COHERENT FORECASTING OF MORTALITY RATES: A SPARSE VECTOR-AUTOREGRESSION APPROACH

Hong Li, Yang Lu

Coherent Forecasting of Mortality Rates: A Sparse Vector-Autoregression Approach

Hong Li* Yang Lu†

October 29, 2016

Abstract

This paper proposes a spatial-temporal autoregressive model for the mortality surface, where mortality rates of each age depend on the historical values of itself (temporality) and the neighbouring ages (spatiality). The mortality dynamics is formulated as a large, first order vector autoregressive model which encompasses standard factor models such as the Lee and Carter (1992) model. Sparsity and smoothness constraints are then introduced, based on the idea that the nearer the two ages, the more important the dependence between mortalities at these ages. Our model has several novelties. First, it ensures that in the long-run, mortality rates at different ages do not diverge. Second, it provides a natural explanation of the so-called cohort effect without identifiability difficulties. Third, the model is easily extended to the multiple-population case in a coherent way. Finally, the model is associated with a closed form, non-parametric estimation method: the penalized least square, which ensures spatial smoothness of the age-dependent parameters. Using US and UK mortality data, we find that our model produces reasonable projected mortality profile in the long-run, as well as satisfying short-run out-of-sample forecast performance.

Keywords: cohort effect, spatial co-integration, penalized least square.

*hong.li@nankai.edu.cn. School of Finance, Nankai University, Tongyan Road 38, 300350, Tianjin, P.R.China.

†Corresponding author, luyang000278@gmail.com. Aix-Marseille School of Economics, Aix-Marseille University, France.

1 Introduction

Forecasting mortality rates has become an important task for insurance companies, pension funds, as well as policy makers, due to the continuous longevity improvement around the world. In most developing countries, the life expectancy increases at an average rate of 0.25 years per annum, and it is estimated that a three-year increase of the life expectancy spells a 50 % percent increase of the cost for various retirement systems[1].

Nevertheless, despite great recent advances in the literature of mortality modelling, several critical issues remain partially unsolved. First, most models decompose mortality improvements into the product of latent mortality trend factors, which are assumed to be non-stationary, and the corresponding age-specific loadings of these trends. As a consequence, the projected mortality rates at different ages will diverge in the long-run, unless all age-specific sensitivity coefficients are identical. For example, Li et al. (2013) note that the difference in the estimated age-specific loadings in the Lee-Carter model leads to increasingly large proportional differences in the projected death rates at adjacent ages. In particular, the projected mortality rates are implausibly low for infant and younger ages relative to older ages, which are inconsistent with the common belief that the mortality profile should vary smoothly and continuously across age. To ensure non-diverging mortality projection in the long-run, Li et al. (2013) propose to gradually rotate the age-specific loadings, in order for them be the same for the majority of ages at some point in the future. The rotation approach, however, depends on expert judgements rather than be data driven. The issue of diverging projected mortality rates applies also to other mortality models such as the Cairns-Blake-Dowd model [Cairns et al. (2006)] and the Age-period-cohort model [Renshaw and Haberman (2006)]. In recent years, modelling mortality rates of multiple populations has gained greater attentions [see, e.g., Li and Lee (2005); Dowd et al. (2011); Zhou et al. (2014); Hyndman et al. (2013)]. Most multi-population models satisfy a population coherence assumption, i.e., the projected mortality of different populations for the same age will not diverge in the long-run. However, in these multi-population models, projected mortality rates at different age may still diverge.

Second, although empirical works usually document a significantly better fit when

[1]Source: International Monetary Fund Global Financial Stability Report, April 2012.

a cohort effect is included [see e.g. Cairns et al. (2009)] along with the period effect, models with age, period, and cohort (APC) effects usually suffer from difficulties in terms of identifiability and interpretability [see e.g. Heckman and Robb (1985); Kuang et al. (2008); Hunt and Villegas (2015)]. Roughly speaking, different APC models can have the same in-sample fit, but produce substantially different mortality forecasts. This raises doubts on the reliability of such models for forecasting purpose.

Finally, while the widely used factor-based mortality models have rather intuitive structures, the statistical properties of the associated estimation method are not sufficiently understood. Indeed, they are typically non standard due to the non-stationarity of the mortality processes, which renders standard consistency results of, say, maximum likelihood estimation, invalid. In particular, recently, Leng and Peng (2016) show that the two-stage estimation method of the Lee-Carter model is inconsistent in some cases [see also Callot et al. (2016) for a discussion].

This paper addresses the above-mentioned issues within the framework of spatial-temporal autoregressive (STAR) process. The terminology "spatial" is borrowed from the spatial statistics and econometrics literature, which studies, roughly speaking, how time series of quantities observed at various places are correlated. While our paper does not take into account the *physical* distance between different countries in our analysis, our model has a similar spirit, in the sense that the spatial dependence in the space of different *ages* is explored. Thus the narrow interpretation of "spatiality" in our paper is the dependence of mortality between neighbouring ages. More precisely, our model emphasizes the similarities of mortality rates at *neighbouring* ages in terms of: 1) the long-run behaviour; *ii*) the short-term shock they receive; *iii*) the smoothness of the age-dependent parameters. Our starting point is a constrained, first-order Vector Autoregressive (VAR) model for the (log) mortality rates at different ages. The parsimony is achieved via appropriate sparsity constraints, which leads to a spatial dependence pattern, in the sense that the closer the two ages, the stronger the dependence between the mortality rates at these ages. Our model has several advantages. First, the constraints we introduce concern both the autoregressive matrix, which captures the Granger causality between mortality rates at different ages, and the variance-covariance matrix of the residual, which captures the instantaneous causality, that is the dependence between

residuals[2]. This improves the standard mortality literature, which usually omits the latter. Second, the model implies co-integrated mortality rates at different ages, with the same co-integration vector $(1, -1)$. In other words, the difference of mortality between any two different ages is stationary. Moreover, unlike common factor models, the long-run co-movement constraint is accompanied by a flexible short-term joint dynamics. Indeed, thanks to the spatial dependence pattern of our model, the long-run error correction mechanism takes effect *gradually* across the age range, which guarantees the flexibility of the short-term dynamics. Third, our model provides a natural interpretation of the so-called cohort effect, without introducing, in an *ad hoc* way, a separate cohort-specific factor, as is done in the existing literature. Fourth, the model can be readily extended to the simultaneous modelling of two or more populations, such as males and females in a same country, or a group of countries with similar socio-economic characteristics, when the inter-population dependence is introduced. Our model ensures that the (log-) mortality differences between any two ages in any two populations is stationary, with a age- and population-specific equilibrium mean. Finally, in terms of the statistical inference, in order to ensure smoothness of the age-dependent parameters of the model, we propose a two-stage, penalized least square (PLS) estimator. This approach is essentially non-parametric, and contributes to the growing literature on the use of flexible, functional time series to forecast mortality [see e.g. Hyndman et al. (2013); Li et al. (2015)]. We derive the *closed form* formula of the PLS estimate, and show how to determine the optimal choice of the penalization term via cross-validation.

Our model is applied to US and UK mortality data, and is compared to the benchmark models of Lee and Carter (1992) and Li and Lee (2005). In the empirical analysis, we find that our model outperforms these benchmarks in terms of forecast error, and is able captures the fact that the mortality improvements during the last decades has been faster than before.

The rest of the paper is organized as follows. Section 2 presents the model. Section 3 provides a simple extension of the model for multiple populations. Section 4 discusses the statistical inference. Section 5 applies the model to US and UK populations. Section 6 concludes. Technical details and large figures are gathered in Appendix.

[2]See Appendix 4 for the definitions of the Granger causality and the instantaneous causality.

2 A spatial-temporal autoregressive (STAR) model

2.1 The standard literature

The aim of this subsection is *not* to provide a systematic review of the literature on mortality models [see e.g. Cairns et al. (2008); Barrieu et al. (2012)], but rather to present the mainstream methodology, and discuss some of the downsides that our paper seeks to improve.

The Lee-Carter model. Let us denote by $m_{i,t}$ the crude death rate at age i and date t, and $y_{i,t}$ the log transformed death rate, $y_{i,t} = \log m_{i,t}$. For expository purpose we re-index the observable ages by $i = 1, 2, ..., I$, and the observable dates by $t = 1, 2..., T$.

In a seminal paper, Lee and Carter (1992) postulate that the (log) mortality rates at different ages are captured by a common factor, multiplied by age-specific sensitivity coefficients with respect to the common trend. More precisely, they assume that:

$$y_{i,t} = a_i + b_i k_t + \epsilon_{i,t}, \qquad \forall i, t, \tag{2.1}$$

where a_i is the time average of $y_{i,t}$'s, k_t is the common factor, and b_i is the age-specific sensitivity coefficient for age i. One inconvenience of this approach is that the mortality forecast at different ages diverge in the long run, unless the sensitivity coefficients b_i are equal for all ages. As a result, the projected mortality profile would be increasingly discontinuous. For example, assume $b_i < b_{i'}$ for some $i \neq i'$. Then the ratio $\frac{y_{i,t}}{y_{i',t}}$ increases unlimitedly when t goes to infinity.

Recently, Li et al. (2013) suggest to gradually rotate the sensitivity coefficients, so that these exposures would be equal for the majority of age at some point in the future. Nevertheless, this approach is rather *ad hoc* and does not allow a clear justification. Moreover, it is not clear how to generalize this approach to multi-population mortality models, and/or to take into account the cohort effect.

Models with cohort effect. The demographic literature has documented the existence of a cohort effect that is unexplained by typical age-period based mortality models, such as the Lee-Carter model. For instance, the so-called UK golden cohort(s), who are

born between 1925 and 1934, have enjoyed a faster mortality improvement compared to neighbouring cohorts, and this is widely credited to the better childhood health, as well as the war time rationing in their adolescent[3]. This phenomenon has motivated various extensions of the Lee-Carter model. For instance[4], Renshaw and Haberman (2006) propose:

$$y_{i,t} = a_i + b_i k_t + c_i \gamma_{t-i} + \epsilon_{i,t}, \tag{2.2}$$

where γ_{t-i} captures the cohort effect. One disadvantage of such extensions, despite their gain in terms of goodness-of-fit, is that they suffer from the identifiability difficulty, due to the simple relation that Calendar year = Birth date (that is the cohort Id) + current age of the cohort [see e.g. Heckman and Robb (1985), Kuang et al. (2008), Hunt and Villegas (2015)]. Roughly speaking, there exist more than one set of *observationally equivalent* parameters. This raises doubts of the suitability of such models for forecasting purpose, as the projected mortality rates depend on the (arbitrarily chosen) extra identification constraints.

Another downside of factor models is that due to the non-stationarity of the mortality rates, the consistency of maximum likelihood estimate on factor based models is a tricky issue. For instance, Leng and Peng (2016) show that, in some cases, the two-stage estimate of the Lee-Carter model is inconsistent.

2.2 The vector-autoregressive model and the Granger causality

In this paper, we propose a (sparse) vector autoregressive (VAR) model for the process $\boldsymbol{y}_t = (y_{1,t}, y_{2,t}, ..., y_{I,t})$. More precisely, let us assume that:

$$y_{1,t+1} = y_{1,t} + m_1 + \epsilon_{1,t}, \tag{2.3}$$

$$y_{2,t+1} = (1 - \alpha_2) y_{2,t} + \alpha_2 y_{1,t} + m_2 + \epsilon_{2,t}, \tag{2.4}$$

for the two initial ages, and

$$y_{i+1,t+1} = (1 - \alpha_{i+1} - \beta_{i+1}) y_{i+1,t} + \alpha_{i+1} y_{i,t} + \beta_{i+1} y_{i-1,t} + m_{i+1} + \epsilon_{i+1,t}, \tag{2.5}$$

[3]Source: Population Trends N.145, Autumn 2011, Office of National Statistics.
[4]See also Cairns et al. (2009), Plat (2009), etc.

for $i = 2, ..., I-1$ and $t = 1, ..., T-1$. In Equation (2.5), α_{i+1} and β_{i+1} are positive parameters that are smaller than 1, and the vector residual process $(\epsilon_{i,t})_t$ is a strong Gaussian white noise. That is, it does not feature temporal dependence and follows a Gaussian distribution $\mathcal{N}(0, \Sigma)$.

The necessity to specify different dynamic equations for $(y_{2,t})$ and $(y_{1,t})$ is due to the lack of observation of mortality rates for ages $i \leq 0$. However, we will show later that in our model, for each i, process $(y_{i,t})_t$ has a random walk dynamics. This, as a consequence, rationalises the specification (2.3) and (2.4).

The term $\alpha_{i+1} y_{i,t}$ on the RHS of (2.5) captures the cohort effect, that is the persistent effect of the mortality shocks on the same cohort, born in year $t-i$: the larger α_{i+1}, the more persistent the impact of the shocks. The term $(1 - \alpha_{i+1} - \beta_{i+1}) y_{i+1,t}$ can be either interpreted as the "learning" effect of the cohort $t-i$ from the (older) neighbouring cohort $t-i-1$; or as the period effect, that is, it captures the serial correlation of the mortality for the fixed age i. Finally, the term $\beta_{i+1} y_{i-1,t}$ captures the "learning" effect of the cohort $t-i$ from the (younger) neighbouring cohort born in $t-i+1$.

Thus the Granger causality between the mortality rates of different cohorts only exists between neighbouring cohorts, and within the same cohort. As we will show later, this assumption generalizes the common factor model [see e.g. Lee and Carter (1992)], while still keeping parsimony of the VAR model. Finally, the assumption that the three coefficients, $(1 - \alpha_{i+1} - \beta_{i+1})$, α_{i+1}, and β_{i+1} sum up to one ensures that the differences between different component processes of (y_t) are stationary [see Section 2.2.1].

Under this specification, the vector process $(y_{1,t}, ..., y_{I,t})_t$ has a (large, but sparse) first-order VAR representation:

$$\begin{bmatrix} y_{1,t+1} \\ y_{2,t+1} \\ y_{3,t+1} \\ \vdots \\ y_{I,t+1} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & \dots & \dots \\ \alpha_2 & 1-\alpha_2 & 0 & \dots & \dots \\ \beta_3 & \alpha_3 & 1-\alpha_3-\beta_3 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & 0 & \beta_I & \alpha_I & (1-\alpha_I-\beta_I) \end{bmatrix}}_{R} \begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \\ \vdots \\ y_{I,t} \end{bmatrix} + \begin{bmatrix} m_{1,t} \\ m_{2,t} \\ m_{3,t} \\ \vdots \\ m_{I,t} \end{bmatrix} + \begin{bmatrix} \epsilon_{1,t+1} \\ \epsilon_{2,t+1} \\ \epsilon_{3,t+1} \\ \vdots \\ \epsilon_{I,t+1} \end{bmatrix}, \tag{2.6}$$

where the (Granger causality) matrix R is lower triangular, and only has non null ele-

ments on the principal diagonal, as well as the first two lower sub-diagonals.

Such a constrained VAR can be regarded as a spatial-temporal autoregressive model [see e.g. Pace et al. (1998)], in the sense that different component processes $(y_{i,t})_t$ are smoothed not only on the time domain, but also on the "space", that is the age domain. As a consequence, we expect that for a given t, components $(y_{i,t})$ and $(y_{j,t})$ are closed so long as i and j are closed. Indeed, the shock $\epsilon_{i,t}$ enters, at time $t+1$, in the expressions of $y_{i,t+1}$, $y_{i+1,t+1}$, $y_{i+2,t+1}$; and enters, at time $t+2$, in the expressions of $y_{i,t+1}, y_{i+1,t+1}, y_{i+2,t+2}$, $y_{i+3,t+2}$ and $y_{i+4,t+2}$. More generally, it takes h years for a shock to reach out *progressively* to $2h+1$ different cohorts. As a consequence, although our VAR model is of first-order, it allows for the effect of the shocks to persist over several periods. This property is not satisfied by standard factor models, and explains the larger flexibility of our model with respect to the latter. Figure 1 illustrates the propagation of a mortality shock $\epsilon_{i,t}$ between different cohorts.

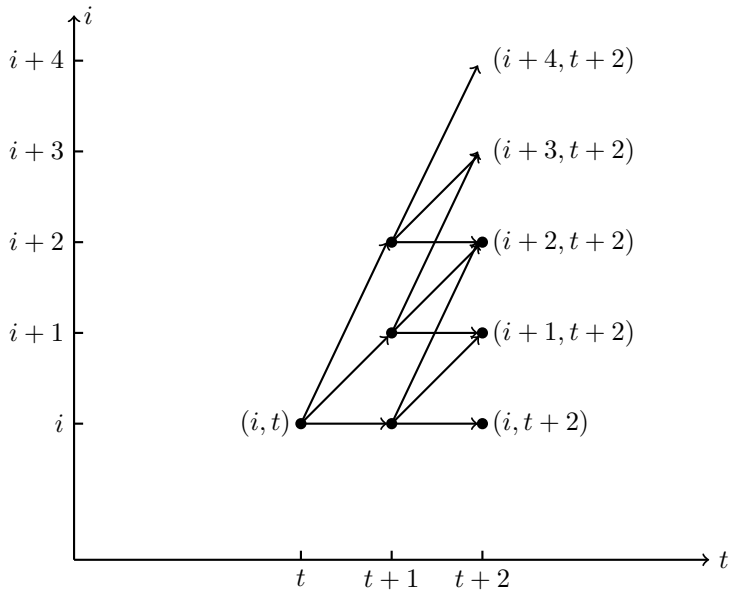


Figure 1: Propagation of the shock $\epsilon_{i,t}$ during the subsequent periods. At time $t+1$, ages i, $i+1$ and $i+2$ are impacted; at time $t+2$, ages i to $i+4$ are impacted.

Therefore, our model provides a natural interpretation of the cohort effect, which arises due to the persistence of shocks received by the same cohort during the early life of this cohort. Moreover, due to the learning effect between different cohorts, the cohort

effect is expected to be similar for neighbouring cohorts. To summarize, our framework permits correlation not only between cohort effects of different cohorts, but also between the cohort effects and past mortality shocks. These dependencies are typically ignored in standard mortality models, which specify the period and cohort factors *independently*, and usually do not specify the evolution of the cohort effect.

In mortality modelling, VAR models (or similar vector error correction models) have been suggested by, for example, Lazar and Denuit (2009) and Salhi and Loisel (2016). These authors propose to fit *unconstrained*, large dimensional vector autoregressive models to the time series of mortality rates. However, given the limited sample size of most mortality data, this approach suffers from the curse of dimensionality [see, e.g. Litterman (1986); Tibshirani (1996) for discussions]. Our constrained model mitigates this problem by incorporating only a small number of lagged regressors that are easily interpretable.

A VAR model has two major ingredients: the Granger causality matrix R in (2.6), which determines the long-run co-movement, or co-integration of the mortality rates, as well as the variance-covariance matrix of the residual, which captures the short-term dependence of the mortality rates. In the mortality literature, these two dependencies are usually captured by different models. For instance, Biffis and Millossovich (2006); Debón et al. (2008); Mavros et al. (2016); Chuliá et al. (2015) propose to decompose the mortality rates into a "central tendency" process, and a residual process. The first is assumed to follow a standard common factor mortality model such as the Lee-Carter, or the CBD model, whereas the residual is specified using similar spatial constraints. Our approach has several advantages compared to this conventional approach. First, factor models generically do not guarantee the coherence of different age-specific mortality processes $(y_{i,t})_t$. Second, such a decomposition is not necessarily unique. In particular, it is shown in the following examples that two of the most popular common factor models, that are the Lee-Carter model and the CBD model, can be represented as special examples of our constrained VAR model.

Example 1: Consider the case where all α_i, for $i = 2, 3, ..., I$ and β_i, for $i = 3, ..., I$ are equal to zero. Then we have a pure age-period model for the time series $(y_{i+1}, t)_t$, that is:

$$y_{i,t+1} = y_{i,t} + m_i + \epsilon_{i,t+1}, \qquad \forall i \geq 1.$$

In particular, if the residuals $(\epsilon_{i,t})$ satisfies the reduced rank condition:

$$\epsilon_{i,t} = m_i \epsilon_t, \tag{2.7}$$

where sequence (ϵ_t) does not depend on i, then we get

$$y_{i,t} = y_{i,1} + m_i \kappa_t,$$

where $\kappa_t = (t - 1 + \sum_{\tau=1}^{t} \epsilon_{i,\tau})$ is a random walk with drift. In other words we have obtained the Lee and Carter (1992) model.

If, instead of equation (2.7), residuals can be decomposed into:

$$\epsilon_{i,t} = \epsilon_{1,t} + i\epsilon_{2,t} \tag{2.8}$$

then we have

$$y_{i,t} = y_{i,1} + \kappa_{1,t} + i\kappa_{2,t},$$

which is CBD-type model [see e.g. Cairns et al. (2009)]. Thus we have shown that both the Lee-Carter model and the CBD model can be expressed as a special case of our VAR model. This motivates our modelling strategy compared to the decomposition approach of Debón et al. (2008); Mavros et al. (2016).

Example 2: Let us now consider the case where $\alpha_i = 1$ for each age $i = 2, 3, ..., I$, and all β_i are zero, for $i = 3, ..., I$. Then we get:

$$y_{i+1,t+1} = y_{i,t} + m_{i+1} + \epsilon_{i+1,t+1}, \tag{2.9}$$

or equivalently:

$$y_{i,t} = y_{1,t-i} + \sum_{\tau=1}^{t-1} m_{i+1-\tau} + \sum_{\tau=1}^{t-1} \epsilon_{i+1-\tau,t+1-\tau}$$

Thus $y_{i+1,t+1}$ only depends on the past residuals via $\epsilon_{i+1-\tau,t+1-\tau}$, which belong to the single birth cohort $t - i$. Thus we have an age-cohort model. Compared to the other stochastic models with cohort effect, for example, Renshaw and Haberman (2006), in Model (2.9), the cohort effect evolves as the time increases, instead of being fixed at the

birth of the cohort.

2.2.1 Analysis of co-movement

The introduction of the spatial dependence structure has an effect of "smoothing" the vector of mortality rates at a same, fixed time. Let us now discuss its implications for the long-run co-movement of the mortality rates at different ages. Roughly speaking, we want to ensure that in the future, mortality rates at different ages do not diverge. This is an important criteria for a reasonable mortality model, given that longevity risk is usually a very long-term risk.

Such a non-divergence property is called co-integration [see e.g. Engle and Granger (1987)] and has received much attention in the mortality literature [see Gaille and Sherris (2011); Li and Hardy (2011); Yang and Wang (2013); Li et al. (2013); Hyndman et al. (2013); Zhou et al. (2014); Hunt and Blake (2015b); Salhi and Loisel (2016)]. Given different ages i and j, processes $(y_{i,t})_t$ and $(y_{j,t})_t$ are co-integrated if: *i*) both processes are integrated of order one; *ii*) there exists a constant $\beta_{i,j}$ such that process

$$y_{i,t} - \beta_{i,j} y_{j,t} \tag{2.10}$$

is stationary.

In this subsection we will prove a stronger co-integration result:

Proposition 1. *Indeed under the specification of Section 2, different component processes $(y_{i,t})_t$ and $(y_{j,t})_t$ are co-integrated, with co-integration vector $(1, -1)$.*

Proof. First, equation (2.3) indicates that for the lowest age $i = 1$, process $(y_{1,t})_t$ is a random walk with drift m_1. Then we have, for age $i = 2$:

$$y_{2,t+1} - y_{1,t+1} = (1 - \alpha_2)(y_{2,t} - y_{1,t}) + m_2 - m_1 + \epsilon_{2,t+1} - \epsilon_{1,t+1},$$

in other words process $(y_{2,t} - y_{1,t})$ is stationary. Similarly, we have:

$$\begin{aligned} y_{3,t+1} - y_{2,t+1} =& (1 - \alpha_3 - \beta_3) y_{3,t} + \alpha_3 y_{2,t} + \beta_3 y_{1,t} - (1 - \alpha_2) y_{2,t} - \alpha_2 y_{1,t} \\ &+ m_3 - m_2 + \epsilon_{3,t+1} - \epsilon_{2,t+1} \\ =& (1 - \alpha_3 - \beta_3)(y_{3,t} - y_{2,t}) - (\beta_3 - \alpha_2)(y_{2,t} - y_{1,t}) \\ &+ m_3 - m_2 + \epsilon_{3,t+1} - \epsilon_{2,t+1}. \end{aligned}$$

Therefore, $\Big(1-(1-\alpha_3-\beta_3)L\Big)(y_{3,t}-y_{2,t})$ is stationary, where L is the lag operator. Since $(1-\alpha_3-\beta_3)$ lies between 0 and 1, $(y_{3,t}-y_{2,t})$ is also stationary for each i. By induction, we can show that each process $(y_{i+1,t} - y_{i,t})$ is stationary. Thus all the processes $(y_{i,t})_t$ follow a random walk with the same drift m_1. □

The co-integration property of our model ensures the projected mortality curves are non-divergent. At the same time, component processes are allowed to have flexible short-term dynamics, especially between remote ages, since it takes h periods for the impact of a shock $\epsilon_{i,t}$ to reach another cohort that is h years older (resp. younger).

While it is possible to conduct formal test of the assumption on the co-integration rank on the vector of mortality rates ($I - 1$ in model (2.3) to (2.5)) using Johansen's general methodology [see e.g. Lazar and Denuit (2009)], such tests are unlikely to have a high power due to data limitation. In particular, a typical dataset may contain 100 ages and fewer than 70 years of obervations. Therefore, while Lazar and Denuit (2009) found a co-integration rank of only $I - 3$ using an unconstrained co-integration test, we believe that the coherent property implied by our constrained model is valuable, since it can generate reasonable projections of future mortality rates. This remark also applies to other models with a similar spirit of imposing coherence, such as the Li-Lee model [Li and Lee (2005)] and the gravity model [Dowd et al. (2011)].

2.3 The spatial dependence of the residuals

Let us now discuss the specification of the residual (ϵ_t), which captures the instantaneous causality between different component processes $(y_{i,t})_t$. The standard literature assumes independence between its different components. To illustrate the necessity of accounting

for this dependence, we first simulate paths of the mortality processes which share the same causality matrix, but have different variance-covariance matrix Σ of the residuals. We will see that quite different patterns emerge when we alter the specification of this latter matrix. We then introduce a spatial autoregressive model for the joint distribution of the residual process.

2.3.1 An illustration by simulation

Let us now report, in a simulation experiment, the role of the cross-sectional dependence structure of the residual process (ϵ_t) on the dynamics of the observed mortality rates. For expository purpose, we only plot three component processes $\boldsymbol{y}_t = (y_{1,t}, y_{2,t}, y_{3,t})$.

Example 2 continued. Let us first re-consider Example 2, which has a pure age-cohort representation. We assume that $\alpha_2 = \alpha_3 = 1$, and $\beta_2 = \beta_3 = 0$, and $m_1 = 0.02$, $m_2 = m_3 = 0.06$. Firstly, we assume that $\epsilon_{1,t}, \epsilon_{2,t}, \epsilon_{3,t}$ are i.i.d. normal with $\mathcal{N}(0, 0.05^2)$. We plot in Figure 2 the path of $y_{1,t}, y_{2,t}$ and $y_{3,t}$, for periods $t = 1, ..., 100$.

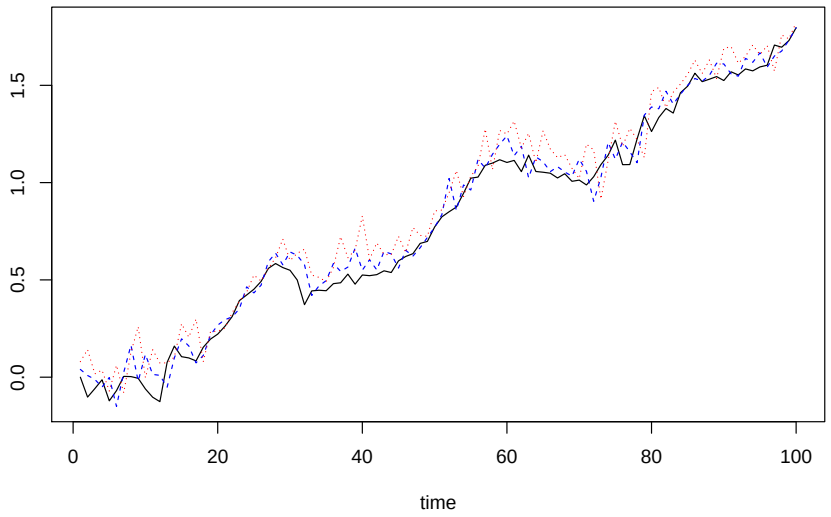


Figure 2: Simulated trajectory of y_1 (in black full line), y_2 (in blue dashed line), y_3 (in red dotted line), when $\alpha_2 = \alpha_3 = 1$, and $\beta_2 = \beta_3 = 0$, and the residuals are independent.

We can observe that it takes one (resp. two) periods for large shocks on $(y_{1,t})$ to propagate to $(y_{2,t})_t$ (resp. $(y_{3,t})_t$). We then keep the values of α_2, α_3, β_2, β_3 but assume

that the residuals are identical. The resulting simulated paths of $\boldsymbol{y}_t$ are plotted in Figure 3. We can see that the identical residuals make the co-movement between the three processes much more pronounced than the independent residuals shown in Figure 2.

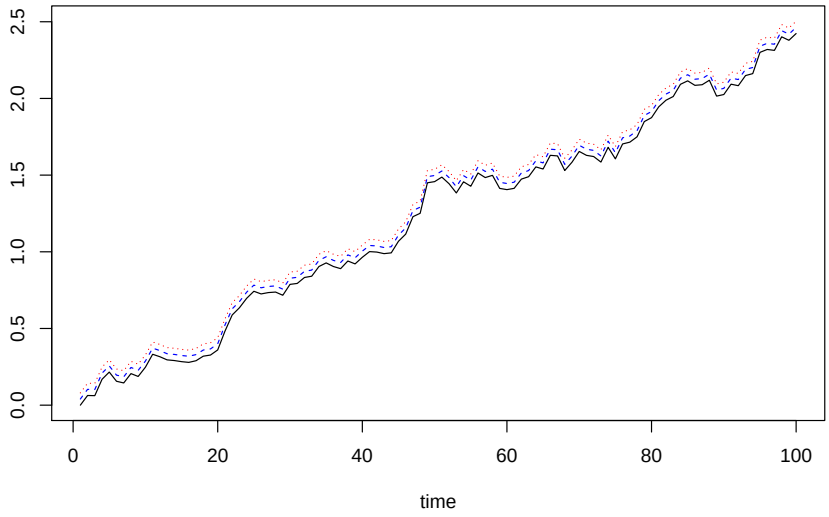


Figure 3: Simulated trajectory of y_1 (in black full line), y_2 (in blue dashed line), y_3 (in red dotted line), when $\alpha_2 = \alpha_3 = 1$, and $\beta_2 = \beta_3 = 0$, and the residuals are identical.

A similar case as Example 1. Let us now illustrate the impact of the residual structure on a model with both a period effect and a cohort effect. Figure 4 plots the simulated paths of $\boldsymbol{y}_t$, when the residuals are independent, but $\alpha_2, \alpha_3, \beta_2, \beta_3$ are close to 0. For illustration, we have chosen the value 0.05 for all of these four parameters. The reason of not choosing the exact value of zero is that in this limiting case, the co-integration relationship between the three time series no longer holds (except when the reduced rank condition of Example 1 is satisfied) and we will have instead three independent random walks.

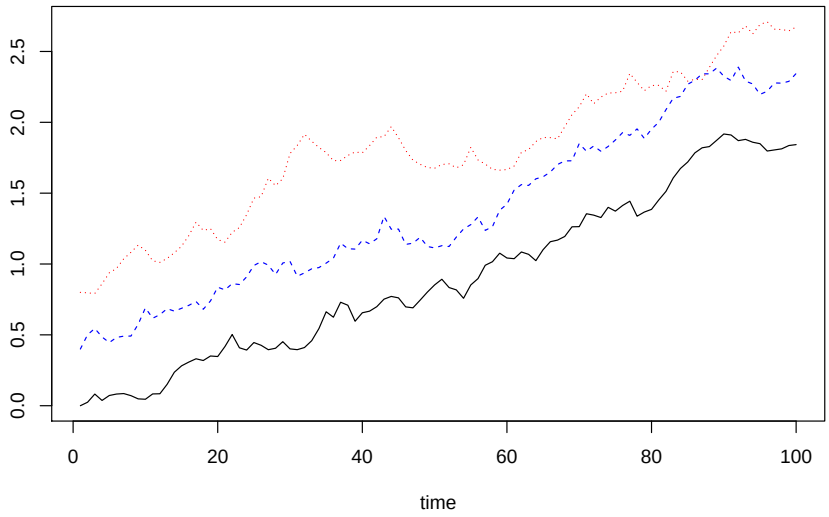


Figure 4: Simulated trajectory of y_1 (in black full line), y_2 (in blue dashed line), y_3 (in red dotted line), when $\alpha_2 = \alpha_3 = \beta_2 = \beta_3 = 0.05$, and the residuals are independent.

We can remark that compared to the Figure 2, the co-movement is much less pronounced. This is expected, since the inter-cohort transmission term α_i, β_i are all close to zero. In Figure 5, we keep the same assumption as in Figure 4, except that we assume the components of the residual are identical. Similar to the age-cohort representation, identical residuals lead to a much more pronounced co-movement between the three components than under the assumption of independent residuals.

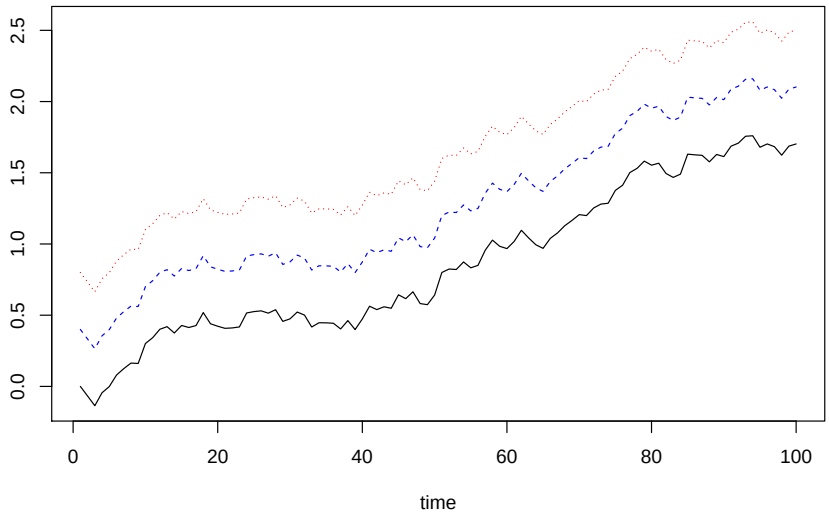


Figure 5: Simulated trajectory of y_1 (in black full line), y_2 (in blue dashed line), y_3 (in red dotted line), when $\alpha_2 = \alpha_3 = \beta_2 = \beta_3 = 0.05$, and the residuals are identical.

2.3.2 The model

Let us now propose a specification of the residual processes. There are two approaches to account for the dependence among residuals. The first one is to use parametric models with a *small* number of parameters to characterize the dependence structure. For instance, Biffis and Millossovich (2006) and Debón et al. (2008) propose a spatial random field model in which the covariance $\mathbb{E}[\epsilon_{i,t}\epsilon_{j,t}]$ is only a function of the distance $|i-j|$, but not of ages i and j. Such assumptions may, however, be too restrictive, as residuals at high ages tend to be much more erratic. Another approach is to leave the variance-covariance matrix completely unconstrained [see e.g. Mavros et al. (2016)]. The downside of the latter is the curse of dimensionality[5], since the number of coefficients to estimate is equal to $\frac{I(I+1)}{2}$, which is quite large so long as I reaches 30 or above.

As a trade-off between these two approaches, we propose a constrained spatial autoregressive model. More precisely, let us consider the following simultaneous equation

[5] A recent paper by Chuliá et al. (2015) do not specify the variance-covariance matrix, but the joint copula function, using a vine structure. Although this specification leads also to $O(I)$ parameters, the structure of the vine, such as which nodes are connected and which are not, should be chosen in a somehow *ad hoc* way.

model[6]:

$$\begin{aligned}
\epsilon_{I,t} &= c_I \epsilon_{I-1,t} + \eta_{I,t} \\
\epsilon_{I-1,t} &= a_{I-1}\epsilon_{I,t} + c_{I-1}\epsilon_{I-2,t} + \eta_{I-1,t} \\
\epsilon_{I-2,t} &= a_{I-2}\epsilon_{I-1,t} + c_{I-2}\epsilon_{I-3,t} + \eta_{I-1,t} \\
&\cdots\cdots \\
\epsilon_{1,t} &= a_1 \epsilon_{2,t} + \eta_{1,t},
\end{aligned} \tag{2.11}$$

where $\boldsymbol{\eta}_t = (\eta_{1,t}, \eta_{2,t}, ..., \eta_{i,t})$ is a Gaussian white noise whose components are independent and have variance σ_i^2 for each i, and parameters a_i, c_i are nonnegative and satisfy the stationary condition:

$$a_i + c_i < 1, \qquad \forall i = 1, 2, ..., I. \tag{2.12}$$

In other words, we assume that each component $\epsilon_{i,t}$ can be predicted by its neighbouring terms $\epsilon_{i+1,t}$ and $\epsilon_{i-1,t}$ alone. This is a first-order, (non-homogeneous) spatial autoregressive model, and is the analogy of the (spatial-temporal) autoregressive model specified in (2.3) to (2.5). A direct consequence is that given two components $\epsilon_{i,t}$ and $\epsilon_{j,t}$, the larger $|i-j|$, the looser the association between them[7].

This specification is more flexible than the parametric approach of Biffis and Millossovich (2006) and Debón et al. (2008), in the sense that parameters a_i, c_i are age-dependant and can capture a variety of patterns. In the empirical application, we also propose a non-parametric smoothing technique, which allows these parameters to feature smooth variation across different ages. At the same time, this specification involves $2 \times (I-1)$ parameters, which is significantly smaller than $\frac{I(I+1)}{2}$, that is the number of parameters, when the matrix Σ is left unconstrained.

[6]The terminology "simultaneous equation" is in the sense that ϵ_t's appear on both sides of the equations.

[7]However, in general, even when $|i-j| \neq 1$ or 0, the correlation between $\epsilon_{i,t}$ and $\epsilon_{j,t}$ is non zero. Indeed, intuitively, $\epsilon_{i,t}$ and its second-order neighbouring $\epsilon_{i+2,t}$ are dependent since both terms are dependent with $\epsilon_{i+1,t}$

In matrix terms, we have:

$$\underbrace{\begin{bmatrix} 1 & -c_I & 0 & \dots & \dots \\ -a_{I-1} & 1 & -c_{I-1} & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & 0 & -a_2 & 1 & -c_2 \\ \dots & \dots & 0 & -a_1 & 1 \end{bmatrix}}_{M} \epsilon_t = \eta_t, \tag{2.13}$$

where matrix M is invertible under assumption (2.12)[8]ãĂĆ Therefore, the previous simultaneous system can be rewritten in the reduced form:

$$\epsilon_t = M^{-1}\eta_t,$$

and the variance-covariance matrix of ϵ_t is equal to:

$$\Sigma = \mathbb{E}[\epsilon_t \epsilon_t'] = (M^{-1})' \text{Diag}(\sigma_1^2, \sigma_2^2, ..., \sigma_I^2) M^{-1}.$$

Note that although only elements on the principal diagonal and the upper and lower sub-diagonals of M are non null, in general all the elements of Σ are non null. However, for suitable values of a_i, c_i and σ_i, we can expect that the larger $|i-j|$, the smaller the covariance $\mathbb{E}[\epsilon_{i,t}\epsilon_{j,t}]$.

3 A multi-population extension

The aim of this section is to show how the spatial-temporal autoregressive model introduced in Section 2 can be extended to the simultaneous analysis of several populations. Such joint models are quickly gaining popularity in the literature due to: *i*) The need for multi-national insurers to evaluate the financial benefits of a geographically diversified portfolio of policyholders [see e.g. Hunt and Blake (2015b)]; *ii*) The need for policy mak-

[8]Indeed matrix M is diagonally dominant. Note that this condition is sufficient but not necessary. Nevertheless, we have chosen to impose it since *i*) it is the stationary condition in spatial autoregressive models, when the cross-sectional dimension I goes to infinity. Roughly speaking, it means that neighbouring observations $\epsilon_{i+1,t}$ and $\epsilon_{i-1,t}$ can only *partially* predict the value of $\epsilon_{i,t}$. *ii*) In the empirical application, we can check that this condition is satisfied.

ers to conduct demographic forecasting on a, say, region-by-region basis. For expository purpose, in this section, we first discuss the two-population case, and generalize the idea to situations with more than two populations.

Let us denote by $y_{1,i,t}$ and $y_{2,i,t}$ the log of the mortality rates of the two populations 1 and 2, and assume the following joint dynamic specification:

$$y_{j,i+1,t+1} = \rho_{j,i+1}\Big[(1-\alpha_{j,i+1})y_{j,i+1,t} + \alpha_{j,i+1}y_{j,i,t}\Big] + (1-\rho_{j,i+1})y_{-j,i+1,t} + m_{j,i+1} + \epsilon_{j,i+1,t}, \tag{3.1}$$

and

$$y_{j,1,t+1} = \rho_{j,1}y_{j,1,t} + (1-\rho_{j,1})y_{-j,1,t} + m_{j,1} + \epsilon_{j,1,t}, \tag{3.2}$$

with $j = 1, 2$, and $-j = 1$ when $j = 2$ and vice versa, $(m_{j,i})_i$-s are constants, and $(\epsilon_{j,i,t})_t$ are Gaussian white noises.

Thus compared to the single population model (2.5), the learning effect from younger cohort, that is $\beta_{j,i+1}y_{j,i-1,t}$, is replaced by a learning effect from the other population[9].

Let us discuss some special cases. **Example 1**: if parameters $\rho_{1,i+1}$, $\rho_{2,i+1}$ are equal to 1 for all values of i, then there is no cross causality in Equation (3.1). In this case we get two separate spatial-temporal autoregressive models[10].

Example 2: if, say, $\rho_{1,i+1} = 1 \neq \rho_{2,i+1}$ for all i, then Population 1 is the dominant population and Granger causes the mortality rates of population 2, whereas Population 2 is a smaller population such as a sub-population of Population 1. This corresponds, roughly speaking, to the gravity model [see Cairns et al. (2011); Dowd et al. (2011)].

In general, when $\rho_{1,i+1}$, $\rho_{2,i+1}$ are within $(0, 1)$ for all i, these two parameters capture the relative importance of one population's own past with respect to the past of the other population, when determining the future mortality rates.

[9]It would clearly be possible to retain also the learning effect from the younger cohort. In this case we would have a model of the form, say:

$$y_{1,i+1,t+1} = \rho_{1,i+1}\Big[(1-\alpha_{1,i+1}-\beta_{1,i+1})y_{1,i+1,t}+\alpha_{1,i+1}y_{1,i,t}+\beta_{1,i+1}y_{1,i-1,t}\Big]+(1-\rho_{1,i+1})y_{2,i+1,t}+m_{1,i+1}+\epsilon_{1,i+1,t},$$

however, empirically we found that our current version has nearly the same prediction power as its extension.

[10]However, the residuals of the two populations can still be correlated.

Similarly as the single population case, we have the following property concerning the co-movement of mortality of the two populations:

Proposition 2. *The difference of any two components of the* $(2\times I)-$*dimensional process* $(y_{1,i,t}, y_{2,i,t}, i = 1, 2, ..., I)$ *is stationary.*

Proof. See Appendix 1.1. □

Therefore, our specification ensures both within-, and inter-population co-integration. At the same time, the dependence of parameters $\rho_{1,i}$, $\rho_{2,i}$ on age i provides flexibility when it comes to capturing potential short-term mortality deviation at certain ages between the two populations. This necessity of reconciling the desirable long-term property and the empirical data is an important criteria for mortality models [see e.g. Hunt and Blake (2015a) for a discussion]. Our model provides such a trade-off.

The specification of the residuals is also extended to allow for inter-population dependence. In particular, let $\epsilon_{1,i,t}$ and $\epsilon_{2,i,t}$ be the residuals from population 1 and 2 for age i and time t, the dynamics of the residual is given by:

$$\begin{aligned}
\epsilon_{j,I,t} &= \theta_{j,I} c_{j,I} \epsilon_{j,I-1,t} + (1-\theta_{j,I})\epsilon_{-j,I,t} + \eta_{j,I,t} \\
\epsilon_{j,I-1,t} &= \theta_{j,I-1}\Big(a_{j,I-1}\epsilon_{j,I,t} + c_{j,I-1}\epsilon_{j,I-2,t}\Big) + (1-\theta_{j,I-1})\epsilon_{-j,I-1,t} + \eta_{j,I-1,t} \\
\epsilon_{j,I-2,t} &= \theta_{j,I-2}\Big(a_{j,I-2}\epsilon_{j,I-1,t} + c_{j,I-2}\epsilon_{j,I-3,t}\Big) + (1-\theta_{j,I-2})\epsilon_{-j,I-2,t} + \eta_{j,I-2,t} \qquad (3.3) \\
&\cdots\cdots \\
\epsilon_{j,1,t} &= \theta_{j,1} a_{j,1}\epsilon_{j,2,t} + (1-\theta_{j,1})\epsilon_{-j,1,t} + \eta_{1,t}.
\end{aligned}$$

A J-population extension with $J \geq 3$. Let us denote by $y_{j,i,t}$ the mortality of population j, $j = 1, ..., J$, at age x and time t. Then a natural idea is to generalize the two-population model [see equation (3.1)] into:

$$y_{j,i+1,t+1} = \rho_{j,j,i+1}\Big[(1-\alpha_{j,i+1})y_{j,i+1,t} + \alpha_{j,i+1}y_{j,i,t}\Big] + \sum_{k\neq j}\rho_{j,k,i+1}y_{k,i+1,t} + m_{j,i+1} + \epsilon_{j,i+1,t}, \qquad (3.4)$$

and

$$y_{j,1,t+1} = \sum_{k=1}^{J} \rho_{j,k,1} y_{j,k,t} + m_{j,1} + \epsilon_{j,1,t}, \tag{3.5}$$

where coefficients $\rho_{j,k,i}$ are positive and sum up to 1, that is, $\sum_{k=1}^{J} \rho_{j,k,i+1} = 1$ for each population j and age $x+1$.

Let us now discuss the co-integration in the multi-population case. The following proposition says that the mortality rates are still co-integrated.

Proposition 3. *The difference of any two components of the* $(I \times J)-$*dimensional process* $(y_{j,i,t}, j = 1, ..., J, i = 1, 2, ..., I)$ *is stationary.*

Proof. See Appendix 1.2. □

These three Propositions [1, 2 and 3] constitute the core contribution of our paper. They show that in order to ensure the co-integration of different mortality processes, it is no necessary use a common factor model such as Lee-Carter, or an error correction model [such as the gravity model Cairns et al. (2011)]. Rather, the family of VAR models offer much more flexibility, even with a rather sparse autoregressive matrix.

4 Statistical inference

In practice, the value of I lies between 40 and 100, thus it is computationally cumbersome to conduct maximum likelihood estimation. In this section we propose a two-stage penalized-least square approach. In the first stage, parameters of the autoregression function, α_i, β_i, and m_i, are estimated. These estimates are then plugged into the initial equation to recover the empirical residuals, which are used in the second stage to estimate the variance-covariance matrix. We begin with the simple, benchmark Ordinary Least Square (OLS) estimation, and then generalize to the Penalized Least Square (PLS) estimation, which provides smooth curves of the parameters along the age dimension.

The two-stage approach is similar to the method of Lee and Carter (1992), in the sense that both are motivated by the existence of a closed form estimate[11]. Although a

[11] In the Lee-Carter model, in the first stage, parameters can be estimated via a single value decomposition.

one-stage estimation would be more efficient, it falls well beyond the scope of the present paper since it poses much more computational challenges due to the large number of parameters.

4.1 The benchmark OLS of the Granger causality parameters

Let us first consider a single population. We rewrite (2.5) as a *stationary* autoregression:

$$y_{i+1,t+1}-y_{i+1,t} = \alpha_{i+1}(y_{i,t}-y_{i+1,t})+\beta_{i+1}(y_{i-1,t}-y_{i+1,t})+m_{i+1}+\epsilon_{i+1,t}, \qquad \forall i \geq 2, \quad (4.1)$$

and

$$y_{1,t+1} - y_{1,t} = m_1 + \epsilon_{1,t} \quad (4.2)$$

$$y_{2,t+1} - y_{2,t} = \alpha_2(y_{1,t} - y_{2,t}) + m_2 + \epsilon_{2,t}. \quad (4.3)$$

Both sides of these equations are stationary, since process $(y_{i+1,t} - y_{i,t})$ is stationary for each i. For each i, the pair of parameters (m_{i+1}, ρ_{i+1}) can be estimated *separately*, since it only appears in one equation. Such a strategy is called Seemingly Unrelated Regression (SUR) [see e.g. Zellner (1962)], in the sense that each OLS is conducted without taking into account the contemporaneous dependence of the residuals $(\epsilon_{i,t})$. Under mild conditions, this benchmark OLS is asymptotically consistent when the sample size T goes to infinity.

4.2 PLS estimation of the Granger causality parameters

The previous estimates of age-dependent parameters, α_i, β_i and m_i are obtained without any smoothness constraint. Such raw estimates are not satisfactory. Indeed, given the relative small sample size of most mortality data (generically, T is of the order of $30-100$), there is a significant risk that the OLS over-fits the data, and leads to erratic curves for these age-dependent parameters.

In this subsection we propose a (spatially) penalized least square estimation method, also called Tikhonov (or L^2) regularization [see e.g. Tsybakov and Zaiats (2009)], to ensure the smoothness of the curves α_i, β_i and m_i. The idea is to add a penalty term to

the sum of residuals, i.e., the error function of the OLS estimation, in order to penalize abrupt changes of these parameters across different ages. The penalty term is quadratic in the parameters of the model, which leads to an explicit solution for the optimization problem. More precisely, we consider the following penalized least square estimator:

$$\min L_1\Big((\alpha_i)_{i=2,\dots,I}, (\beta_i)_{i=3,\dots,I}, (m_i)_{i=1,\dots,I}\Big) \tag{4.4}$$

where

$$L_1 = \sum_{i=2}^{I-1}\sum_{t=1}^{T-1} \epsilon_{i+1,t+1}(\alpha_{i+1}, \beta_{i+1}, m_{i+1})^2 + \sum_{t=1}^{T-1} \epsilon_{2,t+1}(\alpha_2, \beta_2, m_2)^2 + \sum_{t=1}^{T-1} \epsilon_{1,t+1}(m_1)^2 \tag{4.5}$$

$$+ \lambda_\alpha \sum_{i=3}^{I-1}(\alpha_{i+1} - \alpha_i)^2 + \lambda_\beta \sum_{i=3}^{I-1}(\beta_{i+1} - \beta_i)^2 + \lambda_m \sum_{i=3}^{I-1}(m_{i+1} - m_i)^2, \tag{4.6}$$

with λ_ρ, λ_β, and λ_m being *known* and nonnegative parameters. The larger these parameters, the smoother the curves of α_i, β_i and m_i. In the special case where $\lambda_\rho = \lambda_\beta = \lambda_m = 0$, the unique solution of this minimization problem is the OLS estimator.

The right hand side of (4.5) is the sum of squared errors for each age and year, which are obtained using equations (2.3) to (2.5). Moreover, (4.6) corresponds to the penalty terms. Due to the boundary effect, the penalty does not concern the two lowest ages $i = 1$ and $i = 2$.

This penalization scheme is similar to Delwarde et al. (2007), who add a quadratic penalty term to the log-likelihood function, to estimate the (Poisson) Lee-Carter model. Then they propose a numerical, Newton-Raphson based algorithm to estimate the values of the parameters. In our least-square based optimization problem, the objective error function L is quadratic in each parameter α_i, β_i and m_i, which leads to a closed form solution. This result is summarized in the following proposition.

Proposition 4. *Estimates of α_i, β_i and m_i can be obtained in closed form formula by solving a linear system.*

Proof. See Appendix. □

4.3 PLS estimation of the variance-covariance matrix

Once we obtain the estimate of the causality matrix, we can recover the empirical values of the residuals, which we use to estimate the variance-covariance matrix. Similar as in the previous case, we use the following PLS estimator:

$$\min L_2\Big((a_i)_{i=1,\dots,I-1}, (c_i)_{i=2,\dots,I}\Big) \tag{4.7}$$

where

$$L_2 = \sum_{i=2}^{I-1}\sum_{t=1}^{T}\eta_{i,t}(a_i, c_i)^2 + \sum_{t=1}^{T}\eta_{1,t}(a_1)^2 + \sum_{t=1}^{T}\eta_{I,t}(c_I)^2 \tag{4.8}$$

$$+ \lambda_a\sum_{i=2}^{I-1}(a_{i+1} - a_i)^2 + \lambda_c\sum_{i=2}^{I-1}(c_{i+1} - c_i)^2. \tag{4.9}$$

The right hand side of (4.8) is the sum of squared errors from the dynamics of the residuals, which are obtained from (2.11). Again, λ_a and λ_c are known and nonnegative parameters. When these two parameters are equal to zero, the model can be estimated by seemingly unrelated ordinary least square, on an equation by equation basis. In the general case, we can show that [see Appendix 1] the estimates of a_i, c_i allow for closed form expressions.

4.4 The two-population model

Similarly, we rewrite equation (3.1) into:

$$\begin{aligned} y_{j,i+1,t+1} - y_{j,i+1,t} =& -\alpha_{j,i+1}\rho_{j,i+1}(y_{j,i+1,t} - y_{j,i,t}) + (1-\rho_{j,i+1})(y_{-j,i+1,t} - y_{j,i+1,t}) \\ &+ m_{j,i+1} + \epsilon_{j,i+1,t}, \end{aligned} \tag{4.10}$$

and

$$y_{j,1,t+1} - y_{j,1,t} = (1-\rho_{j,1})(y_{-j,1,t} - y_{j,1,t}) + m_{j,1} + \epsilon_{j,1,t}$$

for $j = 1, 2$. From Equation (4.10), we see that if we know $\rho_{j,i}$ and $\alpha_{j,i+1}\rho_{j,i+1}$ for $j = 1, 2$, we can in turn compute $\alpha_{j,i+1}$. Furthermore, if we treat $\alpha_{j,i+1}$ as a separate estimator

in the penalized estimation, the parameters are not quadratic in the objective function (with a form similar to (4.5) to (4.6)), and a closed form representation of the estimator is not available. Therefore, to preserve the quadratic property of the estimators, we denote by $\tilde{\alpha}_{j,i} = \alpha_{j,i+1}\rho_{j,i+1}$. Moreover, the estimation and the smoothing is done with respect to $\tilde{\alpha}_{j,i}$ instead of $\alpha_{j,i+1}$. In this case, the objective function can be written as

$$\min L((\tilde{\alpha}_{j,i})_{j=1,2;i=2,\ldots,I}, (m_{j,i})_{j=1,2;i=1,\ldots,I}, (\rho_{j,i})_{j=1,2;i=1,\ldots,I}), \tag{4.11}$$

where

$$\begin{aligned} L = &\sum_{j=1}^{2}\sum_{i=1}^{I-1}\sum_{t=1}^{T-1}\Big(y_{j,i+1,t+1} - y_{j,i+1,t} - \tilde{\alpha}_{j,i+1}(y_{j,i,t} - y_{j,i+1,t}) - (1-\rho_{j,i+1})(y_{-j,i+1,t} - y_{j,i+1,t}) - m_{j,i+1}\Big)^2 \\ &+ \sum_{j=1}^{2}\sum_{t=1}^{T-1}\Big(y_{j,1,t+1} - y_{j,1,t} - (1-\rho_{j,1})(y_{-j,1,t} - y_{j,1,t}) - m_{j,1}\Big)^2 \\ &+ \sum_{j=1}^{2}\lambda_{j,\alpha}\sum_{i=2}^{I-1}(\tilde{\alpha}_{j,i+1} - \tilde{\alpha}_{j,i})^2 + \sum_{j=1}^{2}\lambda_{j,m}\sum_{i=2}^{I-1}(m_{j,i+1} - m_{j,i})^2 + \sum_{j=1}^{2}\lambda_{j,\rho}\sum_{i=2}^{I-1}(\rho_{j,i+1} - \rho_{j,i})^2, \end{aligned} \tag{4.12}$$

where $-j$ means 1 when $j = 2$ and 2 when $j = 1$. Again, the estimate is obtained by solving a linear system, which is detailed in the Appendix.

The parameters from the variance-covariance matrix are estimated in a way similar to the single-population case as well. In particular, let $\tilde{a}_{j,i,t} = a_{j,i,t}\theta_{j,i,t}$ and $\tilde{c}_{j,i,t} = c_{j,i,t}\theta_{j,i,t}$ for all j, i, and t, the objective function is given by

$$\min L_2\Big((\tilde{a}_{j,i})_{j=1,2;i=2,\ldots,I}, (\tilde{c}_{j,i})_{j=1,2;i=2,\ldots,I}, (\theta_{j,i})_{j=1,2;i=2,\ldots,I}\Big), \tag{4.13}$$

where

$$\begin{aligned} L_2 = &\sum_{j=1}^{2}\sum_{i=2}^{I-1}\sum_{t=1}^{T}\Big(\epsilon_{j,i,t} - \tilde{a}_{j,i}\epsilon_{j,i+1,t} - \tilde{c}_i\epsilon_{j,i-1,t} - (1-\theta_{j,i})\epsilon_{-j,i,t})\Big)^2 \\ &+ \sum_{j=1}^{2}\sum_{t=1}^{T}\Big(\epsilon_{j,1,t} - \tilde{a}_{j,1}\epsilon_{j,2,t} - (1-\theta_{j,2}\epsilon_{-j,2,t}\Big)^2 + \sum_{j=1}^{2}\sum_{t=1}^{T}\Big(\epsilon_{j,I,t} - \tilde{c}_{j,I}\epsilon_{j,I-1,t} - (1-\theta_{j,I,t})\epsilon_{-j,I,t}\Big)^2 \\ &+ \sum_{j=1}^{2}\lambda_{j,a}\sum_{i=2}^{I-1}(\tilde{a}_{j,i+1} - \tilde{a}_{j,i})^2 + \sum_{j=1}^{2}\lambda_{j,c}\sum_{i=2}^{I-1}(\tilde{c}_{j,i+1} - \tilde{c}_{j,i})^2 + \sum_{j=1}^{2}\lambda_{j,\theta}\sum_{i=2}^{I-1}(\tilde{\theta}_{j,i+1} - \tilde{\theta}_{j,i})^2. \end{aligned}$$

4.5 The optimal choice of the smoothing parameter: a cross-validation approach

Clearly, the PLS estimate depends on the choice of the nonnegative smoothing parameters:

$$\boldsymbol{\lambda}_1 = (\lambda_\alpha, \lambda_\beta, \lambda_m), \qquad \boldsymbol{\lambda}_2 = (\lambda_a, \lambda_c).$$

When they are all zero, that is when no smoothing constraint is imposed, the age-dependant parameters are under-smoothed. On the other hand, when they tend to infinity, these parameters become over-smoothed. In the following, we propose a cross-validation method to choose the optimal smoothing parameters.

The basic idea is to estimate the model, when observations related to one birth cohort are omitted. The resulting estimates are then used to forecast the mortality rates for this omitted birth cohort. The forecasting practices is done for all birth cohorts in the dataset, and the optimal values of $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are the ones which generate the minimal overall forecasting error for all cohorts. Let us first consider the optimal choice of $\boldsymbol{\lambda}_1$.

The choice of $\boldsymbol{\lambda}_1$ proceeds as follows. Denote by $\mathcal{G}$ the set of all birth cohorts included in the dataset. For each $g \in \mathcal{G}$, we estimate equation (4.4) by leaving out all residuals $\epsilon_{i,t}$ corresponding to the same cohort g, that is, whose indices satisfy $t-i=g$. This method is called leave-one-out, and the literature has also proposed other similar cross-validation methods such as the $h-$block method [see e.g. Burman et al. (1994); Racine (1997)]. It is also possible to leave out the observation of one single year. However, given the emphasize of cohort effect in our model, the cohort-based cross validation method seems to be more intuitive[12].

We denote the resulting estimate by $\theta(-g, \boldsymbol{\lambda}_1)$, where $-g$ indicates the omission of birth cohort g. Then we compute the out-of-sample forecast of the $y_{i,t}$-s related to the birth cohort g, which are collected in a column vector, $\hat{\boldsymbol{y}}(-g, \boldsymbol{\lambda})$, and the corresponding forecast error:

$$\hat{\boldsymbol{\epsilon}}(-g, \boldsymbol{\lambda}) := \boldsymbol{y}(g) - \hat{\boldsymbol{y}}(-g, \boldsymbol{\lambda}), \tag{4.14}$$

where $\boldsymbol{y}(g)$ is a column vector containing all observations $y_{i,t}$ with $(i,t) \in g$. The dimension of $\boldsymbol{y}(g)$ and $\hat{\boldsymbol{y}}(-g, \boldsymbol{\lambda})$ depends on g. The optimal $\boldsymbol{\lambda}_1$ is obtained by minimizing

[12] We thank an anonymous referee for this insightful suggestion.

the following measure of the forecast error:

$$\boldsymbol{\lambda}_{1,opt} = \arg\min e_1(\boldsymbol{\lambda}_1) = \arg\min \sum_{g \in \mathcal{G}} \hat{\boldsymbol{\epsilon}}(-g, \boldsymbol{\lambda})' \hat{\boldsymbol{\epsilon}}(-g, \boldsymbol{\lambda}). \tag{4.15}$$

Once the value of $\boldsymbol{\lambda}_1$ is fixed, we fit the model (2.3) - (2.5) to the whole dataset, and obtain the empirical values of the residuals $\epsilon_{i,t}$-s. For each $g \in \mathcal{G}$, we then minimize (4.7) while leaving out all $\eta_{i,t}$-s with $(i,t) \in g$. Then we compute the out-of-sample forecast of $\epsilon_{i,t}$ related to g, which we collect in a column vector $\hat{\boldsymbol{\epsilon}}(-g, \boldsymbol{\lambda}_2)$,[13] and calculate the corresponding forecast error:

$$\hat{\boldsymbol{\eta}}(-g, \lambda_2) := \boldsymbol{\epsilon}(g) - \hat{\boldsymbol{\epsilon}}(-g, \boldsymbol{\lambda}_2).$$

Similarly, $\epsilon(g)$ is a column vector containing the empirical values of $\epsilon_{i,t}$-s related to birth cohort g. The optimal value of $\boldsymbol{\lambda}_2$ is the one which minimizes the following measure of forecast error:

$$\boldsymbol{\lambda}_{2,opt} = \arg\min e_2(\boldsymbol{\lambda}_2) = \arg\min \sum_{g \in \mathcal{G}} \hat{\boldsymbol{\eta}}(-t, \lambda_2)' \hat{\boldsymbol{\eta}}(-t, \lambda_2). \tag{4.16}$$

For the two-population model, the optimal smoothing parameters are obtained in an analogous manner.

5 Empirical application

The mortality data used in this study is downloaded from the Human Mortality Database.[14] We focus on the uni-sex populations of the US and of England & Wales (henceforth UK). We use the logarithm of the single-age crude death rates for ages 0 to 99, and from 1950 to 2013 for both populations. We first fit the single population model described in Section 2.2 and 2.3 to the two populations separately, and then fit the two-population model in Section 3.

[13] The estimate can be computed in closed form, in the same way as the estimate from the whole sample.

[14] See: http://www.mortality.org/

5.1 The single population case

Let us first report the results of the single-population model. Figure 11 and 12 display the PLS-smoothed and the non-smoothed (with all $\lambda = 0$) estimates for the two populations. For the PLS estimate, the optimal values of $\boldsymbol{\lambda}$ for the two populations are reported in Table 1. From the figures, we see that the parameters estimated from PLS are much smoother than the non-smoothed estimates for both populations. Moreover, for the smoothed estimates, the conditions $\alpha_i + \beta_i < 1$ and $a_i + c_i < 1$ hold for all ages and both populations.

Country	λ_α	λ_β	λ_m	λ_a	λ_c
US	0.18	0.79	0.01	0.08	0.20
UK	0.42	0.79	1.10	0.43	1.21

Table 1: The optimal values of $\boldsymbol{\lambda}$ for the US and the UK population in the single population case.

5.2 The two-population case

Now we apply the two-population model discussed in Section 3 to the US and the UK population at the same time. Figure 13 and 14 display the smoothed and the original parameter estimates for the product and ratio measures, respectively, in the two-population case. For the PLS estimate, the optimal values of $\boldsymbol{\lambda}$ for the two populations are shown in Table 2.

Country	λ_α	λ_ρ	λ_m	λ_a	λ_c	λ_θ
US	0.06	0.07	2.39	0.19	0.18	0.07
UK	0.47	0.34	5.21	0.50	0.86	0.32

Table 2: The optimal values of $\boldsymbol{\lambda}$ for the average and difference mortality rates in the two-population case.

5.3 Forecasting performance

Let us now evaluate the out-of-sample forecasting performance of our model compared to the benchmarks, that are the Lee-Carter model [Lee and Carter (1992)] in the single population case, and the Li-Lee model [Li and Lee (2005)] in the two-population case. We first fit the single- and two-population STAR model and the benchmark models to

data up to year 2000 for each population, and forecast the future mortality rates from year 2001 to 2013. The forecasting results are compared in terms of their respective root mean squared forecast error (RMSFE). More precisely, if we denote by $\widehat{y}_{j,i,t}$ the corresponding forecast, where $j = 1$ for US and $j = 2$ for UK, then the RMSFE for population i is defined as:

$$RMSFE(j) = \sqrt{\frac{1}{I(U-\hat{u})} \sum_{u=\hat{u}}^{U} \sum_{i=1}^{I} (y_{j,i,u} - \widehat{y}_{j,i,u})^2}, \tag{5.1}$$

where $\hat{u}$ and U are year 2001 and 2013, respectively. The RMSFE ratio measures the average of the forecasting error for each year u and age x for each population.

Figure 6 and 7 show the RMSFE for both populations in the single-population and the two-population case, respectively. These RMSFE's are functions of the forecast horizon. We see that the STAR model provides more accurate forecasts than the respective benchmarks for all forecast horizons. Therefore, the STAR model appears to better capture the recent mortality trends.

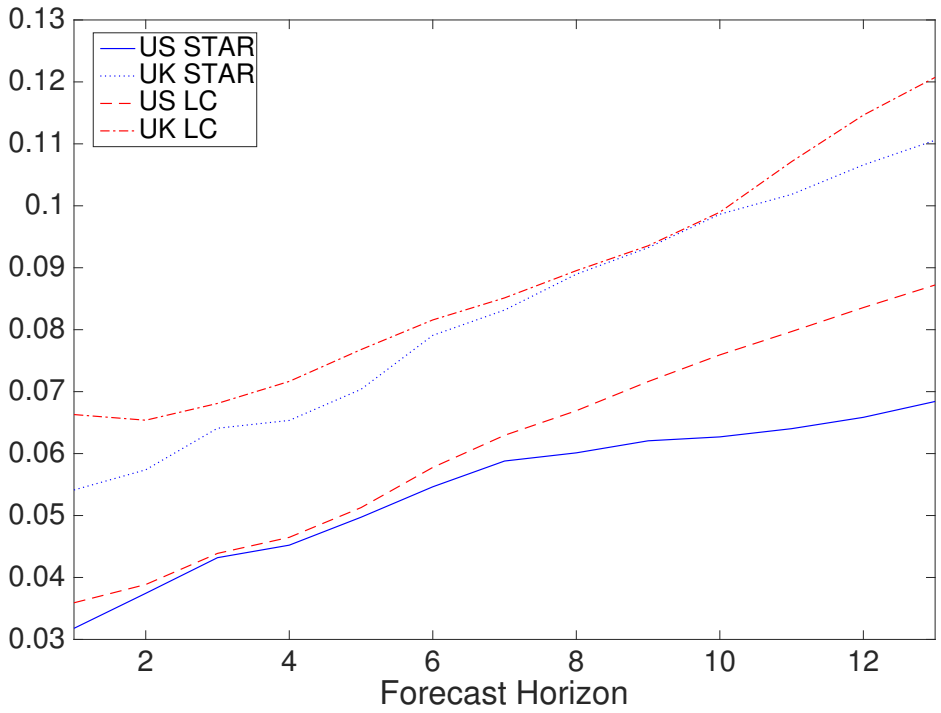


Figure 6: The RMSFE of the Lee-Carter and STAR model applied to two populations separately.

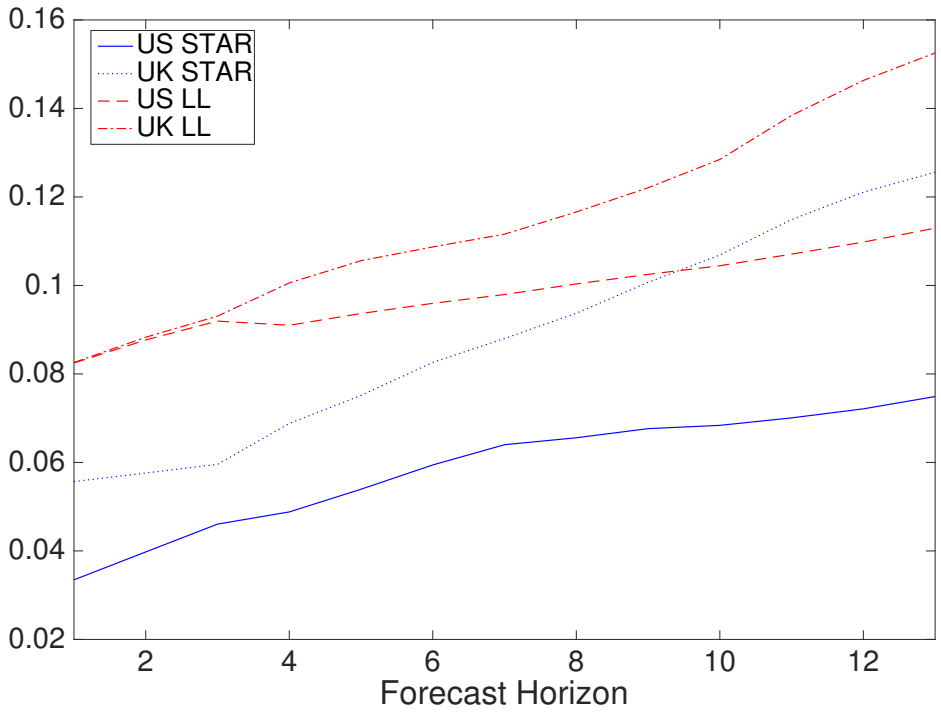
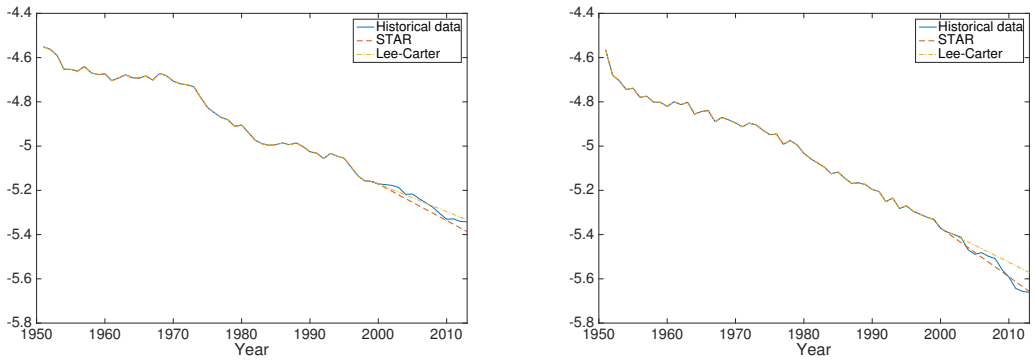


Figure 7: The RMSFE of the Li-Lee and STAR model applied simultaneously to two populations.

In order to understand this superiority, let us compare the forecast of the mortality rates we obtained from the two models. Figure 8 displays the out-of-sample forecasts of the the age-average of $y_{i,t}$-s from the STAR model and the Lee-Carter model for both populations.



(a) US (b) UK

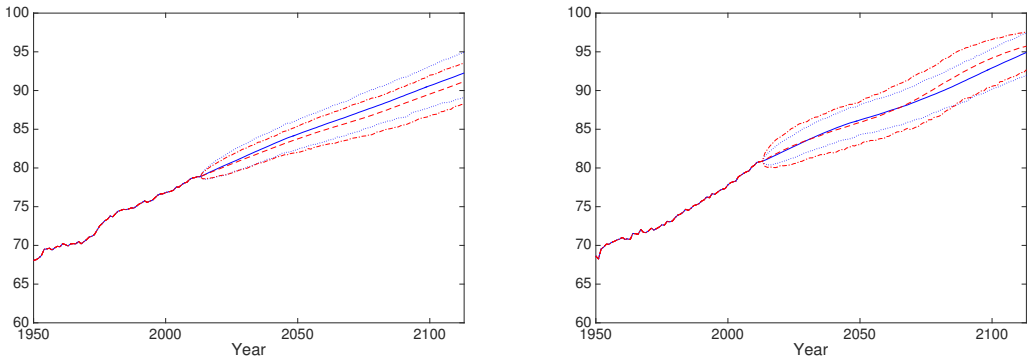
Figure 8: The out-of-sample forecasts of the average of $y_{i,t}$ from the STAR model and the Lee-Carter model.

From Figure 8, we see that the Lee-Carter model consistently under-forecasts the future mortality improvement. Indeed, in our model, future improvement is a weighted average of past improvements rates and the weights are more important for recent observations. On the other hand, the Lee-Carter model assumes an uncorrelated random walk for the aggregate mortality trend.[15] In other words, past improvements have equal weight regardless of the duration elapsed. Therefore, our model has a better ability to capture the most recent trend than the Lee-Carter and Li-Lee model, when the mortality improvement has accelerated during the observation period. This is clearly the case for the two populations in question.

Let us now examine the long-term forecast of the STAR models. In Figure 9 we project the period life expectancies at birth for the next 100 years using both the single-population and the two-population STAR models. We can see that the historical increase of life expectancy is faster for the UK population. Hence, the projected increase of life expectancy from the single-population model is also faster for the UK population. Due to the coherence property, the projected increase of life expectancy from the two-population model should be similar for the two populations in the long-run. As a result,

[15]The possibility of introducing correlated random walk has been recently examined by Leng and Peng (2016). They show that introducing such a correlation leads to potentially biased estimate of the Lee-Carter model. This is due to the non-stationary factor model structure of the Lee-Carter model. On the other hand, our model, which is a Vector Autoregression, does not suffer from this inconsistency. In particular, under mild conditions, the estimates are all asymptotically consistent, when T goes to infinity.

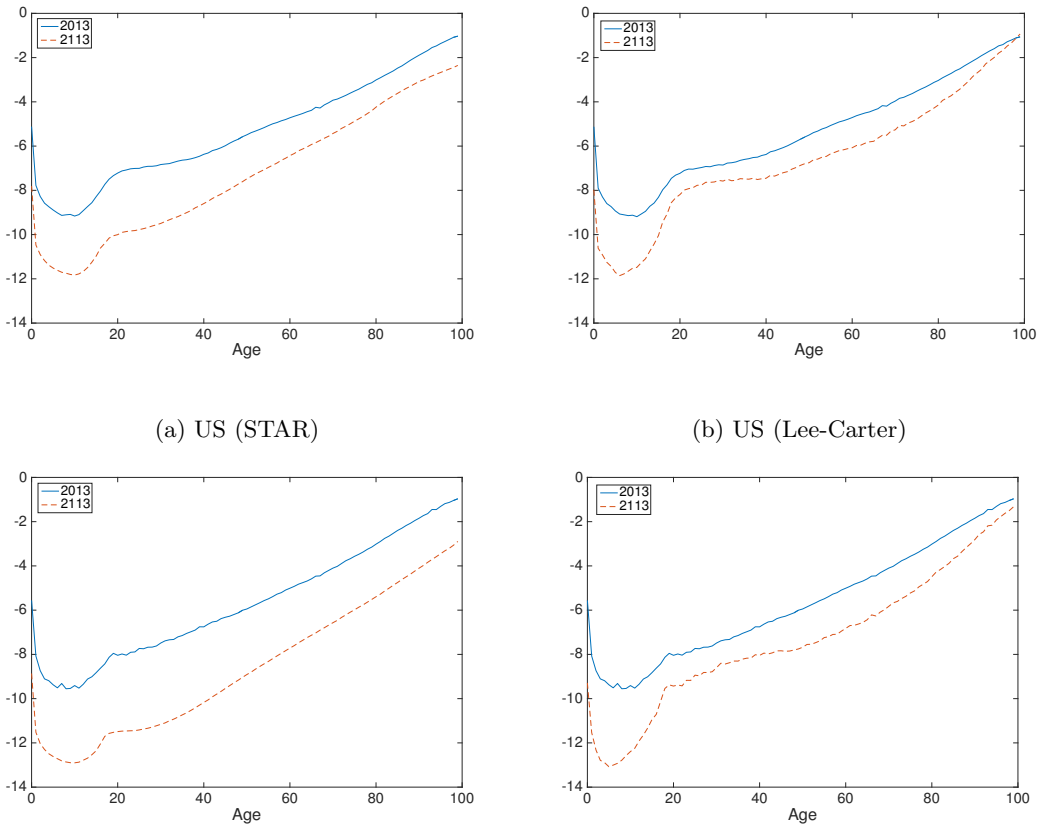
compared to the independent projections, the two-population model leads to, in the long-run, higher projected life expectancies for the US population, and lower projected life expectancies for the UK population. Nevertheless, as we can see, the forecasts of the single-population and two-population STAR model are quite similar in the short term, and the discrepancy only appears after 2050. This demonstrates the ability of the two-population STAR model to reconcile the long-term convergence with the short-term individual dynamics of the two populations.



(a) US (b) UK

Figure 9: The projected life expectancies at birth of the US and the UK population. The solid lines are the historical life expectancies and the mean forecasts from the two-population STAR model; the dashed lines are the mean forecasts of the single-population STAR model; the dotted lines are the 2.5% and the 97.5% quantiles of the two-population STAR model; dotted-dashed lines are the 2.5% and the 97.5% quantiles of the single-population STAR model.

As noted in Section 2.2.1, the STAR model ensures that projected mortality rates at different ages are co-integrated. This property is illustrated in Figure 10, where we project the logarithm of crude death rates using both the single-population STAR model and the Lee-Carter model 100 years into the future. We see that the projected decrease of mortality rates generated by the Lee-Carter model is much slower at older ages. As a consequence, in the long-run, the Lee-Carter model suggests nearly no improvement of mortality at, say, age 100. On the contrary, the projected mortality rates from the STAR model have the same asymptotic improvement rate. Hence, the mortality improvement rates at different ages are similar.



(a) US (STAR)

(b) US (Lee-Carter)

(c) UK (STAR)

(d) UK (Lee-Carter)

Figure 10: The projected logarithm of crude death rates from the Lee-Carter model and the STAR model.

6 Conclusion

This paper has proposed a spatial-temporal autoregressive (STAR) approach for mortality modelling. The model is based on the idea that neighbouring ages have higher degree of interdependence, both in terms of the Granger causality and the instantaneously causality. It provides more flexibility than the benchmark factor model, while at the same time ensures suitable long-run properties of the mortality processes. A closed form, penalized least square estimator has been proposed to ensure the smoothness of the age-dependent parameters both in the Granger causality matrix and the variance-covariance matrix. Finally, the STAR model can be easily extended to model multiple

populations simultaneously.

In the empirical study, we have fitted both the single-population and the two-population STAR model to the US and the UK population. We found that the STAR models are able to generate reasonable mortality forecasts in the long-run. Moreover, in the out-of-sample forecast, the STAR models outperforms the benchmark models (the Lee and Carter (1992) model in the single-population case, and the Li and Lee (2005) model in the two-population case).

A Appendix 1: Proofs

A.1 Proof of Proposition 2

First, from Equation (3.2) we get:

$$y_{1,1,t+1} - y_{2,1,t+1} = (\rho_{1,1} + \rho_{2,1} - 1)(y_{1,1,t} - y_{2,1,t}) + \alpha_{1,1} - \alpha_{2,1} + \epsilon_{1,1,t} - \epsilon_{2,1,t}.$$

Thus $(y_{1,1,t} - y_{2,1,t})$ is stationary.

Next, by equations (3.1) and (3.2), we have

$$\begin{aligned} y_{j,2,t+1} - y_{j,1,t+1} = \rho_{j,2}(1-\alpha_{j,2})(y_{j,2,t} - y_{j,1,t}) + (1-\rho_{j,2})(y_{-j,2,t} - y_{-j,1,t}) \\ + \underbrace{(\rho_{j,2} - \rho_{j,1})(y_{j,1,t} - y_{-j,1,t}) + m_{j,2} - m_{j,1} + \epsilon_{j,2,t} - \epsilon_{j,1,t}}_{\text{stationary}}, \end{aligned} \quad \text{(A.1)}$$

for $j = 1, 2$. From (A.1), we get a VAR model for the joint process:

$$(y_{1,2,t+1} - y_{1,1,t+1}, y_{2,2,t+1} - y_{2,1,t+1}),$$

with autoregressive matrix

$$A = \begin{bmatrix} \rho_{1,2}(1-\alpha_{1,2}) & 1-\rho_{1,2} \\ 1-\rho_{2,2} & \rho_{2,2}(1-\alpha_{1,2}) \end{bmatrix}.$$

Thus the stationary condition is that polynomial $\det(A - zId)$ has no roots outside the

unit circle $|z| \geq 1$. This is satisfied since:

$$\begin{aligned}|\Big[\rho_{1,2}(1-\alpha_{1,2})-z\Big]\Big[\rho_{2,2}(1-\alpha_{1,2})-z\Big]| &\geq |\Big[1-\rho_{1,2}(1-\alpha_{1,2})\Big]\Big[1-\rho_{2,2}(1-\alpha_{2,2})\Big]| \\ &\geq (1-\rho_{1,2})(1-\rho_{2,2})\end{aligned}$$

for all z such that $|z| \geq 1$.

Similarly, by induction we can show that process

$$(y_{1,i+1,t+1} - y_{1,i,t+1}, y_{2,i+1,t+1} - y_{2,i,t+1})$$

is stationary.

A.2 Proof of Proposition 3

The proof of Proposition 3 is similar as in the bivariate case. Let us first consider the vector of mortalities at initial ages $Y_{1,t} = (y_{1,1,t}, y_{2,1,t}, ..., y_{J,1,t})$. Equation (3.5) says that $Y_{1,t}$ follows a VAR(1) model, with autoregressive matrix

$$R_1 = (\rho_{j,k,1}). \tag{A.2}$$

Since all its entries are positive[16], and the sum of each row is equal to 1, this is a stochastic matrix. By Perron-Frobenius's theorem [see e.g. Seneta (2006)], it has a unique eigenvalue on the unit circle, with all other eigenvalues smaller than 1 in modulus. Thus vector $(y_{1,1,t}, y_{2,1,t}, ..., y_{J,1,t})$ is co-integrated and the rank of co-integration is $J-1$.

Let us write the Jordan-Chevalley-Dunford decomposition of the autoregressive matrix R_1:

$$R_1 = P^{-1}DP,$$

where P is a real invertible matrix, and D is triangular superior. Since by definition the sum of each row of $R_1 = (\rho_{j,k,1})$ is equal to one, namely $R_1 e = e$, where e is the unitary vector, we can deduce that 1 is the eigenvalue of R_1. Therefore, without loss of

[16]The case of analysis of square matrix with nonnegative entries is similar to the case of positive matrix, see e.g. Seneta (2006)

generality, we can assume that $d_{1,1}$ is equal to 1. Then we can rewrite (3.5) into:

$$PY_t = DPY_{t-1} + m_1 + \epsilon_t.$$

Thus the $2, 3, .., J-$th components of vector PY_t are all stationary. They constitute the linear span of co-integration vectors of Y_t, which is of dimension $J-1$.

Let us now use the condition $R_1 e = e$, that is $D(Pe) = Pe$. In other words Pe is a non zero (since P is invertible) eigenvector associated with the eigenvalue 1 of D. Since D is triangular superior, and 1 is its simple eigenvalue, we deduce that the $2, 3, .., J-$th components of $P'e$ are all zero. Therefore, we have found $J-1$ co-integration vector, which are the $2, 3, .., J-$th components of matrix P. They are linearly independent (since P is invertible), and share the common property that all the entries sum up to zero. Since the vector space of vectors whose components sum up to zero is of dimension $J-1$, we conclude that any vector with exactly one entry with value 1, one entry with value -1, and $J-2$ entries with value 0, belongs to this vector space. Thus each of the two mortality rates $y_{j,1,t}$ and $y_{j',1,t}$ are co-integrated.

Let us now consider the mortality rates at *all* ages. First, we remark that the vector $Z_t = (y_{j,i,t})_{1 \leq j \leq J, 1 \leq i \leq I}$ forms a large VAR system with autoregressive matrix R_2 that is, by definition, row stochastic $R_2 e_2 = e_2$, where e_2 is the $I \times J$ dimensional unitary vector. Let us show that all eigenvalues of this large matrix is smaller than 1 in modulus.

To this end, let us write R_2 in blocks:

$$R_2 = \begin{bmatrix} A_1 & 0 & \cdots & \cdots \\ * & A_2 & \cdots & \cdots \\ * & * & A_3 & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}. \tag{A.3}$$

This is a matrix of blocks with I rows and I columns, corresponding to the I different ages, and each entry, such as A_1, is itself a matrix of dimension $J \times J$. Moreover, as a block matrix, R_2 is triangular superior, in the sense that in equation (A.3), all the $J \times J$ dimensional block matrices above the diagonal are zero. This is due to the fact that in our multi-population mortality specification, the mortality at age i depends only on past

mortality at age i or lower.

We can also remark that the block matrix A_1 is nothing else but the matrix R_1 we considered in equation (A.2). In particular, A_1 has 1 as a simple eigenvalue, and all other eigenvalues are smaller than 1 in modulus.

In order to prove that all other eigenvalues of R_2 are smaller than 1 in modulus, we remark that since R_2 is (block) triangular superior, the characteristic polynomial $\det(R_2 - xId)$ can be factorized into:

$$\det(R_2 - xId) = \det(A_1 - xId)\det(A_2 - xId)\cdots\det(A_I - xId). \tag{A.4}$$

Thus the spectrum of R_2 is the union of the spectrum of the I different $J \times J$ matrices $A_1, ..., A_I$. It can be easily shown that $A_2, ..., A_I$ all have nonnegative entries only, and the sum of each row is strictly smaller than 1[17]. Hence they do not have eigenvalues with modulus smaller than 1. Therefore all the roots of $\det(A_i - xId)$ are larger than unity in modulus, for $i = 2, ..., I$. As a consequence, we conclude that 1 is the single, simple eigenvalue with modulus non smaller than 1 of R_2. Thus the vector $(y_{j,i,t})_{1\leq j\leq J, 1\leq i\leq I}$ has $(I \times J - 1)$ linearly independent co-integration relationship. Then by the same argument as for the vector of mortality at initial age $(y_{j,1,t})_{1\leq j\leq J}$, we can deduce that there exists $(I \times J - 1)$ co-integration vector whose sums of components are zero, and then, all $I \times J$ dimensional vector of the type $(0, ..., 1, 0, ..., -1, 0, ...)$ are co-integration vectors.

[17] This is because the mortality rates at higher ages $i \geq 2$ depend not only on past mortality rates of other populations at the same age i, but also past mortality rates at an inferior age $i - 1$.

B Appendix 2: Closed form formula of the PLS estimate

PLS estimation of the causality matrix Let us solve the minimization problem (4.4). The minimiser is a solution of the following equations:

$$
\begin{aligned}
0 = \frac{\partial L_1}{\partial \alpha_{i+1}} &= -2\sum_{t=1}^{T-1}(y_{i,t}-y_{i+1,t})\Big(y_{i+1,t+1}-y_{i+1,t}-\alpha_{i+1}(y_{i,t}-y_{i+1,t}) \\
&\quad -\beta_{i+1}(y_{i-1,t}-y_{i+1,t})-m_{i+1}\Big)+2\lambda_\alpha\Big[2\alpha_{i+1}-\alpha_i-\alpha_{i+2}\Big], \\
0 = \frac{\partial L_1}{\partial \beta_{i+1}} &= -2\sum_{t=1}^{T-1}(y_{i-1,t}-y_{i+1,t})\Big(y_{i+1,t+1}-y_{i+1,t}-\alpha_{i+1}(y_{i,t}-y_{i+1,t}) \\
&\quad -\beta_{i+1}(y_{i-1,t}-y_{i+1,t})-m_{i+1}\Big)+2\lambda_\beta\Big[2\beta_{i+1}-\beta_i-\beta_{i+2}\Big], \\
0 = \frac{\partial L_1}{\partial m_{i+1}} &= -2\sum_{t=1}^{T-1}\Big(y_{i+1,t+1}-y_{i+1,t}-\alpha_{i+1}(y_{i,t}-y_{i+1,t})-\beta_{i+1}(y_{i-1,t}-y_{i+1,t})-m_{i+1}\Big) \\
&\quad +2\lambda_m\Big[2m_{i+1}-m_i-m_{i+2}\Big],
\end{aligned}
$$

for $i = 2, ..., I-1$, and for the boundary ages:

$$
\begin{aligned}
0 = \frac{\partial L_1}{\partial \alpha_2} &= -2\sum_{t=1}^{T-1}(y_{1,t}-y_{2,t})\Big(y_{2,t+1}-y_{2,t}-\alpha_2(y_{1,t}-y_{2,t})-m_2\Big), \\
0 = \frac{\partial L_1}{\partial m_2} &= -2\sum_{t=1}^{T-1}\Big(y_{2,t+1}-y_{2,t}-\alpha_2(y_{1,t}-y_{2,t})-m_2\Big), \\
0 = \frac{\partial L_1}{\partial m_1} &= -2\sum_{t=1}^{T-1}\Big(y_{I,t+1}-y_{I,t}-m_1\Big).
\end{aligned}
$$

Thus parameters m_1, m_2, and α_2 are estimated directly from the linear system composed of the three boundary conditions. This is equivalent to solve the OLS for the regression models (2.3) and (2.4).

The rest of the parameters satisfy the linear system in $(\alpha_i, \beta_i, m_i)_{i=2,...,I-1}$. This system has $3 \times (I-2)$ unknown parameters and as many equations. Thus the solution of this linear system is unique, whenever the determinant of the associated matrix is non null. But this determinant is polynomial in $\lambda_\alpha, \lambda_\beta$ and λ_m, and when $\lambda_m = \lambda_\alpha = \lambda_\beta = 0$, the solution of the system is unique (and is, by definition, the OLS estimator). Therefore, the determinant is non null when $\lambda_m = \lambda_\alpha = \lambda_\beta = 0$. As a consequence, the determinant

is non null for almost all $\lambda_\alpha, \lambda_\beta, \lambda_m$,[18] and the unique minimizer of the optimization problem can be computed in closed form, by solving the linear system.

PLS estimation of the variance-covariance matrix Let us now solve the minimization problem (4.7). The minimizer satisfies the following linear system:

$$\begin{aligned} 0 = \frac{\partial L_2}{\partial a_i} &= -2\sum_{t=1}^{T} \epsilon_{i+1,t}\Big(\epsilon_{i,t} - a_i\epsilon_{i+1,t} - c_i\epsilon_{i-1,t}\Big) + 2\lambda_a\Big[2a_i - a_{i-1} - a_{i+1}\Big], \\ 0 = \frac{\partial L_2}{\partial c_i} &= -2\sum_{t=1}^{T} \epsilon_{i-1,t}\Big(\epsilon_{i,t} - a_i\epsilon_{i+1,t} - c_i\epsilon_{i-1,t}\Big) + 2\lambda_c\Big[2c_i - c_{i-1} - c_{i+1}\Big], \end{aligned} \tag{B.1}$$

for all ages $i = 2, ..., I-1$, as well as:

$$\begin{aligned} 0 = \frac{\partial L_2}{\partial a_1} &= -2\sum_{t=1}^{T} \epsilon_{i+1,t}(\epsilon_{1,t} - a_1\epsilon_{2,t}) \\ 0 = \frac{\partial L_2}{\partial c_I} &= -2\sum_{t=1}^{T} \epsilon_{i+1,t}(\epsilon_{I,t} - c_I\epsilon_{I-1,t}) \end{aligned}$$

for the two boundary ages. As in the previous subsection, parameters a_1 and c_I are estimated using OLS, whereas the rest of the parameters $(a_i, c_i)_{i=2,...,I-1}$ are obtained by solving a linear system (B.1). By construction, the determinant of the latter is polynomial in λ_a and λ_c, and is non null when $\lambda_a = \lambda_c = 0$. Therefore, it is non null for almost surely all values λ_a and λ_c.

[18] In the sense of Lebesgue measure.

The two-population case Let us now solve the minimization problem (4.11). The minimizer satisfies the following linear system:

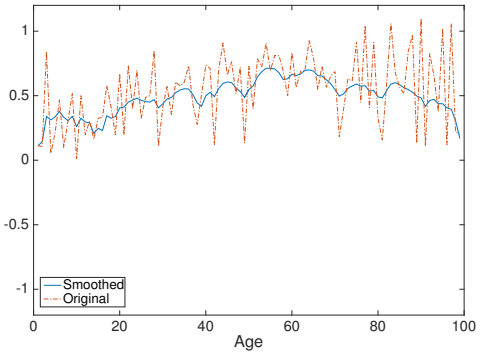
$$
\begin{aligned}
0 = \frac{\partial L}{\partial \tilde{\alpha}_{j,i+1}} &= -2 \sum_{t=1}^{T-1} (y_{j,i,t} - y_{j,i+1,t}) \Big(y_{j,i+1,t+1} - y_{j,i+1,t} - \tilde{\alpha}_{j,i+1}(y_{j,i,t} - y_{j,i+1,t}) \\
&\quad - (1 - \rho_{j,i+1})(y_{-j,i+1,t} - y_{j,i+1,t}) - m_{j,i+1} \Big) + 2\lambda_{j,\alpha} \Big[2\tilde{\alpha}_{j,i+1} - \tilde{\alpha}_{j,i} - \tilde{\alpha}_{j,i+2} \Big], \\
0 = \frac{\partial L}{\partial m_{i+1}} &= -2 \sum_{t=1}^{T-1} \Big(y_{j,i+1,t+1} - y_{j,i+1,t} - \tilde{\alpha}_{j,i+1}(y_{j,i,t} - y_{j,i+1,t}) \\
&\quad - (1 - \rho_{j,i+1})(y_{-j,i+1,t} - y_{j,i+1,t}) - m_{j,i+1} \Big) + 2\lambda_{j,m} \Big[2m_{j,i+1} - m_{j,i} - m_{j,i+2} \Big], \\
0 = \frac{\partial L}{\partial \rho_{j,i+1}} &= -2 \sum_{t=1}^{T-1} (y_{j,i+1,t} - y_{-j,i+1,t}) \Big(y_{j,i+1,t+1} - y_{j,i+1,t} - \tilde{\alpha}_{j,i+1}(y_{j,i,t} - y_{j,i+1,t}) \\
&\quad - (1 - \rho_{j,i+1})(y_{-j,i+1,t} - y_{j,i+1,t}) - m_{j,i+1} \Big) + 2\lambda_{j,\beta} \Big[2\rho_{j,i+1} - \rho_{j,i} - \rho_{j,i+2} \Big],
\end{aligned}
$$

for $j = 1, 2$ and $i = 1, ..., I - 1$. For age 1, the linear system is given by

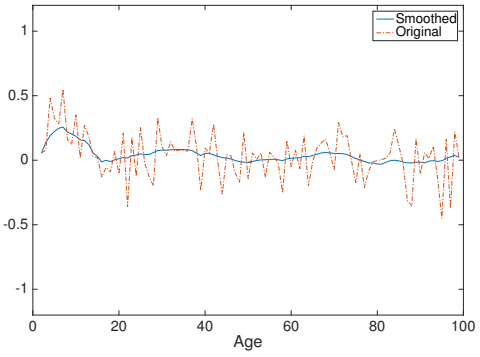
$$
\begin{aligned}
0 = \frac{\partial L}{\partial m_{j,1}} &= -2 \sum_{t=1}^{T-1} \Big(y_{j,1,t+1} - y_{j,1,t} - (1 - \rho_{j,1})(y_{-j,1,t} - y_{j,1,t}) - m_{j,1} \Big), \\
0 = \frac{\partial L}{\partial \rho_{j,1}} &= -2 \sum_{t=1}^{T-1} (y_{j,1,t} - y_{-j,1,t}) \Big(y_{j,1,t+1} - y_{j,1,t} - (1 - \rho_{j,1})(y_{-j,1,t} - y_{j,1,t}) - m_{j,1} \Big),
\end{aligned}
$$

for $j = 1, 2$.

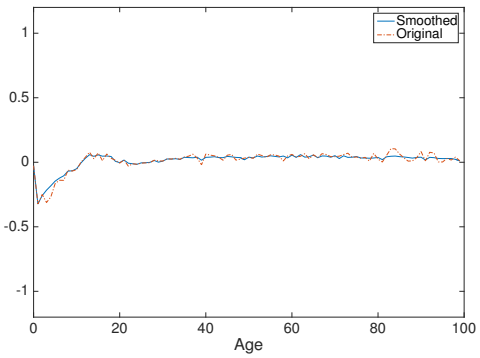
C Appendix 3: Additional figures



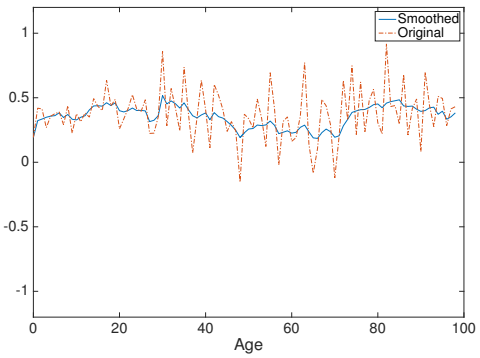
(a) α



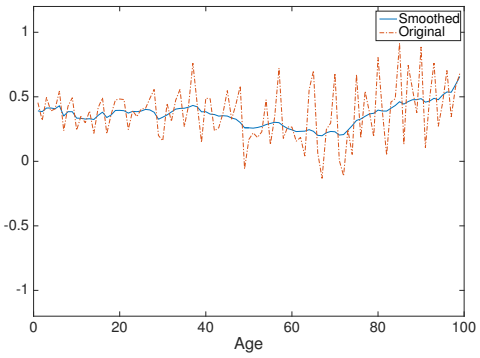
(b) β



(c) m

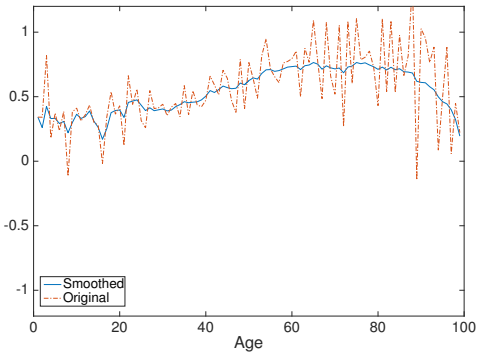


(d) a

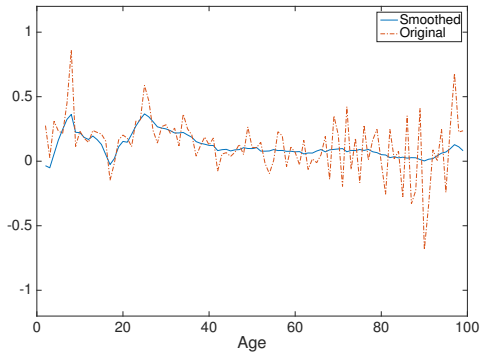


(e) c

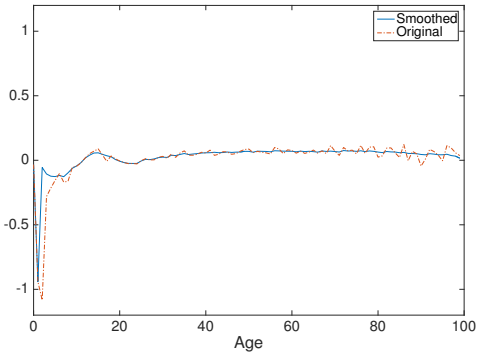
Figure 11: The smoothed and the original estimates for the US population from the single-population model.



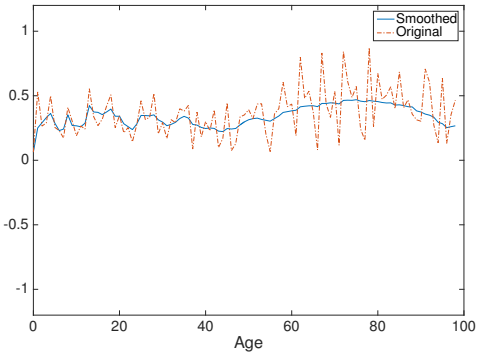
(a) α



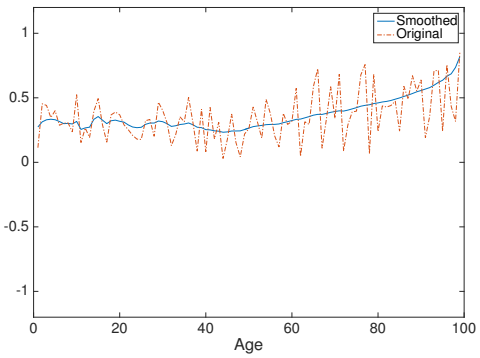
(b) β



(c) m

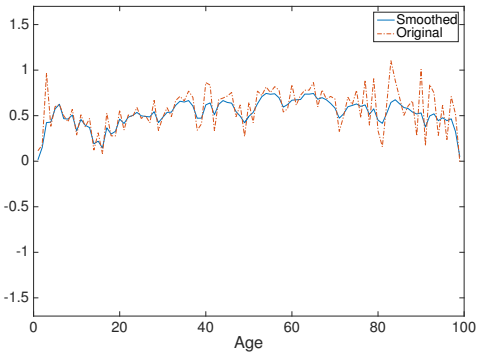


(d) a

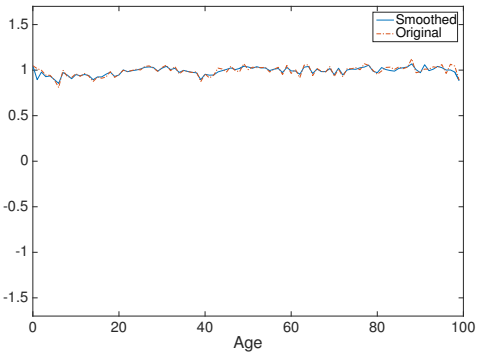


(e) c

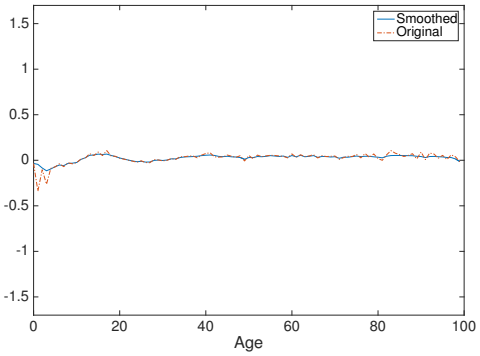
Figure 12: The smoothed and the original estimates for the UK population from the single-population model.



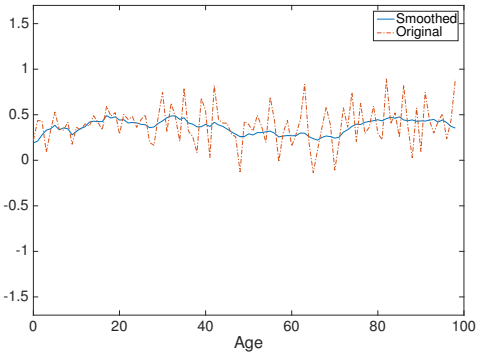
(a) α



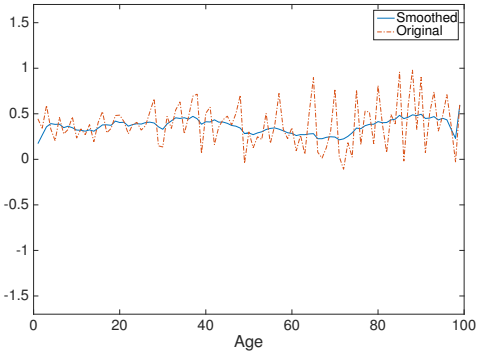
(b) ρ



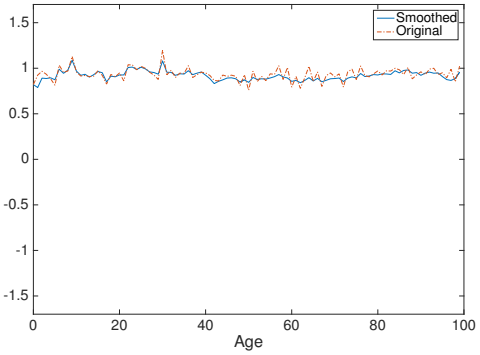
(c) m



(d) a

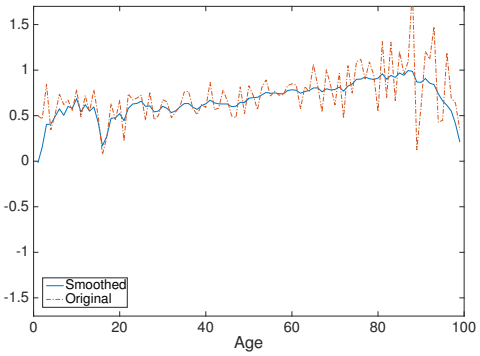


(e) c

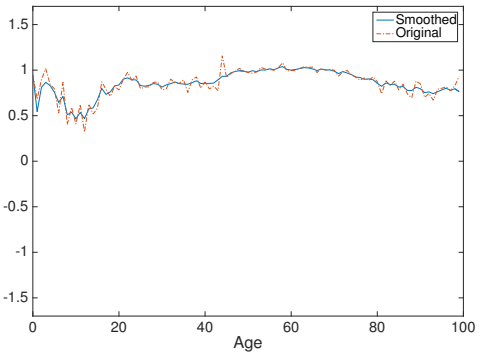


(f) θ

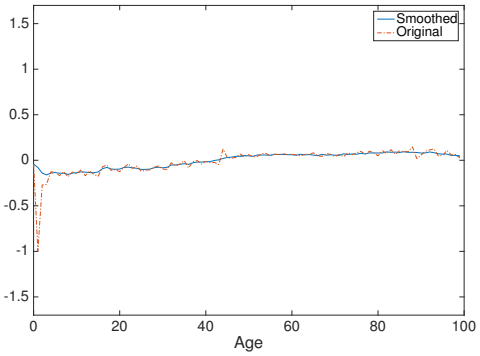
Figure 13: The smoothed and the original estimates for the US population from the two-population model.



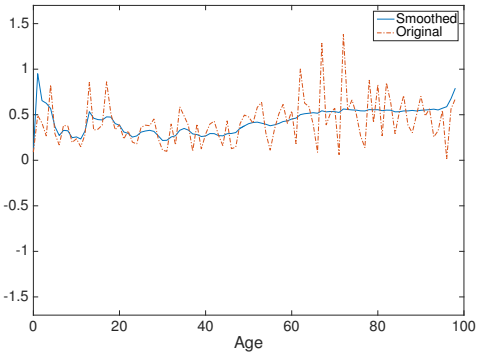
(a) α



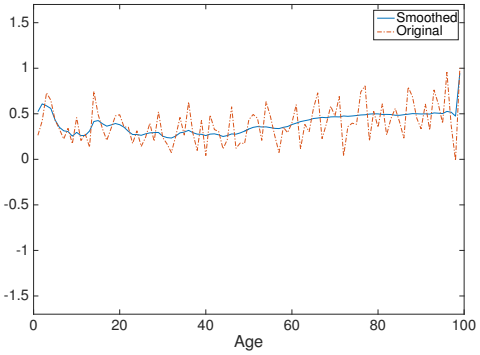
(b) ρ



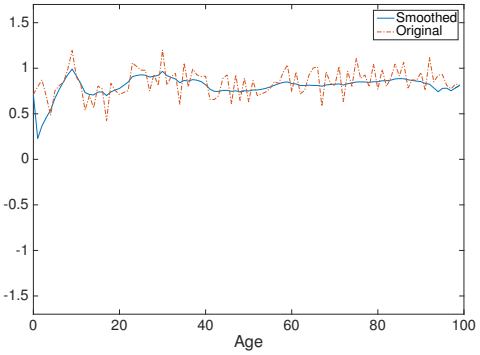
(c) m



(d) a



(e) c



(f) θ

Figure 14: The smoothed and the original estimates for the UK population from the two-population model.

D Appendix 4: Granger and instantaneous causality

The aim of this appendix is to review the notions of causality introduced in the (discrete time) literature. The first concept is the directional causality, or Granger's causality, which has been first introduced by Granger (1969).

Definition 1 (Granger's (non)causality)**.** Let $F = (F_t)$ and $Z = (Z_t)$ be two discrete time processes, then the process F does not Granger cause the process Z if and only if:

$$Z_t \perp\!\!\!\perp \underline{F_{t-1}} \mid \underline{Z_{t-1}}, \qquad \forall t > 0, \tag{D.1}$$

where the symbol $\cdot \perp\!\!\!\perp \cdot \mid \cdot$ means conditional independence.

Similarly, if the noncausality condition is not satisfied, then we say there is Granger causality from F towards Z.

Thus the process F does not Granger cause the process Z, if, at any date, the current value of Z_t is independent of the lagged values of F given the lagged values of Z. Equivalently the best predictor of any nonlinear transformation $g(Z_t)$ of Z_t given the whole information set $(\underline{F_{t-1}}, \underline{Z_{t-1}})$ does not depend on the past history of F.

In the context of a VAR model of (F_t, Z_t), say:

$$\begin{bmatrix} F_t \\ Z_t \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{bmatrix} \begin{bmatrix} F_{t-1} \\ Z_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{bmatrix}, \tag{D.2}$$

where $(\epsilon_{1t}, \epsilon_{2t})$ is an i.i.d. sequence of residuals, the Granger causality is equivalent to the following:

Definition 2 (Granger's causality in a VAR model)**.** Process (F_t) (resp. (Z_t)) Granger causes (Z_t) (resp. (F_t)) if and only if the coefficient r_{21} (resp. r_{12}) is non zero.

Thus this definition is directional. We can have situations where (F_t) Grange causes (Z_t), but not the converse. Moreover, this causality measures the cross-dependence between different periods. The next concept of causality is not directional, and concerns the instantaneous dependence between different time series.

Definition 3 (Instantaneous causality in a VAR model)**.** We say that there is instantaneous causality between processes (F_t) and (Z_t) if and only if the residuals $\epsilon_{1t}, \epsilon_{2t}$ are not jointly independent.

For instance, if $(\epsilon_{1t}, \epsilon_{2t})$ is jointly normal, then there is instantaneous causality if and only if the variance-covariance matrix of the residuals is not diagonal.

References

Barrieu, P., Bensusan, H., El Karoui, N., Hillairet, C., Loisel, S., Ravanelli, C., and Salhi, Y. (2012). Understanding, Modelling and Managing Longevity Risk: Key Issues and Main Challenges. *Scandinavian Actuarial Journal*, 2012(3):203–231.

Biffis, E. and Millossovich, P. (2006). A Bidimensional Approach to Mortality Risk. *Decisions in Economics and Finance*, 29(2):71–94.

Burman, P., Chow, E., and Nolan, D. (1994). A Cross-Validatory Method for Dependent Data. *Biometrika*, 81(2).

Cairns, A. J. G., Blake, D., and Dowd, K. (2006). A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration. *Journal of Risk and Insurance*, 73(4):687–718.

Cairns, A. J. G., Blake, D., and Dowd, K. (2008). Modelling and Management of Mortality Risk: a Review. *Scandinavian Actuarial Journal*, 2008(2-3):79–113.

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G., Epstein, D., Ong, A., and Balevich, I. (2009). A Quantitative Comparison of Stochastic Mortality Models Using Data from England and Wales and the United States. *North American Actuarial Journal*, 13(1):1–35.

Cairns, A. J. G., Blake, D., Dowd, K., Coughlan, G., and Khalaf-Allah, M. (2011). Bayesian Stochastic Mortality Modelling for Two Populations. *ASTIN Bulletin*, 41(1):29–59.

Callot, L., Haldrup, N., and Kallestrup-Lamb, M. (2016). Deterministic and Stochastic Trends in the Lee-Carter Mortality Model. *Applied Economics Letters*, 23(7):486–493.

Chuliá, H., Guillén, M., and Uribe, J. M. (2015). Modeling Longevity Risk With Generalized Dynamic Factor Models and Vine Copulae. *ASTIN Bulletin*, 46(01):1–26.

Debón, A., Montes, F., Mateu, J., Porcu, E., and Bevilacqua, M. (2008). Modelling Residuals Dependence in Dynamic Life Tables: A Geostatistical Approach. *Computational Statistics & Data Analysis*, 52(6):3128–3147.

Delwarde, A., Denuit, M., and Eilers, P. (2007). Smoothing the Lee–Carter and Poisson Log-bilinear Models for Mortality Forecasting: A Penalized Log-likelihood Approach. *Statistical Modelling*, 7(1):29–48.

Dowd, K., Cairns, A. J. G., Blake, D., Coughlan, G., and Khalaf-Allah, M. (2011). A Gravity Model of Mortality Rates for Two Related Populations. *North American Actuarial Journal*, 15(2):334–356.

Engle, R. F. and Granger, C. W. (1987). Co-integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, 55(2):251–276.

Gaille, S. and Sherris, M. (2011). Modelling Mortality with Common Stochastic Long-Run Trends. *The Geneva Papers on Risk and Insurance-Issues and Practice*, 36(4):595–621.

Granger, C. (1969). Investigating Causal Relations by Econometric Models and Cross-Spectral Methods. *Econometrica*, 37(3):424–38.

Heckman, J. and Robb, R. (1985). Using Longitudinal Data to Estimate Age, Period and Cohort Effects in Earnings Equations. In *Cohort Analysis in Social Research*, pages 137–150. Springer.

Hunt, A. and Blake, D. (2015a). Identifiability, Cointegration and the Gravity Model. *Pension Institute DP*.

Hunt, A. and Blake, D. (2015b). Modelling Longevity Bonds: Analysing the Swiss Re Kortis Bond. *Insurance: Mathematics and Economics*, 63:12–29.

Hunt, A. and Villegas, A. M. (2015). Robustness and Convergence in the Lee–Carter Model with Cohort Effects. *Insurance: Mathematics and Economics*, 64:186–202.

Hyndman, R. J., Booth, H., and Yasmeen, F. (2013). Coherent Mortality Forecasting: the Product-Ratio Method with Functional Time Series Models. *Demography*, 50(1):261–283.

Kuang, D., Nielsen, B., and Nielsen, J. (2008). Identification of the Age-period-Cohort Model and the Extended Chain-Ladder Model. *Biometrika*, 95(4):979–986.

Lazar, D. and Denuit, M. M. (2009). A Multivariate Time Series Approach to Projected Life Tables. *Applied Stochastic Models in Business and Industry*, 25(6):806–823.

Lee, R. and Carter, L. (1992). Modeling and Forecasting US Mortality. *Journal of the American Statistical Association*, 87(419):659–671.

Leng, X. and Peng, L. (2016). Inference Pitfalls in Lee-Carter Model for Forecasting Mortality. *Insurance: Mathematics and Economics*, 70:58–65.

Li, H., O'Hare, C., and Zhang, X. (2015). A Semiparametric Panel Approach to Mortality Modeling. *Insurance: Mathematics and Economics*, 61:264–270.

Li, J. S.-H. and Hardy, M. R. (2011). Measuring Basis Risk in Longevity Hedges. *North American Actuarial Journal*, 15(2):177–200.

Li, N. and Lee, R. (2005). Coherent Mortality Forecasts for a Group of Populations: An Extension of the Lee-Carter Method. *Demography*, 42(3):575–594.

Li, N., Lee, R., and Gerland, P. (2013). Extending the Lee-Carter Method to Model the Rotation of Age Patterns of Mortality Decline for Long-Term Projections. *Demography*, 50(6):2037–2051.

Litterman, R. B. (1986). Forecasting with Bayesian Vector Autoregressions-Five Years of Experience. *Journal of Business & Economic Statistics*, 4(1):25–38.

Mavros, G., Cairns, A. J. G., Kleinow, T., and Streftaris, G. (2016). Stochastic Mortality Modelling: Key Drivers and Dependent Residuals. *Heriot-Watt University DP*.

Pace, R. K., Barry, R., Clapp, J. M., and Rodriquez, M. (1998). Spatiotemporal Autoregressive Models of Neighborhood Effects. *Journal of Real Estate Finance and Economics*, 17(1):15–33.

Plat, R. (2009). Stochastic Portfolio Specific Mortality and the Quantification of Mortality Basis Risk. *Insurance: Mathematics and Economics*, 45(1):123–132.

Racine, J. (1997). Feasible Cross-Validatory Model Selection for General Stationary Processes. *Journal of Applied Econometrics*, 12(2):169–179.

Renshaw, A. and Haberman, S. (2006). A Cohort-Based Extension to the Lee-Carter Model for Mortality Reduction Factors. *Insurance: Mathematics and Economics*, 38(3):556–570.

Salhi, Y. and Loisel, S. (2016). Basis Risk Modelling: a Co-integration Based Approach. *forthcoming in Statistics*.

Seneta, E. (2006). *Non-Negative Matrices and Markov Chains*. Springer Science & Business Media.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tsybakov, A. B. and Zaiats, V. (2009). *Introduction to Nonparametric Estimation*, volume 11. Springer.

Yang, S. S. and Wang, C.-W. (2013). Pricing and Securitization of Multi-Country Longevity Risk with Mortality Dependence. *Insurance: Mathematics and Economics*, 52(2):157–169.

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American statistical Association*, 57(298):348–368.

Zhou, R., Wang, Y., Kaufhold, K., Li, J. S.-H., and Tan, K. S. (2014). Modeling Period Effects in Multi-Population Mortality Models: Applications to Solvency II. *North American Actuarial Journal*, 18(1):150–167.