result. First, though symmetric on the log scale, the sex ratio and its inverse are asymmetric in their effects on product death rates: the larger effect occurs for males. Second, the effect of a constant ratio is greater for lower death rates. Both factors produce a greater difference from geometric mean life expectancy for male mortality than for female mortality. However, these differences are slight and the main determinant of the trends in male and female forecast life expectancy is the forecast of geometric mean mortality. In the Swedish example, this explains why male forecast life expectancy decelerates more rapidly than does female life expectancy.

The forecast life expectancies for Australian states show slow convergence of some pairs of states and slow divergence of others. Several trajectories intersect, for example NSW and Queensland. These results are the outcome of the combined influences of overall mortality decline, convergence of age-specific ratios to constant means, and the age pattern of these means, as discussed above. The average forecast rate of increase is less than the observed average. As in the case of Sweden, the dominant first principal component of the product function (a linear coefficient with fixed age pattern of change) produces slightly convex forecasts of life expectancy.

4.4   Forecast accuracy

It has been shown that the product-ratio functional method ensures coherence without compromising the overall (average) accuracy of the forecasts. In the two-sex case, the accuracy of the male mortality forecast was improved at the expense of accuracy in female mortality forecast; in other words, by adopting the coherent method, the accuracy of the forecasts for the two sexes was (partly) equalised. Similarly in the six-state example, forecast accuracy is more homogeneous in the coherent method, with an improvement in overall accuracy. This

feature of the method is useful in practical applications such as population forecasting, where it is preferable to maintain a balanced margin of error and hence a more balanced forecast population structure than might occur in the independent case. Similarly in financial planning, it may be preferable to accept moderate error in all sub-populations rather than risk a large error in one sub-population.

Coherent forecasting incorporates additional information into the forecast for a single sub-population. The additional information acts as a frame of reference — limiting the extent to which a sub-population forecast may continue a trend that differs from that of trends in related sub-populations. A similar approach has previously been adopted by Ortega & Poncela (2005) in forecasting fertility, as well as by Li & Lee (2005) for mortality.

The use of additional information raises the question of the choice of information. How important is it that the sub-populations be appropriately chosen? In the two-sex case, where data for both females and males are used to forecast mortality for each sex, the choice is pre-determined. Similarly, there is little doubt about the appropriateness of geographic sub-populations of a country. The grouping of several countries, however, is less clear (Li & Lee 2005). We note that the 'outlier' state of Tasmania has been successfully forecast, and similar results can be expected for countries that might be considered outliers in a larger region. The product-ratio method is relatively insensitive to the composition of the product function because of the simpler orthogonal relationship between the product and ratio functions (compared with the more complex correlated relationship between the group and individual rates used by Li and Lee).

Evidence suggests that the product-ratio functional method may produce more accurate forecasts than other methods. The evaluations of Booth et al. (2006) show that the functional data model of Hyndman & Ullah (2007) produces more accurate forecasts of death rates than the Lee-Carter method and its variants in 13 out of 20 populations. In this research we have used an improved functional data model which incorporates weights (Hyndman & Shang 2009), and has been shown to produce more accurate forecasts than any other method based on a principal components decomposition including the Lee-Carter variant used by Li and Lee (Shang et al. 2011). To compare our results for Sweden with those of Li & Lee (2005) we extended our forecasts to 2100: the product-ratio method produced a sex gap in life expectancy at birth of 2.2 years compared with 3.0 years by the Li-Lee method. Forecast life expectancy in 2100 was 91.3 years for females and 89.1 years for males, compared with approximately 93 and 90 years

respectively from Li & Lee (2005). Further research is needed to fully evaluate methods for coherent mortality forecasting.

4.5 Generalisation of the Li-Lee method

The product-ratio functional method can be viewed as a generalisation of the Li & Lee (2005) approach. Li & Lee (2005) employed the following model, expressed here for the two-sex case using notation consistent with that used in Section 2:

$$y_{t,j}(x) = \mu_j(x) + \beta_t \phi(x) + \gamma_{t,j} \phi_j(x) + e_{t,j}(x) \tag{10}$$

where $j \in \{F, M\}$, β_t is a random walk with drift, $\phi(x)$ is the first principal component of the common factor Lee-Carter model (based on combined male and female data), $\gamma_{t,j}$ is an AR(1) process, $\phi_j(x)$ is the first principal component of the residual matrix of the common factor model, and $e_{t,j}(x)$ is the error. The residual matrix is based on β_t after adjustment to match annual life expectancy (Lee & Miller 2001). Coherency is achieved by modelling $\gamma_{t,j}$ as an AR(1) process.

There are several points of difference between Li & Lee (2005) and the product-ratio functional method. First, Li & Lee (2005) use a single principal component in each model, while we use up to six (Hyndman & Ullah 2007) — though in practice only the first two have much effect for mortality. The additional principal components capture patterns in the data thatwhich are supplementary to the main trend and which may be significant. Booth et al. (2006, Table 3) showed that methods involving additional principal components were at least as accurate in 16 of 20 populations. The incorporation of additional principal components has the advantage that modelled death rates are no longer perfectly correlated across age (which is a feature of single component models). The use of the multiple principal components does not add significantly to the complexity of the model because the $\gamma_{t,\ell}$ coefficients ($\ell = 1, \ldots, L$) are uncorrelated by construction; there is thus no need to employ a vector approach.

Second, Li & Lee (2005) restrict the time series models to a random walk with drift for β_t and an AR(1) process for $\gamma_{t,j}$, whereas we allow more varied dynamics to be modelled through more general ARIMA processes for each $\beta_{t,\ell}$, $\ell = 1, \ldots, L$ and any stationary ARMA(p,q) or ARFIMA(p,d,q) process for each $\gamma_{t,\ell}$, $\ell = 1, \ldots, L$. Third, in the Li & Lee (2005) approach, the $\gamma_{t,j}$ coefficients are highly correlated with each other (over j), necessitating a vector approach

(Reinsel 2003) for modelling the $\gamma_{t,j}$ coefficients, which adds considerably to the complexity of the method.

Finally, by modelling $y_{t,j}$ instead of $f_{t,j}$, the Li & Lee (2005) model includes observational error which will be propagated in the forecasts. In the functional approach, observational error is removed by smoothing and added back in after modelling as part of variance estimation.

Thus the Li & Lee (2005) model can be considered a special case of the more general product-ratio functional method — the case where there is no smoothing, each model has only one component, and the time series models used are either a random walk with drift or a first order autoregression.

4.6   Extensions and practical application

A possible extension of the method would be to incorporate two or more dimensions in defining the product, so as to obtain coherency between and within each dimension; for example, sex and state might be used to coherently forecast male and female mortality within state, and mortality by state within each sex. A further possible but less elegant extension is the use of an external standard mortality schedule in place of the product, combined with ratios of sub-populations to this standard. This option may be useful where data for the appropriate group are missing. Best practice mortality (Oeppen & Vaupel 2002) may provide a useful standard. Further research is needed to examine these possibilities in greater detail.

Finally, it should be acknowledged that the development of improved forecasting methods for mortality, and by extension for fertility and migration (Hyndman & Booth 2008), represents a step towards more reliable and more easily automated demographic forecasting and the acceptance of these stochastic methods by national statistical offices responsible for producing official population projections. Though more complex than traditional methods, these methods are easily accessible through user-friendly code now available on CRAN (Hyndman 2010). Application of the methods is considerably simplified by this free software.

References

Bengtsson, T. (2006), Linear increase in life expectancy: past and present, *in* T. Bengtsson, ed., 'Perspectives on Mortality Forecasting III. The Linear Rise in Life Expectancy: History and Prospects', Social Insurance Studies No.3, Swedish Social Insurance Agency, Stockholm, pp. 83–99.

Booth, H., Hyndman, R. J., Tickle, L. & de Jong, P. (2006), 'Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions', *Demographic Research* **15**, 289–310.

Booth, H., Maindonald, J. & Smith, L. (2002*a*), 'Applying Lee-Carter under conditions of variable mortality decline', *Population Studies* **56**(3), 325–336.

Booth, H., Maindonald, J. & Smith, L. (2002*b*), 'Applying Lee-Carter under conditions of variable mortality decline', *Population Studies* **56**(3), 325–336.

Booth, H. & Tickle, L. (2008), 'Mortality modelling and forecasting: A review of methods', *Annals of Actuarial Science* **3**(1-2), 3–43.

Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. (2008), *Time series analysis: forecasting and control*, 4th edn, John Wiley & Sons, Hoboken, NJ.

Cairns, A., Blake, D. & Dowd, K. (2006), 'A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration', *Journal of Risk and Insurance* **73**, 687–718.

Cairns, A., Blake, D., Dowd, K., Coughlan, G. & Khalaf-Allah, M. (2008), 'Mortality density forecasts: An analysis of six stochastic mortality models', *Pensions Institute Discussion Paper PI-0801.* **64**.

Currie, I. D., Durban, M. & Eilers, P. H. (2004), 'Smoothing and forecasting mortality rates', *Statistical Modelling* **4**, 279–298.

Granger, C. & Joyeux, R. (1980), 'An introduction to long-memory time series models and fractional differencing', *Journal of Time Series Analysis* **1**, 15–39.

Greenshtein, E. (2006), 'Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint', *Annals of Statistics* **34**(5), 2367–2386.

Greenshtein, E. & Ritov, Y. (2004), 'Persistence in high-dimensional linear predictor selection and the virtue of overparameterization', *Bernoulli* **10**(6), 971–988.

Haslett, J. & Raftery, A. E. (1989), 'Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource (with discussion)', *Applied Statistics* **38**, 1–50.

Hosking, J. (1981), 'Fractional differencing', *Biometrika* **68**, 165–176.

Human Mortality Database (2010), University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). `www.mortality.org`. Downloaded on 14 October 2010.

Hyndman, R. J. (2008), *addb: Australian Demographic Data Bank*. R package version 3.222. `robjhyndman.com/software/addb`

Hyndman, R. J. (2010), *demography: Forecasting mortality, fertility, migration and population data*. R package version 1.07. Contributions from Heather Booth and Leonie Tickle and John Maindonald. `robjhyndman.com/software/demography`

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

Hyndman, R. J. & Khandakar, Y. (2008), 'Automatic time series forecasting: The forecast package for R', *Journal of Statistical Software* **26**.

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

Hyndman, R. J. & Shang, H. L. (2010), 'Rainbow plots, bagplots and boxplots for functional data', *Journal of Computational & Graphical Statistics* **19**, 29–45.

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.

Kalben, B. B. (2002), "Why Men Die Younger: Causes of Mortality Differences by Sex". Society of Actuaries.

Lee, R. D. (2000), 'The Lee-Carter method for forecasting mortality, with various extensions and applications', *North American Actuarial Journal* **4**(1), 80–92.

Lee, R. D. (2006), Mortality forecasts and linear life expectancy trends, *in* T. Bengtsson, ed., 'Perspectives on Mortality Forecasting: Vol. III. The Linear Rise in Life Expectancy: History and Prospects', Social Insurance Studies No.3, Swedish National Social Insurance Board, Stockholm, pp. 19–39.

Lee, R. D. & Carter, L. R. (1992), 'Modeling and forecasting US mortality', *Journal of the American Statistical Association* **87**(419), 659–671.

Lee, R. D. & Miller, T. (2001), 'Evaluating the performance of the Lee-Carter method for forecasting mortality', *Demography* **38**(4), 537–549.

Lee, R. & Nault, F. (1993), 'Modeling and forecasting provincial mortality in Canada'. Paper presented at the World Congress of the IUSSP, Montreal, Canada.

Li, N. & Lee, R. (2005), 'Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method', *Demography* **42**(3), 575– 594.

Oeppen, J. (2008), Coherent forecasting of multiple-decrement life tables: a test using Japanese cause of death data, Technical report, Universitat de Girona. Departament d'Informàtica i Matemàtica Aplicada. `hdl.handle.net/10256/742`

Oeppen, J. & Vaupel, J. W. (2002), 'Broken limits to life expectancy', *Science* **296**(5570), 1029–1031.

Ortega, J. A. & Poncela, P. (2005), 'Joint forecasts of southern European fertility rates with non-stationary dynamic factor models', *International Journal of Forecasting* **21**(3), 539–550.

Peiris, M. & Perera, B. (1988), 'On prediction with fractionally differenced ARIMA models', *Journal of Time Series Analysis* **9**(3), 215–220.

Preston, S. H., Heuveline, P. & Guillot, M. (2001), *Demography: measuring and modelling population processes*, Blackwell, Oxford.

Reinsel, G. C. (2003), *Elements of Multivariate Time Series Analysis*, second edn, Springer.

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

Shang, H. L., Booth, H. & Hyndman, R. J. (2011), 'Point and interval forecasts of mortality rates and life expectancy: a comparison of ten principal component methods', *Demographic Research* . to appear.

Shumway, R. H. & Stoffer, D. S. (2006), *Time Series Analysis and Its Applications: With R Examples*, 2nd edn, Springer, New York.

Tukey, J. W. (1977), *Exploratory Data Analysis*, Addison-Wesley, London.

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

Wilmoth, J. R. (1995), 'Are mortality projections always more pessimistic when disaggregated by cause of death?', *Mathematical Population Studies* **5**(4), 293–319.

Wood, S. N. (1994), 'Monotonic smoothing splines fitted by cross validation', *SIAM J. Sci Comput* **15**(5), 1126–1133.