ARTICLE IN PRESS

International Journal of Forecasting xxx (xxxx) xxx

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting mortality with a hyperbolic spatial temporal VAR model

Lingbing Feng [a], Yanlin Shi [b,*], Le Chang [c]

[a] Institute of Industrial Economics, Jiangxi University of Finance and Economics, Nanchang, Jiangxi, China
[b] Department of Actuarial Studies and Business Analytics, Macquarie University, Sydney, NSW, Australia
[c] Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, ACT, Australia

ARTICLE INFO

Keywords:
Mortality forecasting
Vector autoregressive
Co-integration
Penalized least squares
Lee–Carter model

ABSTRACT

Accurate forecasts of mortality rates are essential to various types of demographic research like population projection, and to the pricing of insurance products such as pensions and annuities. Recent studies have considered a spatial–temporal vector autoregressive (STVAR) model for the mortality surface, where mortality rates of each age depend on the historical values for that age (temporality) and the neighboring cohorts ages (spatiality). This model has sound statistical properties including co-integrated dependent variables, the existence of closed-form solutions and a simple error structure. Despite its improved forecasting performance over the famous Lee–Carter (LC) model, the constraint that only the effects of the same and neighboring cohorts are significant can be too restrictive. In this study, we adopt the concept of hyperbolic memory to the spatial dimension and propose a hyperbolic STVAR (HSTVAR) model. Retaining all desirable features of the STVAR, our model uniformly beats the LC, the weighted functional demographic model, STVAR and sparse VAR counterparties for forecasting accuracy, when French and Spanish mortality data over 1950–2016 are considered. Simulation results also lead to robust conclusions. Long-term forecasting analyses up to 2050 comparing the four models are further performed. To illustrate the extensible feature of HSTVAR to a multi-population case, a two-population illustrative example using the same sample is further presented.

1. Introduction

Life expectancy has consistently improved around the world, predominately caused by mortality improvements (Guibert, Lopez, & Piette, 2019). This has spurred serious concerns about the corresponding mortality and longevity risks. Mortality (longevity) risk describes the fact that people are surviving shorter (longer) than expected (Feng & Shi, 2018). For instance, advances in medical science, technological improvements and lifestyle changes may very likely result in higher mortality improvements, which increases the exposure to longevity risk. For the insurance industry, misestimating such improvements will lead to inaccurate premium ratemaking and growing insolvency risks.

Mortality modeling and forecasting have become standard mitigation tools to reduce mortality and longevity risks. Among existing methods, one popular stream is based on the seminal work of Lee and Carter (1992), which is known as the famous Lee–Carter (LC) model. Many extensions of the LC model have been extensively discussed in the literature—see, for example, Booth, Hyndman, Tickle, and De Jong (2006), Renshaw and Haberman (2006) and Barrieu, Bensusan, El Karoui, Hillairet, Loisel, Ravanelli, and Salhi (2012). Among those based on functional principal component analysis (FPCA), the weighted functional demographic model (FDM.W) studied in Shang, Booth, and Hyndman (2011) demonstrates outstanding

* Corresponding author.
E-mail address: yanlin.shi@mq.edu.au (Y. Shi).

https://doi.org/10.1016/j.ijforecast.2020.05.003
0169-2070/

forecasting performance. Recently proposed FPCA-type methods also work well for grouped populations (Shang & Hyndman, 2017). Another important stream focuses on applying high-dimensional vector autoregressive (VAR) models, which can allow more flexible temporal modeling than the LC approach (Guibert et al., 2019). Our research contributes to significant extensions of those VAR-type models.

There are two major concerns with all VAR-type models for mortality modeling and forecasting. In terms of statistical modeling, the issue that the number of variables (e.g., age) is usually greater than the sample size (e.g., year) is widely problematic. Consequently, there are insufficient data to fit a full VAR-type model. Hence, the first concern is an appropriate dimension-reduction approach. A recent paper by Li and Lu (2017) addresses this by focusing on the cohort effects and adopting a spatial–temporal VAR (STVAR) model with a restrictive coefficient matrix, such that only mortality rates for neighboring ages can interact. Second, VAR-type models need to be stationary to produce meaningful estimates and forecasts. Considering that the age-wide mortality rates are non-stationary time series, Li and Lu (2017) employ an age-coherent structure to impose a powerful constraint on the coefficient matrix and impose smoothness penalties. Hence, the forecast mortality rates will be smoothed and those of neighboring age groups will not diverge in the long run. Also, the proposed STVAR model has many attractive statistical properties, such as the co-integration of dependent variables, closed-form solutions and a simple error structure.

Despite the effectiveness of this approach, some new concerns are brought up. The major issue is that the STVAR considers the sparsity of the coefficient matrix in a relatively ad-hoc manner. Within such a framework, only lagged mortality rates of ages $x-2$, $x-1$ and x can affect contemporary rates of age x. However, exclusively assuming significant effects between those neighboring younger cohorts can be overly restrictive, because information from other younger age groups might further contribute to the mortality forecast. By contrast, Guibert et al. (2019) work on the mortality improvements directly (differenced log rates) and adopt a pure data-driven approach. They employ the sparse VAR (SVAR) model with an elastic-net (ENET) penalty estimation method. This essentially allows for a more flexible coefficient matrix than that considered in STVAR. However, the price of such flexibility is the loss of coherence and the questionable interpretation of the cohort effects. For example, unlike STVAR, the SVAR model does not impose any constraints or penalties. Consequently, age coherence is not guaranteed for the forecast rates in STVAR. Moreover, despite the power of the ENET method, the resulting coefficient matrix is fully data-driven and thus may very likely be uninterpretable. For instance, the empirical analysis in Guibert et al. (2019) indicated that the mortality improvements of age 45 were significantly influenced by the lagged evolution of age 95, which is counterintuitive. As a result, the desirable cohort effects could be overwhelmed and unexplainable in the SVAR model.

To address these concerns, we propose a novel hyperbolic STVAR model (HSTVAR), using the concept of hyperbolic memory. This idea has been widely studied in financial time series to describe the persistent influence of shocks (see Baillie, Bollerslev, and Mikkelsen (1996), Davidson (2004) and Ho and Shi (2020) for examples of related models; and see Choi, Yu, and Zivot (2010), Feng and Shi (2017) and Gao, Ho, and Shi (2020) for examples of their applications). Unlike short-memory autoregressive models, which allow for only geometric decay, hyperbolic memory also allows for hyperbolic decay, for which dependency reduces slowly and is still persistent in the long term. For our purpose, the benefit of hyperbolic memory (also known as long memory) is its flexibility of dependency structure with only one required parameter. Applying it to the STVAR framework, this parametric structure will enable the influence of the mortality rates of younger cohorts on that of the current age to decay slowly or quickly. As preliminary evidence of its usefulness, we consider the logged total French mortality rates from ages 0–100 over the years 1950–2016. The averaged correlations between those rates of the current and younger cohorts are plotted in Fig. 1[1] Clearly, those correlations demonstrate a declining trend slower than a geometric pattern but faster than a unit-root fashion. Such a trend belongs to a hyperbolic structure and can be effectively modeled via our proposed approach.

To demonstrate the effectiveness of the proposed model, we provide empirical evidence of total mortality data of France, sourced from Human Mortality Database. Using the crude rates ranging from 1950 to 2016, we systematically compare the forecasting performance of the LC, FDM.W, STVAR, HSTVAR and SVAR models on the age groups of 0 to 100. As measured by the root mean squared error (RMSE), the HSTVAR model consistently beats other competing models up to the 16-steps-ahead forecasting horizon for all four datasets. This consistently holds when the mortality data of Spain, a geographic neighbor of France, is further considered over the same sample period, with additional robust simulation results. We then conduct long-term forecasting analysis to compare the life expectancies calculated by different models up to 2050. Finally, when French and Spanish data are jointly modeled, HSTVAR is extensible to model this case and still consistently outperforms the Li–Lee model (a multi-population extension of LC (Li & Lee, 2005)), the weighted coherent functional demographic model (a multi-population extension of FDM.W (Hyndman, Booth, & Yasmeen, 2013)), the STVAR model and the SVAR model.

The contributions of this paper are mostly attributed to the introduction of hyperbolic memory to mortality forecasting. The proposed HSTVAR model significantly complements the recent study of Li and Lu (2017), and largely improves the flexibility to model the effects of all younger cohorts. Also, the merits of the STVAR are all retained in the new model, including the desirable co-integration features, the existence of closed-form solutions, an effective error structure and a straightforward extension to the

[1] For instance, the first lag is the average correlations between logged mortality rates of ages 1–100 over 1951–2016 and those of ages 0–99 over 1950–2015.

Please cite this article as: L. Feng, Y. Shi and L. Chang, Forecasting mortality with a hyperbolic spatial temporal VAR model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.05.003.

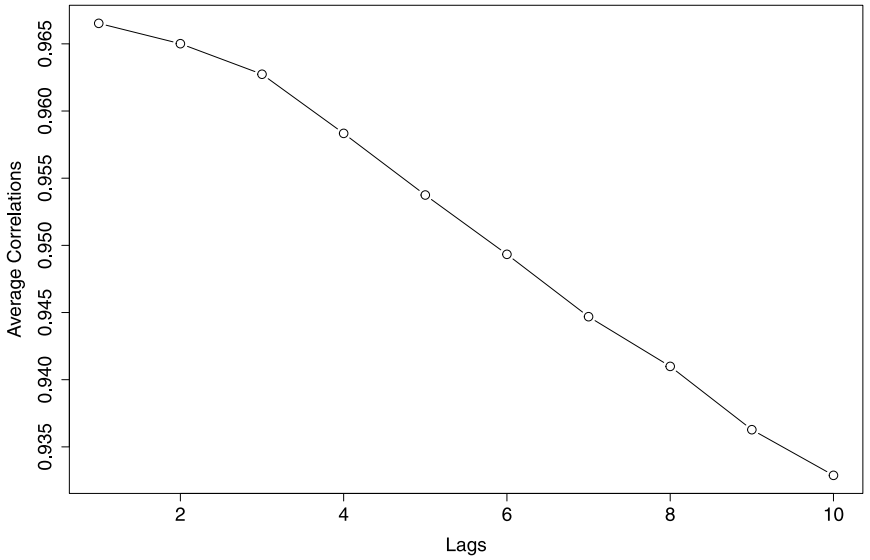


Fig. 1. Averaged correlations between logged mortality rates of the current and younger (lagged) cohorts.

multi-population scenario. In addition, we systematically study the forecasting performance of the HSTVAR model for French and Spanish mortality data. Its superiority over all other competing models indicates the potential and usefulness of our method to model and forecast mortality rates in other contexts.

The rest of this paper is organized as follows. In Section 2, we describe the FDM.W model. The STVAR and proposed HSTVAR models are discussed in Section 3. We present the statistical properties and estimation procedure of the HSTVAR model in Section 4. In Section 5, we conduct empirical and simulation studies with a single population case, and in Section 6, we discuss the multi-population extension. Finally, Section 7 concludes the paper.

2. The weighted functional demographic model

Utilizing a functional data paradigm Ramsay and Silverman (2007), and Hyndman and Ullah (2007) proposed a nonparametric method for modeling and forecasting log mortality rates, namely, the Hyndman–Ullah (HU) method. To further allow the mean and the basis function in the HU method to focus more on recent data, Shang et al. (2011) introduced a weighted HU method, namely, the weighted functional demographic model (FDM.W), which adds geometrically decaying weights in estimation of these quantities. FDM.W substantially extends the popular LC method in the following ways.

1. The log mortality rate is assumed to be a smooth function of age that is observed with error. That is,

$$\ln m_{x,t} = f_t(x) + \sigma_t(x)\epsilon_{x,t}, \tag{1}$$

where $m_{x,t}$ is the observed mortality rate at age x and year t, $f_t(x)$ is the underlying continuous and smooth function and $\sigma_t(x)$ further allows noise to vary with x, whereas $\epsilon_{x,t}$ is an independently and identically distributed normal random variable. In both the HU and the FDM.W methods, $f_t(x)$ is estimated via penalized regression splines with a partial monotonic constraint (see Ramsay et al. (1988)).
2. The weighted functional mean $a(x)$ is then estimated using the weighted average

$$\widehat{a}(x) = \sum_{t=1}^{T} w_t f_t(x), \tag{2}$$

where the weights are defined as $w_t = \eta(1-\eta)^{T-t}$, and η denotes a geometrically decaying weight parameter. Details of estimating η can be found in Hyndman and Shang (2009).
3. Given $\widehat{a}(x)$, a set of weighted curves $w_t(f_t(x) - \widehat{a}(x))$ for $t = 1, \ldots, T$ is decomposed into orthogonal weighted principal components and the corresponding uncorrelated principal component scores using functional principal component analysis (FPCA), such that

$$f_t(x) = \widehat{a}(x) + \sum_{j=1}^{J} b_j(x) k_{t,j} + e_t(x), \tag{3}$$

where $b_j(x), j = 1, \ldots, J$ is a set of first J weighted principal components, $k_{t,j}, j = 1, \ldots, J$ is a set of uncorrelated principal component scores and $e_t(x)$ is the residual function with mean 0. According to Hyndman and Booth (2008) and Shang et al. (2011), we consider $J = 6$, which should be larger than any of the components required.[2]

[2] Booth et al. (2006) and Renshaw and Haberman (2006) also used more than one set of principal components and scores to improve the accuracy of the original LC model.

Please cite this article as: L. Feng, Y. Shi and L. Chang, Forecasting mortality with a hyperbolic spatial temporal VAR model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.05.003.

4. Lastly, a univariate time series model, such as the ARIMA or an exponential smoothing state space model (Hyndman & Shang, 2009), can be implemented to produce the h-steps-ahead forecast of $k_{T+h,j}$, denoted as $\widehat{k}_{T+h,j}$. In this paper, we forecast $k_{T+h,j}$ using an ARIMA (p, d, q) model, with the orders p, d and q automatically chosen to minimize the corrected Akaike Information Criterion (AICc) as described in Hyndman and Khandakar (2008). Then, conditioning on the data $\ln m_{x,t}$, $t = 1, \ldots, T$ and the set of weighted functional principal components $b_j(x)$, $j = 1, \ldots, J$ estimated in (3), the h-steps-ahead forecast of $\ln m_{x,T+h}$ is achieved by

$$\ln \widehat{m}_{x,T+h} = \widehat{a}(x) + \sum_{j=1}^{J} b_j(x)\widehat{k}_{T+h,j}. \tag{4}$$

Shang et al. (2011) compared the point and interval forecast accuracy and bias of various extensions of the LC method, including the HU and FDM.W methods. They found that the FDM.W method provides the most accurate point forecasts of mortality rates based on one-step forecast errors. Therefore, in addition to the benchmark of the LC model, we include the FDM.W method to compare its performance with the VAR-type models considered in our study.

3. The vector autoregressive (VAR) model

In contrast to a factor model like LC, another popular stream to study and forecast mortality rates is the VAR model. However, the application of the VAR model to mortality data brings up two issues. First, the VAR model requires the dependent variables to be stationary. Without modification or constraints, $\ln m_{x,t}$ is clearly trending and therefore non-stationary. Second, there are more unknown parameters (p) than observations (T) in the standard VAR framework. Suppose we have N age groups; even in the simplest VAR(1) case, for each $\ln m_{x,t}$, all N lagged log mortality rates need to be included. Thus, the total number of parameters to be estimated is $p = N(N + 1)$ (including N intercepts). Considering that we usually only have dozens of yearly data to work with, the $p \gg NT$ issue will arise for an intermediate N such as 50.

3.1. The spatial temporal VAR (STVAR) model

To address these two issues, Li and Lu (2017) proposed the STVAR model. On the temporal dimension, it considers the Granger causality and co-integration to resolve the stationarity problem. For age groups, the STVAR model utilizes the sparse spatial information to reduce the dimensionality of p. Let $y_{x,t} = \ln m_{x,t}$; this leads to the following specification.[3]

$$\begin{aligned} y_{1,t} &= m_1 + y_{1,t-1} + \varepsilon_{1,t} \\ y_{2,t} &= m_2 + (1-\alpha_2)y_{2,t-1} + \alpha_2 y_{1,t-1} + \varepsilon_{2,t} \\ y_{i,t} &= m_i + (1-\alpha_2-\beta_i)y_{i,t-1} + \alpha_i y_{i-1,t-1} \\ &\quad + \beta_i y_{i-2,t-1} + \varepsilon_{i,t} \end{aligned} \tag{5}$$

where $i = 3, 4, \ldots, N$, and $t = 1, 2, \ldots, T$. $\varepsilon_{i,t}$ is assumed to follow a multi-Gaussian distribution with $\mathbf{0}$ $(N \times 1)$ means and a $\boldsymbol{\Sigma}$ $(N \times N)$ variance–covariance matrix. Rewritten in a VAR(1) form, we have

$$\mathbf{Y_t} = \mathbf{M} + \mathbf{B}\mathbf{Y_{t-1}} + \boldsymbol{\varepsilon_t} \tag{6}$$

where $\mathbf{Y_t} = (y_{1,t}, y_{2,t}, \ldots, y_{N,t})'$, $\mathbf{M} = (m_1, m_2, \ldots, m_N)'$, $\boldsymbol{\varepsilon_t} = (\varepsilon_{1,t}, \varepsilon_{2,t}, \ldots, \varepsilon_{N,t})'$ and

$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots \\ \alpha_2 & 1-\alpha_2 & 0 & \cdots & \cdots \\ \beta_3 & \alpha_3 & 1-\alpha_3-\beta_3 & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \cdots & 0 & \beta_N & \alpha_N & 1-\alpha_N-\beta_N \end{bmatrix} \tag{7}$$

This specification leads to attractive features. First, as shown in Li and Lu (2017), all neighboring age pairs $y_{i,t}$ and $y_{i+1,t}$ are co-integrated with order (1,-1). This successfully solves the stationarity issue. Second, the total number of parameters is largely reduced to $p = 3N - 3$ and no longer greater than NT.

Forecasting is performed in an iterative fashion, where

$$\begin{aligned} \hat{\mathbf{Y}}_{t+1} &= \hat{\mathbf{M}} + \hat{\mathbf{B}}\mathbf{Y_t} \\ \hat{\mathbf{Y}}_{t+h} &= \hat{\mathbf{M}} + \hat{\mathbf{B}}\hat{\mathbf{Y}}_{t+h-1} \end{aligned} \tag{8}$$

and $h > 1$. Also, to ensure the age coherence, Li and Lu (2017) conducted the estimation in a penalized least-squares (PLS) fashion. The details are discussed in Section 4.

3.2. The hyperbolic spatial temporal VAR (HSTVAR) model

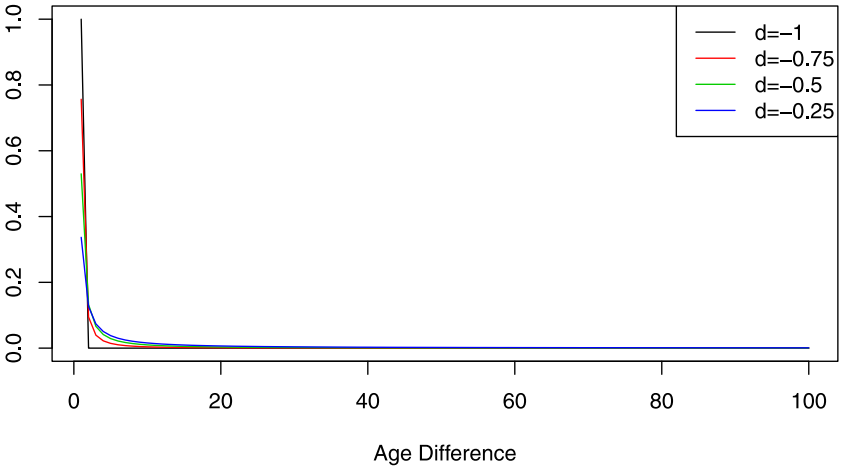
Despite the effectiveness of the STVAR model, the sparsity of the coefficient matrix $\mathbf{B}$ is quite restrictive. For each $y_{j,t}$, by enforcing all coefficients of $y_{i-j,t}$ $(j > 3)$ to be exactly 0, this specification ignores all potential cohort effects for those younger cohorts. In other words, only the same and neighboring cohort effects are considered in the STVAR model. This is an ad-hoc structure and might not suit all types of mortality data.

In order to retain the advantages of the STVAR model, we adopt the idea of hyperbolic memory on the age dimension and propose an HSTVAR model. We follow Li and Lu (2017) but allow for more flexible cohort effects for all younger cohorts, by modifying the coefficient matrix to

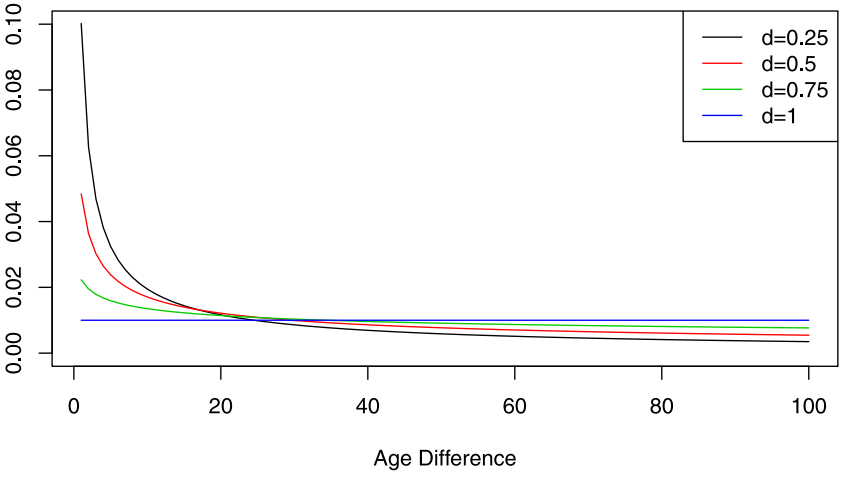
$$\mathbf{B} = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots \\ \beta_2 & 1-\beta_2 & 0 & \cdots & \cdots \\ \beta_3 w_{32} & \beta_3 w_{31} & 1-\beta_3 & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \beta_N w_{N,N-1} & \cdots & \beta_N w_{N,2} & \beta_N w_{N,1} & 1-\beta_N \end{bmatrix} \tag{9}$$

[3] Note that the STVAR and HSTVAR models discussed in this paper only consider one lag in the VAR specification. Including more lags is of great interest to improve the forecasting accuracy further, but may introduce more difficulty in the explanations of parameters. A comprehensive analysis on this is beyond the scope of this paper and remains for future work. We thank the anonymous referee for this suggestion.

Please cite this article as: L. Feng, Y. Shi and L. Chang, Forecasting mortality with a hyperbolic spatial temporal VAR model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.05.003.



(a) $d < 0$



(b) $d > 0$

Fig. 2. Demonstration of hyperbolic weights over younger ages.

where for age $i > 2$, we impose the constraint that $\sum_{k=1}^{i-1} w_{ik} = 1$. In contrast to (7), we distribute the impact of the age $i-2$ and $i-1$ ($\alpha_i+\beta_i$) to all $i-k$ ($0 < k < i-1$). The distribution to age $i-k$ is based on the corresponding weight w_{ik}, where

$$w_{ik} = \frac{\delta_k(d)}{\sum_{l=1}^{i-1}\delta_l(d)},\ \delta_k(d) = \frac{k-1+d}{k}\delta_{k-1}(d) \text{ and } \delta_0(d) = 1 \tag{10}$$

The specification of $\delta_k(d)$ is inspired by hyperbolic memory (also known as long memory) widely studied in finance time series research. First proposed by Hosking (1981), $\delta_k(d)$ allows for the impact of a shock to die out hyperbolically over time. Unlike geometric decay (or short memory), which is featured in autoregressive-type models, the impact of a shock only decays hyperbolically and is therefore much more persistent. The speed of this decay is measured by d, which is commonly known as the hyperbolic-memory (long-memory) parameter. In our case, it can be seen that w_{ik} has inconsistent signs when $d < -1$, w_{ik} increases with k when $d > 1$, and w_{ik} decreases hyperbolically (geometrically) when $d > 0$ ($d < 0$). Hence, we impose the constraint that $|d| \leq 1$, which reasonably assumes that the impact of ages further away should be smaller. To illustrate these differences, we plot $w_{100,k}$ in Fig. 2 with different values of d.

In Fig. 2(a), when d is closer to -1, the decline in $w_{100,k}$ over k is faster with the growth in k. In all cases, $w_{100,k}$ reduces to 0 fairly quickly. When $d > 0$, Fig. 2(b) suggests that a larger d leads to a slower decay in $w_{100,k}$. When $d = 1$, all weights are equal, which is standard for extreme cases where cohort effects are persistent for all younger cohorts. Therefore, using the flexible weights $w_{i,k}$, our proposed model can nest both hyperbolic and geometric spatial relationships on the age dimension. This is expected to outperform the more constrained sparse coefficient matrix **B** as considered in Li and Lu (2017).

Using the HSTVAR model, forecasting is conducted in the same way as described in (8). Estimations and other technical features of the HSTVAR model are similar to those of STVAR and are discussed in Section 4.

Please cite this article as: L. Feng, Y. Shi and L. Chang, Forecasting mortality with a hyperbolic spatial temporal VAR model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.05.003.

3.3. The sparse VAR (SVAR) model

In contrast to Li and Lu (2017), a recent study by Guibert et al. (2019) employed the SVAR model to mortality improvements to address the two common issues of VAR-type models. First, by working on $\triangle y_{x,t} = y_{x,t} - y_{x,t-1}$ and assuming that $y_{x,t}$ is $I(1)$ for all ages, the dependent variables are $I(0)$ and therefore stationary. Second, using an elastic-net (ENET) penalty estimation, the SVAR model adopts a pure data-driven method to select the non-zero coefficients of $\triangle y_{x-j,t}$ of all $\triangle y_{x,t}$. To be consistent with the STVAR and HSTVAR models, we only consider one lag.[4] Thus, the SVAR model has a VAR(1) specification:

$$\triangle \mathbf{Y_t} = \mathbf{M} + \mathbf{B}\triangle \mathbf{Y_{t-1}} + \boldsymbol{\varepsilon_t} \tag{11}$$

$\triangle \mathbf{Y_t} = (\triangle y_{1,t}, \triangle y_{2,t}, \ldots, \triangle y_{N,t})'$ and the sparsity (zeros) of $\mathbf{B}$ is determined by the LASSO (L1) penalty during estimation without any constraints. Forecasting is performed similarly to (8), as follows:

$$\begin{aligned} \triangle \hat{\mathbf{Y}}_{\mathbf{t+1}} &= \hat{\mathbf{M}} + \hat{\mathbf{B}}\triangle \mathbf{Y_t} \\ \triangle \hat{\mathbf{Y}}_{\mathbf{t+h}} &= \hat{\mathbf{M}} + \hat{\mathbf{B}}\triangle \hat{\mathbf{Y}}_{\mathbf{t+h-1}} \end{aligned} \tag{12}$$

where $h > 1$, and $\hat{\mathbf{Y}}_{t+h} = \mathbf{Y_t} + \sum_{l=1}^{h} \triangle \hat{\mathbf{Y}}_{\mathbf{t+l}}$.

Despite the power of the ENET algorithm, there are new issues introduced by the SVAR model. First, the data-driven method can lead to unexplainable estimates of the cohort effects. For instance, the results of Guibert et al. (2019) suggest that mortality improvements of age 45 are significantly affected by those of age 100, which is counterintuitive. Second, without any smoothing penalty, age coherence cannot be guaranteed. Consequently, as pointed out by Feng and Shi (2018), the SVAR model may not necessarily improve the accuracy of mortality forecasting, since irrelevant independent variables may introduce noise.

4. Stationarity, estimation and error structure of the HSTVAR model

As the HSTVAR is a more flexible extension of the STVAR model, estimation is performed in the same fashion and all of the technical features of STVAR are retained.

4.1. Stationarity of the HSTVAR model

As described in (9), the coefficient matrix $\mathbf{B}$ is constrained such that each row sums to exactly 1. More generally speaking, we can rewrite the HSTVAR model as follows:

$$\begin{aligned} y_{1,t} &= m_1 + y_{1,t-1} + \varepsilon_{1,t} \\ y_{2,t} &= m_2 + (1 - b_{21})y_{2,t-1} + b_{21}y_{1,t-1} + \varepsilon_{2,t} \\ y_{3,t} &= m_3 + (1 - b_{31} - b_{32})y_{3,t-1} + b_{31}y_{1,t-1} \\ &\quad + b_{32}y_{2,t-1} + \varepsilon_{3,t} \\ y_{i,t} &= m_i + (1 - \sum_{l=1}^{i-1} b_{il})y_{i,t-1} + \sum_{l=1}^{i-1} b_{il}y_{l,t-1} + \varepsilon_{i,t} \end{aligned} \tag{13}$$

where $i > 3$. Compared to (9), $b_{ik} = \beta_i w_{i,i-k}$ and $\sum_{l=1}^{i-1} b_{il} = \beta_i$.

Proposition 1. *Indeed, under specification (13) and the assumption that all residuals $\varepsilon_{i,t}$ are stationary and all $0 < \beta_{ij} < 1$, different component processes $y_{i,t}$ and $y_{j,t}$ are co-integrated, with co-integration vector (1, −1).*

Proof. See Appendix A.1. □

Therefore, as long as β_i for $i = 2, 3, \ldots, N$ in (9) all fall in the range (0,1), $y_{x,t}$ is co-integrated, as with the STVAR model. This successfully resolves the stationarity issue for VAR-type models.

4.2. Estimation with PLS

To ensure age coherence, Li and Lu (2017) introduced smoothing parameters for αs and βs in (7) and obtained the estimates via PLS. Following the same design, given that the tuning parameter d, which determines w_{ik}, is known, we have the objective function of HSTVAR as described below.

$$\begin{aligned} LF_1 = &\sum_{i=2}^{N}\sum_{t=2}^{T}\Bigg[y_{i,t} - m_i - (1-\beta_i)y_{i,t-1} \\ &- \sum_{l=1}^{i-1}\beta_i w_{i,i-l}y_{l,t-1}\Bigg]^2 \\ &+ \sum_{t=2}^{T}(y_{1,t} - y_{1,t-1} - m_1)^2 + \lambda_m \sum_{i=2}^{N}(m_i - m_{i-1})^2 \\ &+ \lambda_\beta \sum_{i=3}^{N}(\beta_i - \beta_{i-1})^2 \end{aligned} \tag{14}$$

where λ_m and λ_β are pre-selected smoothing parameters of $\mathbf{M}$ and $\boldsymbol{\beta}$, respectively, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_N)'$. The estimates of $\boldsymbol{M}$ and $\boldsymbol{\beta}$ can then be derived by minimizing LF_1. With greater values of λ_m (λ_β), the fitted $\mathbf{M}$ ($\boldsymbol{\beta}$) will be smoother across ages. When both λ_m and λ_β are equal to 0, this is a special case of PLS with no penalty and reduces to the usual ordinary least-squares case. With smoother changed $\mathbf{M}$ and $\boldsymbol{\beta}$, the forecast $\hat{\mathbf{Y}}_{t+h}$ is expected to become smoother from $x-1$ to x. In other words, mortality rates will not diverge for neighboring ages, which ensures age coherence. Similar to STVAR, the estimates of all parameters have closed-form solutions, as demonstrated below.

Proposition 2. *Estimates of all m_i and β_i for $i = 1, 2, \ldots, N$ for the objective function (14) have closed-form solutions.*

Proof. See Appendix A.2. □

From the above proof, the main difference between the HSTVAR and STVAR models can also be seen as the assumed influence of the "exogenous" lagged mortality rates on $y_{x,t}$. More specifically, the STVAR model only

[4] We also evaluated the case when a VAR(7) specification was modeled, as conducted in Guibert et al. (2019). Although the forecasting performance of SVAR did improve in that case, our baseline conclusions of all competing models were unaffected. The results of the SVAR(7) specification are available upon request.

Please cite this article as: L. Feng, Y. Shi and L. Chang, Forecasting mortality with a hyperbolic spatial temporal VAR model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.05.003.

allows two independent impacts of $y_{x-1,t-1}$ and $y_{x-2,t-1}$. By contrast, the HSTVAR model requires only one such "exogenous" rate, $y^{w}_{x,t-1}$, which, however, adaptively contains all information of the younger ages (from 0 to $x-1$). For an appropriately chosen d, we would expect that $y^{w}_{x,t-1}$ is more informative than the lagged rates of just the closest two younger ages ($x-1$ and $x-2$). Hence, HSTVAR improves STVAR with regard to forecasting accuracy, as more effective information is utilized.

Altogether, there are three tuning parameters for our HSTVAR model, d, λ_m and λ_β. They need to be selected before performing the estimation. A common solution for this is to employ cross-validation. However, due to the time-series nature, related methods, such as leave-one-age-group-out, are inapplicable to the HSTVAR model. Hence, we employ the procedure discussed in Hyndman and Athanasopoulos (2018) to perform cross-validation, which is also known as "evaluation on a rolling forecasting origin". The basic algorithm is explained below:

1. Identify the first training sample (e.g., $y_{i,2}$, $y_{i,3}$, …, $y_{i,0.7T}$ for $i = 1, 2, \ldots, N$) out of the entire dataset;
2. Given a set of d, λ_m and λ_β, use the training sample to fit the HSTVAR model and obtain the one-step-ahead forecast $\hat{y}_{i,0.7T+1}$;
3. Extend the training set to include $y_{i,0.7T+1}$ and refit the HSTVAR model to obtain the one-step-ahead forecast $\hat{y}_{x,0.7T+2}$;
4. Repeat steps 2–3 until $\hat{y}_{i,T}$ is generated; and
5. Calculate the root of mean squared error (RMSE) as

$$\sqrt{\frac{1}{0.3T \times N}\sum_{i=1}^{N}\sum_{h=1}^{0.3T}(y_{i,0.7T+h} - \hat{y}_{i,0.7T+h})^2}$$

d, λ_m and λ_β are then chosen as those with the smallest RMSE via a grid search, where the potential ranges of d, λ_m and λ_β are (-1,1), $[0, \infty)$ and $[0, \infty)$, respectively.

Compared to the STVAR model, the HSTVAR model has three tuning parameters (smoothing penalties for ms, αs and βs as in (5)) to choose. Since both the STVAR and HSTVAR models have closed-form solutions, the computational cost and intensity are fairly similar. In other words, by adopting the idea of hyperbolic memory, the proposed STVAR model enables us to examine cohort effects of younger cohorts, with no additional computational cost.

4.3. Error structure of the HSTVAR model

Apart from estimating $\boldsymbol{\theta}$, it is also essential to analyze the dependency structure of the residuals. Such error structures can capture the contemporaneous interactions between the mortality rates for different ages. Similar to most popular mortality models, we assume that the residual vector $\boldsymbol{\varepsilon_t}$ of our proposed HSTVAR model follows a multi-Gaussian distribution with $\mathbf{0}$ means and a $\boldsymbol{\Sigma}$ variance–covariance matrix. That is, $\boldsymbol{\varepsilon_t} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Indicated by $\sigma^{\varepsilon}_{ij}$, this ith row and jth column component of $\boldsymbol{\Sigma}$ then measures the covariance of $\varepsilon_{i,t}$ and $\varepsilon_{j,t}$.

Existing studies have investigated the dependence among residuals in mortality modeling. Biffis and Millossovich (2006), Debón, Montes, Mateu, Porcu, and Bevilacqua (2008) and Guibert et al. (2019) used a parametric covariance function on age distance $|i-j|$ to capture the dependence structure of the residuals. As discussed in Li and Lu (2017), the variance–covariance estimation under a typical VAR model requires $\frac{N(N+1)}{2}$ number of parameters, which is prohibitively large in mortality modeling (e.g., $N = 101$ when ages from 0 to 100 are examined). Therefore, Li and Lu (2017) proposed a constrained spatial autoregressive model to model the residual for the ith age $\varepsilon_{i,t}$ by its neighboring terms $\varepsilon_{i-1,t}$ and $\varepsilon_{i+1,t}$. This specification significantly reduces the number of parameters from $\frac{N(N+1)}{2}$ to $2(N-1)$. However, similar to estimating the coefficient matrix $\boldsymbol{B}$, exclusively assuming the dependency between residuals from neighboring ages can be overly restrictive. As preliminary evidence, we considered the correlations between $\varepsilon_{i,t}$ and $\varepsilon_{i-l,t}$, averaged over $i = 1, 2, \ldots, 100$, where $l = 1, 2, \ldots, 10$. The residuals were extracted from the HSTVAR model fitted for the French mortality data over 1950–2000, as examined in Section 5. We plot these correlations in Fig. 3. Clearly, although the correlations reduce more quickly than those shown in Fig. 1, the reduction is still much slower than a geometric fashion, indicating possible hyperbolic memory.

Without significantly increasing the computational cost, we therefore impose the hyperbolic memory for the residual of the HSTVAR model to account for the patterns shown in Fig. 3. Similar to that considered by Li and Lu (2017), we have

$$\begin{aligned}
\varepsilon_{1,t} &= a_1 \sum_{k=1}^{N-1} w^{U}_{1,k}\varepsilon_{1+k,t} + \eta_{1,t} \\
\varepsilon_{i,t} &= a_i \sum_{k=1}^{N-i} w^{U}_{i,k}\varepsilon_{i+k,t} + c_i \sum_{l=1}^{i-1} w^{L}_{i,l}\varepsilon_{i-l,t} + \eta_{i,t} \\
\varepsilon_{N,t} &= c_N \sum_{l=1}^{N-1} w^{L}_{N,l}\varepsilon_{N-l,t} + \eta_{N,t}
\end{aligned} \tag{15}$$

where $1 < i < N$, $w^{U}_{i,k} = \delta_k(e)/\sum_{k=1}^{N-i}\delta_k(e)$ and $w^{L}_{i,l} = \delta_l(e)/\sum_{l=1}^{i-1}\delta_l(e)$, with $\delta_k(e)$ and $\delta_l(e)$ defined in the same ways as in (10). For each age, $\eta_{i,2}, \eta_{i,3}, \ldots, \eta_{i,T}$ is an independently and identically distributed Gaussian sequence with 0 mean and σ_i^2. To ensure stationarity of $\varepsilon_{i,t}$, we also require that $a_i \geq 0$, $c_i \geq 0$ and $a_i + c_i < 1$ for each age.

After the HSTVAR model is fitted, the residuals $\boldsymbol{\varepsilon_t}$ can be obtained. The objective function can then be defined as follows.

$$\begin{aligned}
LF_2 = &\sum_{i=2}^{N-1}\sum_{t=2}^{T}\left[\varepsilon_{i,t} - a_i\sum_{k=1}^{N-i} w^{U}_{i,k}\varepsilon_{i+k,t} - c_i\sum_{l=1}^{i-1} w^{L}_{i,l}\varepsilon_{i-l,t}\right]^2 \\
&+ \sum_{t=2}^{T}(\varepsilon_{1,t} - a_1\sum_{k=1}^{N-1} w^{U}_{1,k}\varepsilon_{1+k,t})^2 \\
&+ \sum_{t=2}^{T}(\varepsilon_{N,t} - c_N\sum_{l=1}^{N-1} w^{L}_{N,l}\varepsilon_{N-l,t})^2 \\
&+ \lambda_a\sum_{i=2}^{N-1}(a_i - a_{i-1})^2 + \lambda_c\sum_{i=3}^{N}(c_i - c_{i-1})^2
\end{aligned} \tag{16}$$

Please cite this article as: L. Feng, Y. Shi and L. Chang, Forecasting mortality with a hyperbolic spatial temporal VAR model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.05.003.

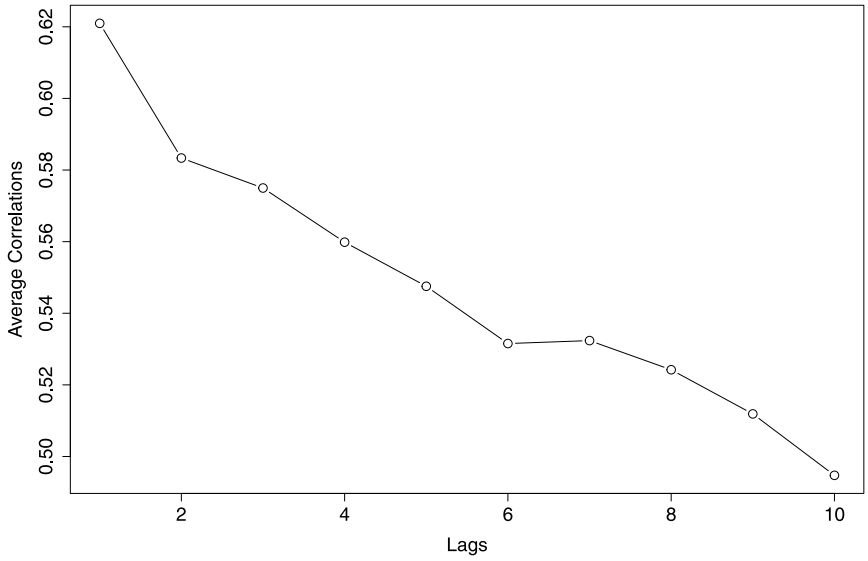


Fig. 3. Averaged correlations between residuals of lagged ages.

where λ_a and λ_c are pre-selected smoothing penalties for a_is and c_is, respectively. As no equality constraint is applied in (15), closed-form solutions of parameters exist as those for the usual PLS problem.

To determine the hyperbolic weights, we also need to know e in advance. Thus, as for LF_1, there are three tuning parameters to choose before performing the estimation. Since no time dependency is imposed, this can be done via the usual cross-validation. In this paper, we choose a 10-fold setting, as one of the standard approaches.

After the estimates of a_is and c_is are obtained, we can produce the estimated $\boldsymbol{\Sigma}$. Following Li and Lu (2017), we can rewrite (15) to $\mathbf{B}_{\boldsymbol{\varepsilon}}\boldsymbol{\varepsilon}_{\mathbf{t}} = \boldsymbol{\eta}_{\mathbf{t}}$, where $\boldsymbol{\varepsilon}_{\mathbf{t}} = (\varepsilon_{1,t}, \varepsilon_{2,t}, \ldots, \varepsilon_{i,t})'$, $\boldsymbol{\eta}_{\mathbf{t}} = (\eta_{1,t}, \eta_{2,t}, \ldots, \eta_{i,t})'$ and $\mathbf{B}_{\boldsymbol{\varepsilon}}$ is as in Box I. The estimated $\boldsymbol{\Sigma}$ is then equal to $(\mathbf{B}_{\boldsymbol{\varepsilon}}^{-1})' Diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)\mathbf{B}_{\boldsymbol{\varepsilon}}^{-1}$. To reduce computational intensity, we can set σ_i^2 to the sample variance of η_i, for $i = 1, 2, \ldots, N$.

The error structure described above nests that explained in Li and Lu (2017) as a special case. In particular, when $w_{i,1}^U = 1$ and $w_{j,1}^L = 1$ for $i = 1, 2, \ldots, N-1$ and $j = 2, 3, \ldots, N$, $\mathbf{B}_{\boldsymbol{\varepsilon}}$ reduces to a square matrix of 0s, except for diagonals of 1 and sub-diagonals of $\hat{a}_i$ and $\hat{c}_j$. This only considers the contemporary impact of mortality of ages $x-1$ and $x+1$ on that of age x. For both the STVAR and HSTVAR models, the estimated $\boldsymbol{\Sigma}$ will eventually be used to generate prediction intervals via simulation, as described in Section 5.

5. Empirical application

In this paper, we focus on mortality data of France obtained from the Human Mortality Database. Following Booth et al. (2006), we selected an opportune range of data starting from 1950 to 2016 in order to have a reliable and complete dataset. Age groups 0–100 were included in the sample. As a geographic neighboring country, the mortality data of Spain over the same period were also investigated for comparison. The crude total mortality rates were studied,[5] and the log rates are plotted in Fig. 4 across all years. Consistent improvements over time can be observed for both countries. It can also be seen that the Spanish data are relatively rougher, with large variations for ages 20–40. For instance, the mortality rates of age 25 in the 1990s are higher than those in the 1980s.

To illustrate the power of our proposed model, we considered a training sample from 1950–2000 and forecasted the mortality rates from 2001–2016. The out-of-sample results for the LC, FDM.W, STVAR, HSTVAR and SVAR models were compared. We then performed a simulation study to check the robustness of our empirical results. Further, a long-term analysis up to the year 2050 was conducted. In this case, the entire sample from 1950–2016 was fitted to estimate the mortality rates and life expectancy.

5.1. Out-of-sample forecasting performance

Before fitting the models, we selected tuning parameters for STVAR and HSTVAR. Using the cross-validation procedure discussed in Section 4.2, the results are presented in Table 1. For both France and Spain, the corresponding smoothing penalties of the STVAR and HSTVAR models were selected to be close to each other. For French mortality data, λ_m was set to 5.56 for STVAR and 6.67 for HSTVAR. The resulting hyperbolic parameter d was -0.78 and -0.83 for France and Spain, respectively. Since both

[5] Existing studies found that cohort effects could be different for male and female mortality (Renshaw & Haberman, 2006). In relation to that, we also studied males and females separately, and our proposed model outperformed all competing models in both cases. The results are available upon request.

Please cite this article as: L. Feng, Y. Shi and L. Chang, Forecasting mortality with a hyperbolic spatial temporal VAR model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.05.003.

$$\mathbf{B}_{\varepsilon} = \begin{bmatrix} 1 & -\hat{a}_1 w^U_{1,1} & -\hat{a}_1 w^U_{1,2} & \cdots & -\hat{a}_1 w^U_{1,N-1} \\ -\hat{c}_2 w^L_{2,1} & 1 & -\hat{a}_2 w^U_{2,1} & \cdots & -\hat{a}_2 w^U_{2,N-2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -\hat{c}_{N-1} w^L_{N-1,N-2} & \cdots & -\hat{c}_{N-1} w^L_{N-1,1} & 1 & -\hat{a}_{N-1} w^U_{N-1,1} \\ -\hat{c}_N w^L_{N,N-1} & \cdots & \cdots & -\hat{c}_N w^L_{N,1} & 1 \end{bmatrix}_{N\times N}.$$

Box I.

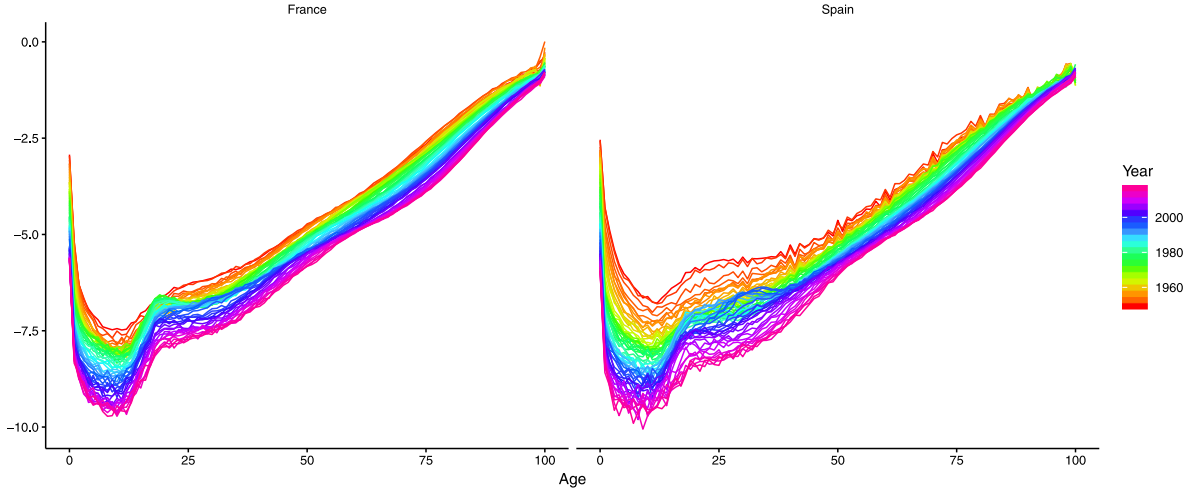


Fig. 4. Mortality data: 1950–2016.

Table 1
Tuning parameter selection.

Country	STVAR					HSTVAR					
	λ_m	λ_α	λ_β	λ_a	λ_c	d	λ_m	λ_β	e	λ_a	λ_c
France	5.56	2.23	4.45	1.34	138.89	−0.78	6.67	2.23	−0.97	1.17	115.56
Spain	18.22	0.45	1.34	0.67	1.12	−0.83	17.33	0.45	−0.94	0.89	0.89

Note: this table displays the chosen tuning parameters for French and Spanish total mortality rates over 1950–2000 for the STVAR models investigated in this paper. For the STVAR (HSTVAR) model, λ_m, λ_α and λ_β (λ_m and λ_β) are smoothing penalties for m, α and β (m and β), respectively; d is the hyperbolic parameter to measure the spatial influence of the younger ages for the HSTVAR model; e is the hyperbolic parameter to measure the spatial influence in its error structure; and λ_a and λ_c are smoothing penalties for a and c, respectively, in the error structures of STVAR and HSTVAR.

are close to -1, given the effects of their own lag, the cohort effects were expected to reduce geometrically in a fairly quick fashion for younger cohorts. Using these tuning parameters, the closed-form solutions could be employed to obtain the estimates. The LC and SVAR models were also fitted as described in Section 2 and Guibert et al. (2019), respectively.

To compare the forecasting performance across all models, we followed Li and Lu (2017) and employed the RMSE. We considered the RMSEs over age groups and time horizons separately and an overall measure as follows:

$$\begin{aligned} RMSE_x &= \sqrt{\frac{1}{16}\sum_{h=1}^{16}(y_{x,T+h} - \hat{y}_{x,T+h})^2} \\ RMSE_h &= \sqrt{\frac{1}{101}\sum_{x=0}^{100}(y_{x,T+h} - \hat{y}_{x,T+h})^2} \\ RMSE_{all,h} &= \sqrt{\frac{1}{101\times h}\sum_{i=1}^{h}\sum_{x=0}^{100}(y_{x,T+i} - \hat{y}_{x,T+i})^2} \end{aligned} \tag{17}$$

Please cite this article as: L. Feng, Y. Shi and L. Chang, Forecasting mortality with a hyperbolic spatial temporal VAR model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.05.003.

Table 2
RMSE over ages summary.

Model	$RMSE_{all,16}$	Mean	p-val.	*Std. Dev.*	Q_1	Q_3
Panel A: France						
LC	0.2158	0.1653	0.0000	0.1394	0.0532	0.2600
FDM.W	0.1396	0.1231	0.0001	0.0661	0.0730	0.1669
STVAR	0.1184	0.1074	0.0000	**0.0500**	0.0606	0.1411
HSTVAR	**0.1098**	**0.0941**	–	0.0567	**0.0454**	**0.1264**
SVAR	0.1411	0.1200	0.0004	0.0746	0.0628	0.1643
Panel B: Spain						
LC	0.2280	0.1828	0.0019	0.1369	0.0665	0.2864
FDM.W	0.2099	0.1616	0.0058	0.1346	0.0509	0.2499
STVAR	0.1834	0.1469	0.0000	0.1104	0.0595	0.2465
HSTVAR	**0.1645**	**0.1334**	–	**0.0968**	**0.0509**	**0.2055**
SVAR	0.2042	0.1674	0.0011	0.1175	0.0692	0.2770

Note: this table displays the RMSE over age groups for the 16-steps-ahead forecasts of French and Spanish total mortality rates. $RMSE_{all,16}$ is the overall RMSE across all ages and time horizons. Mean, *Std. Dev.*, Q_1 and Q_3 are the sample mean, standard deviation, first quartile and third quartile of the RMSEs over age groups, respectively. Bold numbers represent the smallest RMSEs among the four models. p-val. is the *p*-value of the corresponding Diebold–Mariano test, which contrasts the forecasting performance of HSTVAR against that of one of the benchmark models (i.e., LC, FDM.W, STVAR and SVAR) individually.

where $RMSE_x$ ($RMSE_h$) is the RMSE averaged over all 16 forecasting steps (101 age groups) for age group x (time horizon h). $RMSE_{all,h}$ is the overall measure considering both dimensions up to step h. Relevant results for all four models are reported in Tables 2 and 3, as well as in Figs. 5 and 6.

Fig. 5 displays the $RMSE_x$. For the French data, the STVAR and HSTVAR models produced a smaller RMSE than LC and FDM.W for most age groups, especially among the young ages. Although SVAR also led to a smaller RMSE than LC in most cases, its performance was similar to that of FDM.W, and SVAR did not outperform the STVAR and HSTVAR models for almost all ages younger than 60. Comparing the two spatial temporal models, HSTVAR consistently outperformed STVAR across ages 20–60. Their performance for the other age groups was similar. This result suggests that meaningful cohort effects are important when forecasting mortality rates for most French age groups. In addition, allowing for a more flexible influence of younger cohorts can improve the forecasting accuracy of mortality rates, as evidenced by the lower RMSE of HSTVAR compared to STVAR. All of our observations of the French data consistently held for the Spanish case.

Descriptive statistics of $RMSE_x$ are reported in Table 2. For the French data, the mean $RMSE_x$ across all age groups for the HSTVAR model was at least 40%, 20%, 19% and 10% smaller than that for the LC, SVAR, FDM.W and STVAR models, respectively. Q_1 and Q_3 measures further support that the HSTVAR model performed the best among the models. The standard deviation of $RMSE_x$ confirms that the results of HSTVAR were much more narrowly spread (0.0567) than those of the LC (0.1394), FDM.W (0.0500) and SVAR (0.0746). As indicated by $RMSE_{all,16}$, the overall performance of our proposed HSTVAR model was the best among the models. Those observations were largely robust with regard to those for the Spanish mortality rates, in which the HSTVAR model produced the smallest $RMSE_{all,16}$ and the mean, standard deviation, Q_1 and Q_3 of $RMSE_x$. To test the statistical differences, we performed the DM tests described in Diebold and Mariano (2002) for the forecast rates between HSTVAR and each of the other four models. For both the French and Spanish data, the HSTVAR significantly outperformed all the competing models.

Fig. 6 plots the $RMSE_{all,h}$ for h ranging from 1 (2001) to 16 (2016). Distinct differences among all the five models can be observed for both French and Spanish data at all forecast horizons. The STVAR models uniformly beat the LC, FDM.W and SVAR competitors in all scenarios. Compared to the STVAR model, the $RMSE_{all,h}$ for the HSTVAR was smaller in almost all scenarios. Further, with the growth of h (especially from the fifth step onwards), the increment in $RMSE_{all,h}$ was slower for the HSTVAR, suggesting its better performance in the long run. Table 3 reports the $RMSE_h$ at each of the 16 steps. Again, the HSTVAR consistently outperformed the rest, except in very few cases.

Table 3
RMSE over forecasting steps.

Steps	France					Spain				
	LC	FDM.W	STVAR	HSTVAR	SVAR	LC	FDM.W	STVAR	HSTVAR	SVAR
1	0.0971	0.0623	**0.0572**	0.0578	0.0664	0.1414	0.0777	**0.0694**	0.0710	0.0806
2	0.1000	0.0554	0.0540	**0.0537**	0.0597	0.1540	0.0654	0.0618	**0.0604**	0.0724
3	0.1428	0.0838	0.0771	**0.0768**	0.0830	0.1879	**0.0982**	0.1025	0.1041	0.1045
4	0.1707	0.1027	0.0733	**0.0688**	0.1055	0.1611	0.0923	0.0915	**0.0897**	0.1086
5	0.1701	0.1042	0.0702	**0.0661**	0.1072	0.1621	0.0998	0.0998	**0.0947**	0.1083
6	0.1932	0.1150	0.0865	**0.0793**	0.1136	0.1780	0.1263	0.1238	**0.1168**	0.1347
7	0.2054	0.1284	0.0932	**0.0851**	0.1255	0.1868	0.1373	0.1358	**0.1243**	0.1369
8	0.2132	0.1351	0.0988	**0.0901**	0.1390	0.1927	0.1555	0.1577	**0.1435**	0.1649
9	0.2002	0.1220	0.1075	**0.1019**	0.1269	0.2184	0.2034	0.1798	**0.1612**	0.2018
10	0.2194	0.1379	0.1137	**0.1054**	0.1404	0.2495	0.2499	0.2123	**0.1893**	0.2470
11	0.2295	0.1482	0.1104	**0.1011**	0.1484	0.2504	0.2464	0.2192	**0.1954**	0.2370
12	0.2551	0.1637	0.1391	**0.1265**	0.1628	0.2646	0.2638	0.2374	**0.2088**	0.2598
13	0.2761	0.1919	0.1632	**0.1496**	0.1938	0.3023	0.3108	0.2602	**0.2308**	0.2982
14	0.2926	0.2009	0.1591	**0.1413**	0.2022	0.3006	0.3103	0.2538	**0.2253**	0.2951
15	0.2643	**0.1705**	0.1882	0.1775	0.1694	0.2826	0.2943	0.2513	**0.2184**	0.2702
16	0.2908	0.1984	0.1804	**0.1683**	0.2041	0.3038	0.3062	0.2548	**0.2271**	0.2870

Note: this table displays the RMSE over time horizons of the forecast French and Spanish mortality rates for 16 steps. Bold numbers represent the smallest RMSEs among the three models.

Please cite this article as: L. Feng, Y. Shi and L. Chang, Forecasting mortality with a hyperbolic spatial temporal VAR model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.05.003.

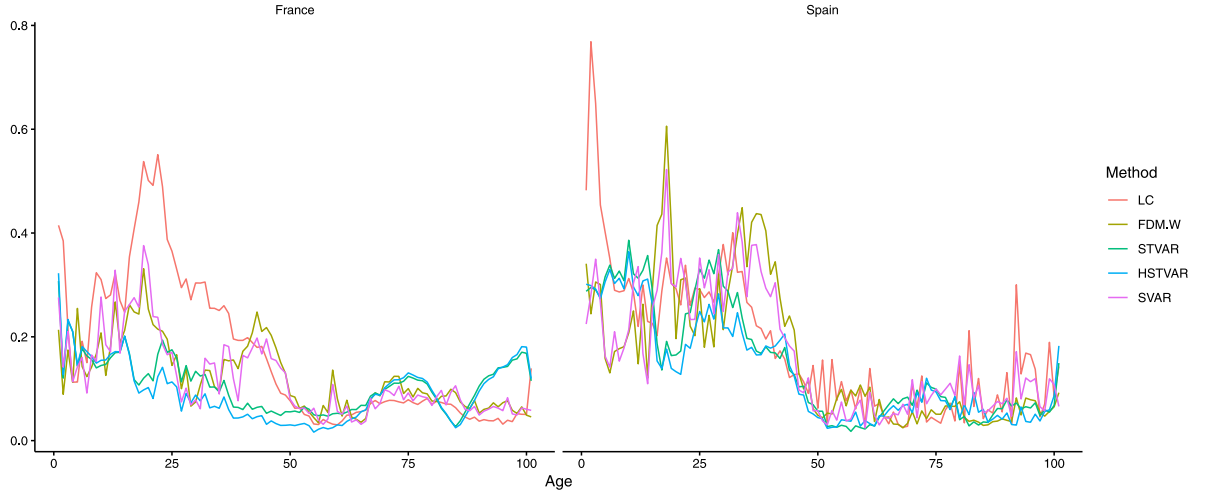


Fig. 5. RMSE over ages.

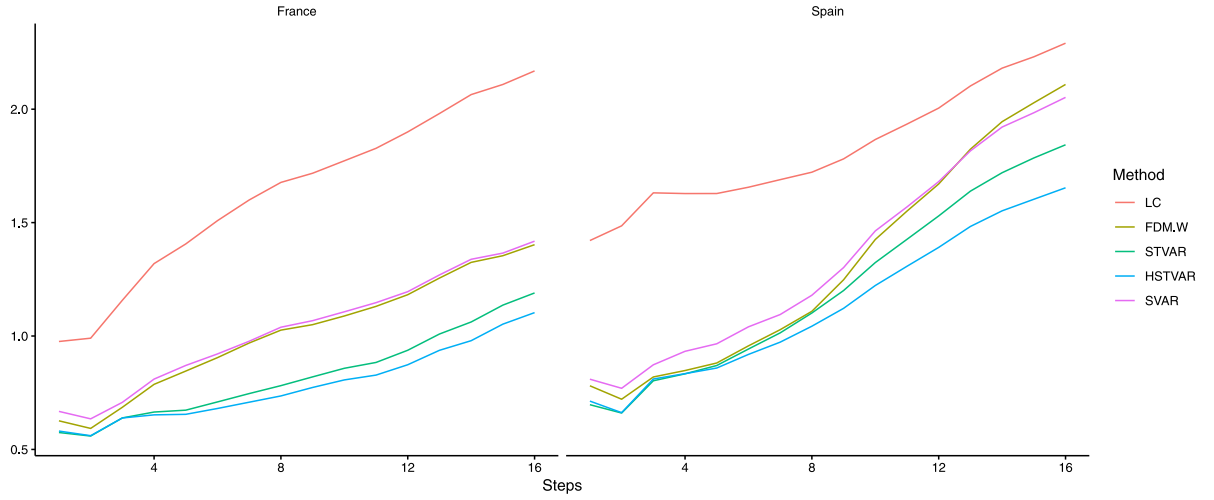


Fig. 6. $RMSE_{all,h}$.

In order to understand this superiority, we compared the forecast mortality rates of all models at the sixteenth step. Fig. 7 shows the five $\hat{y}_{x,2016}$ sequences together with the actual data in 2016. Due to the lack of a penalty scheme, LC, FDM.W and SVAR displayed some incoherent patterns and divergences among ages for both French and Spanish data. Overall, the LC, FDM.W and SVAR models over-forecasted the rates for age groups 20–50 in both cases. By contrast, forecasts of the STVAR and HSTVAR models did not diverge, as the other two did, and they better captured the variations among age groups. Although the results of the two spatial temporal VAR models were fairly similar to each other, those of the HSTVAR were still comparatively closer to the true data, especially for ages 20–40.

Finally, we plotted the forecast mortality rates averaged over all 101 age groups, as shown in Fig. 8. Both the true data (1950–2016) and the four forecast averages over 2001–2016 are displayed. For the French data, LC, FDM.W and SVAR clearly over-forecasted the rates at almost all steps. Despite the similarity, the forecast average rates of HSTVAR were still closer to the true data than those of the STVAR model. This result was largely consistent when applying the Spanish data, except that the LC under-forecasted the mortality rates over 2001–2010 and over-forecasted them for the remaining period.

5.2. Simulation results

To examine the superiority of the proposed HSTVAR model shown above, we followed Feng and Shi (2018) and performed simulation studies. For both French and Spanish data, we generated 1000 replicates. These replicates were generated by weighted penalized regression splines with a monotonicity constraint (Wood, 1994). This method is frequently used to produce smoothed mortality rates based on the crude data. We first fit the entire sample (1950–2000) via the regression splines. The residuals were then collected as the fitted rates for age

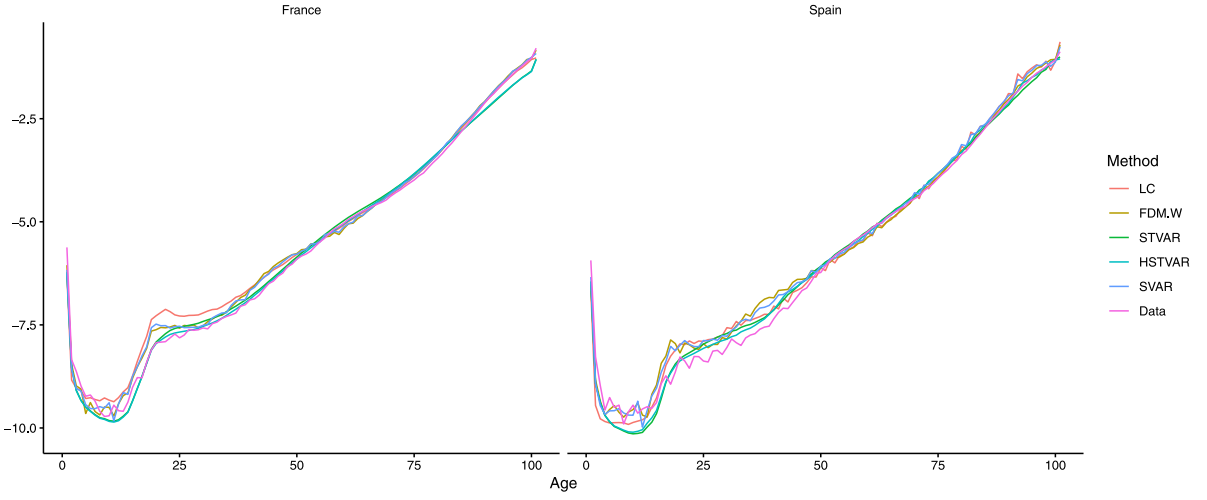


Fig. 7. Forecast vs. actual $\ln m_x$: 2016.

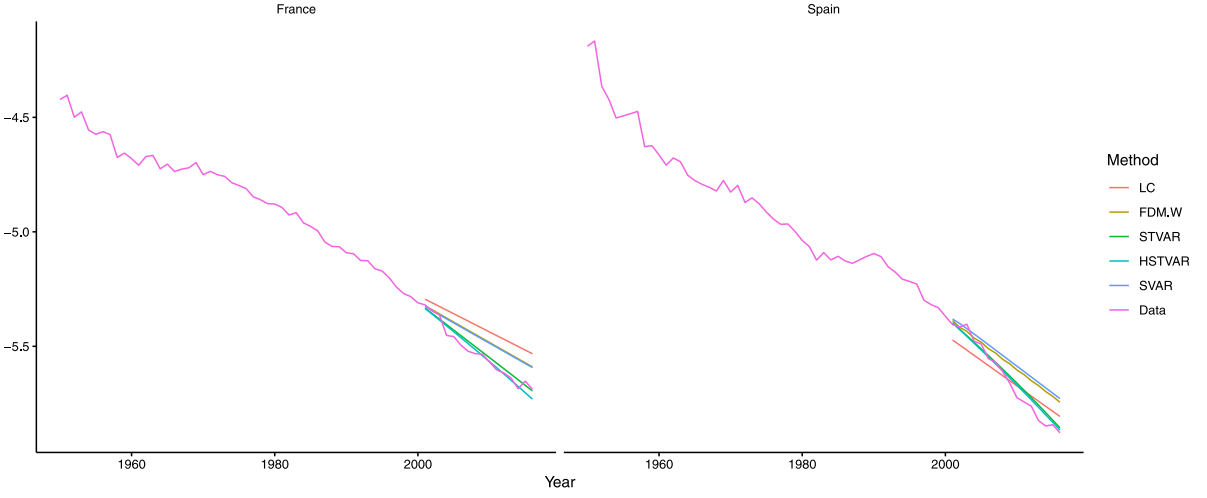


Fig. 8. Forecast vs. actual $\ln \bar{m}_t$ averaged over ages: 1950–2016.

x at time t subtracted from the true log of mortality rates. We assumed a multi-Gaussian distribution with sample mean and covariances of the collected residuals as the population mean and covariances, respectively, and 51×101 errors were then simulated for both French and Spanish data. Further, added to the fitted values of the corresponding splines, one complete simulated sequence was produced. This procedure was repeated until 1000 replicates were created. We then followed the same steps as in Section 5.1 to fit the LC, FDM.W, STVAR, HSTVAR and SVAR models[6] based on the simulated data. The 16-steps-ahead forecasts were finally generated in each case, and the $RMSE_{all,16}$ was calculated using the true sample of 2001–2016. The RMSEs are summarized in Table 4.

The descriptive statistics presented in Table 4 are largely consistent with our previous findings. Although the LC models demonstrated the smallest spread, the resulting mean $RMSE_{all,16}$ was much worse than the rest. Ranked second-best in terms of the standard deviation, HSTVAR performed the best in terms of the average, Q_1 and Q_3 statistics for $RMS_{all,16}$ for all the simulated replicates. To visually compare the simulation results of FDM.W and the three VAR-type models, we plotted the smoothed densities of those RMSEs in Fig. 9. For simulations of the French data, except for the similarity of the results of SVAR and FDM.W, the distributions of VAR-type models were clearly distinct from each other. As for the Spanish case, the distributions of STVAR and SVAR were more similar. Consistent with our observations in Table 4, the HSTVAR model had the narrowest spread with much smaller RMSEs distributed than the FDM.W, STVAR and SVAR models.

To sum up, by analyzing out-of-sample forecasts of the French and Spanish mortality data over 2001–2016, our proposed HSTVAR model consistently outperformed the

[6] For each replicate, the tuning parameters were kept the same as those described in Section 5.1.

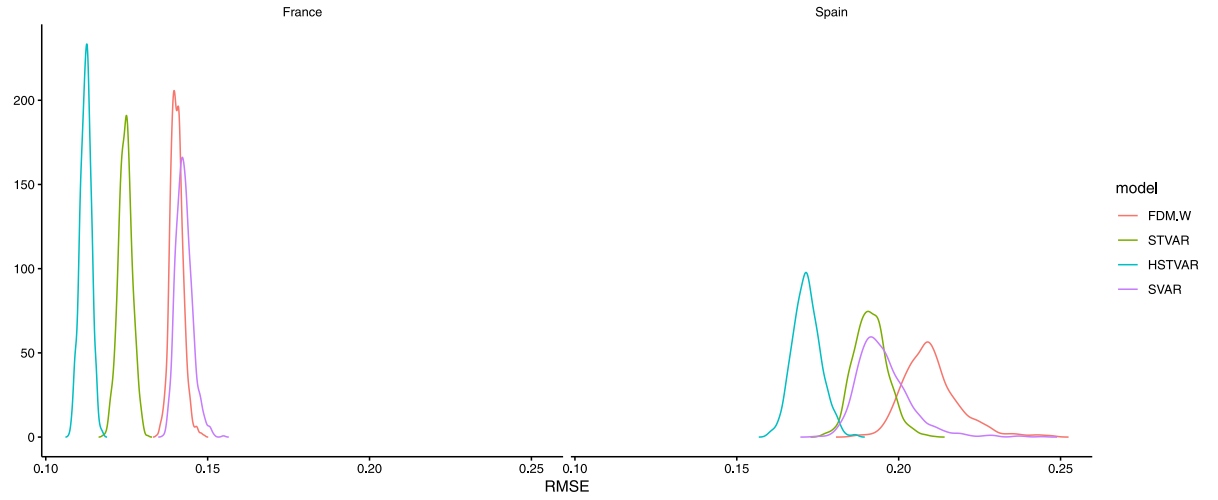


Fig. 9. Density plots of RMSEs of simulated results.

Table 4
Simulation summary.

Model	Mean	*Std. Dev.*	Q_1	Q_3
Panel A: France				
LC	0.2164	**0.0004**	0.2161	0.2167
FDM.W	0.1405	0.0020	0.1391	0.1416
STVAR	0.1246	0.0021	0.1231	0.1259
HSTVAR	**0.1123**	0.0017	**0.1112**	**0.1135**
SVAR	0.1427	0.0025	0.1409	0.1442
Panel B: Spain				
LC	0.2255	**0.0035**	0.2231	0.2278
FDM.W	0.2101	0.0088	0.2042	0.2142
STVAR	0.1914	0.0052	0.1878	0.1945
HSTVAR	**0.1716**	0.0043	**0.1687**	**0.1743**
SVAR	0.1952	0.0082	0.1899	0.1993

Note: this table displays the $RMSE_{all,16}$ of forecasts based on simulated French and Spanish total mortality rates. Mean, Std. Dev., Q_1 and Q_3 are the sample mean, standard deviation, first quartile and third quartile of the RMSEs over the 1000 simulated replicates.

LC, FDM.W, STVAR and SVAR counterparts. Its superiority was demonstrated via the RMSE over age groups and time horizons and the overall $RMSE_{all,16}$. The simulation results also provided robust conclusions. Thus, our proposed inclusion of the hyperbolic parameter to model the effects of younger cohorts is a powerful way to improve the forecasting accuracy of a VAR framework.

5.3. Long-term forecast analysis

We also examined the long-term forecasts of the five models. The complete French and Spanish datasets were employed (from 1950 to 2016), and the mortality rates were forecasted for the year 2050. For the two STVAR models, we employed the same tuning parameters presented in Table 1. Forecasted mortality rates in 2050 and the life expectancy from 2001–2050 are plotted in Figs. 10 and 11, respectively.

Compared with the actual mortality rates in 2016, the forecasted rates in 2050 suggest significant improvements in all cases. Due to the lack of a penalty scheme, LC, FDM.W and SVAR all showed significant divergences among most age groups for both French and Spanish data. By contrast, the forecasted rates of STVAR and HSTVAR were much more smoothed across all ages. Overall, the forecast rates of LC, FDM.W and SVAR were almost uniformly located higher than those of the STVAR models. For the very old ages, it is worth noticing that the forecast rates of LC, FDM.W and SVAR showed little improvement compared to the 2016 data. For the Spanish data, LC and SVAR forecasted unreasonably higher rates for the oldest age group (100-year-olds) than the corresponding rates in 2016. By contrast, both STVAR and HSTVAR produced more improvements to the mortality rates of old age groups. This is consistent with the desirable co-integration property of STVAR models.

Among all information generated from mortality rates, life expectancy is widely studied in demographic research and in actuarial practice. In Fig. 11, we report the actual French and Spanish life expectancy at birth from 2001 to 2016, together with mean/point forecasts (solid line) and 95% prediction intervals (dashed line) of forecasts up to 2050. The prediction intervals (PIs) of all models were calculated using the 2.5th and 97.5th percentiles with simulations of 1000 replicates. For instance, the $\hat{\boldsymbol{\varepsilon}}_{\mathbf{t+h}}$ of the STVAR and HSTVAR models were simulated using a fitted error structure, as discussed in Section 4.3. Specifically, at each forecast step h, these residuals were simulated from $\mathbf{N}(\mathbf{0}, \hat{\boldsymbol{\Sigma}})$. The h forecast $\hat{\mathbf{Y}}_{t+h}$ was then generated by $\hat{\mathbf{M}} + \hat{\mathbf{B}}\hat{\mathbf{Y}}_{t+h-1} + \hat{\boldsymbol{\varepsilon}}_{t+h}$, until all forecasts up to 2050 were produced.

Consistent with our observations in Fig. 10, e_0 forecasted by STVAR and HSTVAR was longer than that by the LC, FDM.W and SVAR models. In 2050, the mean forecasted e_0 for the French (Spanish) data was 90.7, 89.6, 87.6, 87.5 and 87.5 (89.7, 90.1, 88.1, 87.9 and 87.6) by the HSTVAR, STVAR, LC, FDM.W and SVAR models, respectively. For the French data, significant differences in the width of the PIs were observed. We found that the widths of the STVAR model were uniformly wider than those of the other models. Comparing HSTVAR and

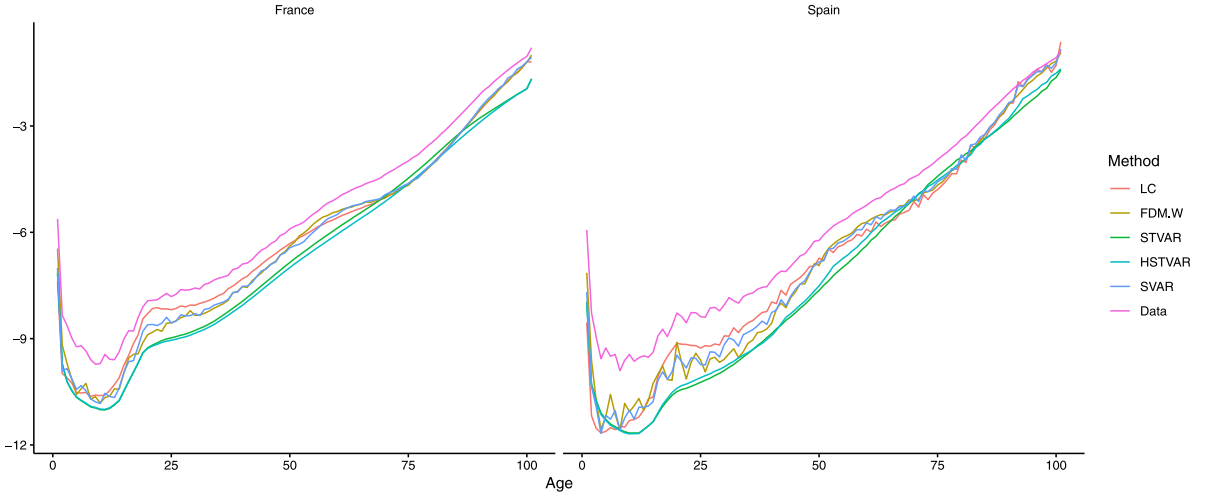


Fig. 10. Forecasted $\ln m_x$ in 2050.

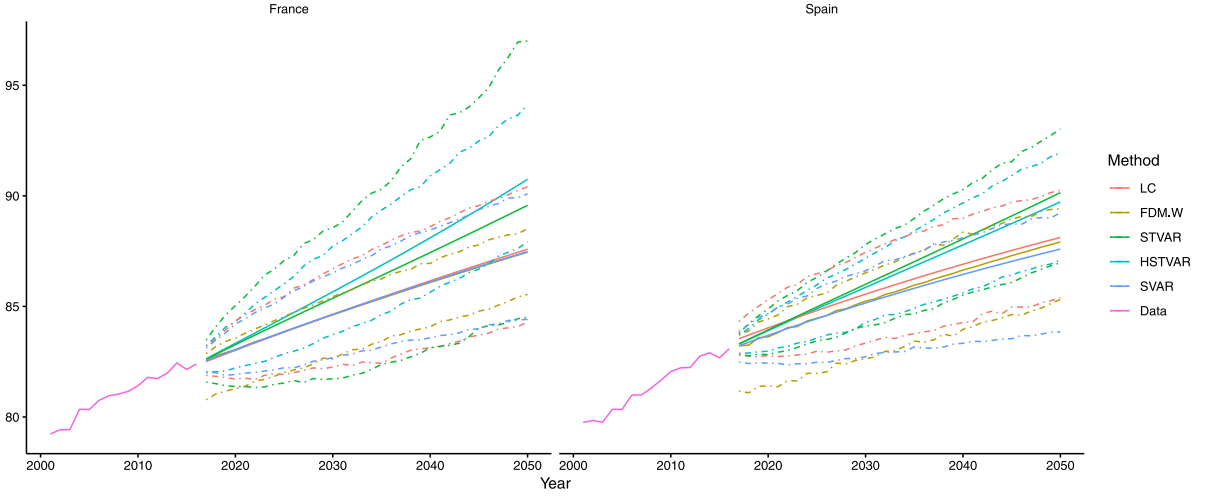


Fig. 11. Forecast vs. actual e_0: 2001–2050.

STVAR, for instance, the 95% PI in 2050 was (87.9, 94.1) and (84.5, 97.0), respectively. Thus, given the closeness of their forecast e_0, this might be evidence that the PIs of the HSTVAR are more efficient than those of the STVAR. The forecasted e_0 for the Spanish data obtained similar results. The widths of PIs generated by the STVAR model were much smaller than those in the French case. However, they were still uniformly wider than those of the HSTVAR model. Therefore, based on the similarity of the estimated life expectancy, our results suggest preliminary evidence that the proposed HSTVAR model can produce more efficient PIs than the STVAR model.

6. An extension to multi-population modeling and forecasting

Given global improvements in public health, medicine, transportation, and technology, it is widely recognized that the mortality rates of different populations tend to be correlated in some manner. Over the last two decades, a variety of multi-population mortality models have been developed to account for common patterns in large groups of populations (see, for example, Li and Lee (2005), Cairns, Blake, Dowd, Coughlan, and Khalaf-Allah (2011), and Li and Hardy (2011)). Li and Lu (2017) and Guibert et al. (2019) explored the possibility of extending the STVAR and SVAR models, respectively, to analyze multiple populations simultaneously. In this section, we briefly discuss the extension of our proposed HSTVAR to multi-population mortality modeling and forecasting.

6.1. Two-population extension

Following Li and Lu (2017), a two-population extension of the HSTVAR model can be specified as follows

Please cite this article as: L. Feng, Y. Shi and L. Chang, Forecasting mortality with a hyperbolic spatial temporal VAR model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.05.003.

(2-HSTVAR model).

$$\begin{aligned} y_{j,1,t} &= m_{j,1} + \rho_{j,1} y_{j,1,t-1} + (1-\rho_{j,1}) y_{-j,1,t-1} + \varepsilon_{j,1,t} \\ y_{j,2,t} &= m_{j,2} + \rho_{j,2}[(1-\beta_{j,2}) y_{j,2,t-1} + \beta_{j,2} y_{j,1,t-1}] \\ &\quad + (1-\rho_{j,1}) y_{-j,2,t-1} + \varepsilon_{j,2,t} \\ y_{j,i,t} &= m_{j,i} + \rho_{j,i}[(1-\beta_{j,i}) y_{j,i,t-1} + \beta_{j,i} y^{w}_{j,i,t-1}] \\ &\quad + (1-\rho_{j,i}) y_{-j,i,t-1} + \varepsilon_{j,i,t} \end{aligned} \tag{18}$$

where $i = 3, 4, \ldots, N$, $y^{w}_{j,i,t-1} = \sum_{k=1}^{i-1} w_{j,i,i-k} y_{j,k,t-1}$; $j = 1, 2$ is the jth population and $-j$ is the other population; and $\rho_{j,i}$ measures the impact of weighted younger cohorts on the mortality rate of age i from the same population j, whereas $1 - \rho_{j,i}$ measures the impact of the lagged mortality rate of age i from the other population $(-j)$ on the mortality rate of age i from the population j. This specification nests the extension of STVAR discussed in Li and Lu (2017) as a special case, when $w_{j,i,i-1} = 1$ for all i and j so that $y^{w}_{j,i,t-1} = y_{j,i-1,t-1}$. Also, we stipulate that $w_{j,i,k} = \delta_k(d_j)/\sum_{l=1}^{i-1} \delta_l(d_j)$, which essentially assigns different d_j to the jth population to measure the hyperbolic dependency of younger cohorts.

To estimate parameters ($m_{j,i}$ and $\beta_{j,i}$), the specification described in (18) contains four tuning parameters for each population, including the hyperbolic parameter d_{-j}. Besides, $\lambda_{j,m}$, $\lambda_{j,\beta}$ and $\lambda_{j,\rho}$ are smoothing penalties of $m_{j,i}$, $\beta_{j,i}$ and $\rho_{j,i}$, respectively. The stationarity property is discussed below.

Proposition 3. *Assume that* $0 < \beta_{j,i} < 1$ *and* $0 < \rho_{j,i} < 1$ *for* $i = 1, 2, \ldots, N$ *and* $j = 1, 2$. *The difference of any two components of the* $(N \times 2)$*-dimensional process* $y_{j,i,t}$ *defined by* (18) *is stationary.*

Proof. See Appendix A.3. □

In terms of estimation, the PLS procedure discussed in Section 4.2 can be straightforwardly applied here. The parameters for both populations are then estimated simultaneously. The eight tuning parameters need to be predetermined following the same cross-validation procedure described in Section 4.2. We can then use the same arguments in the proof for Proposition 2 to show that closed-form solutions exist for this 2-HSTVAR model. Forecasting can then be conducted in the usual way for VAR-type models, which is described in (8).

To obtain the uncertainty of the forecasts, we still need to understand the error structure of the 2-HSTVAR model. Similar to that discussed in Li and Lu (2017), the following specification can be employed.

$$\begin{aligned} \varepsilon_{j,1,t} &= \phi_{j,1} a_{j,1} \sum_{k=1}^{N-1} w^{U}_{j,1,k} \varepsilon_{j,1+k,t} + (1-\phi_{j,1}) \varepsilon_{-j,1,t} + \eta_{j,1,t} \\ \varepsilon_{j,i,t} &= \phi_{j,i}(a_{j,i} \sum_{k=1}^{N-i} w^{U}_{j,i,k} \varepsilon_{j,i+k,t} + c_{j,i} \sum_{l=1}^{i-1} w^{L}_{i,l} \varepsilon_{i-l,t}) \\ &\quad + (1-\phi_{j,i}) \varepsilon_{-j,i,t} + \eta_{j,i,t} \\ \varepsilon_{j,N,t} &= \phi_{j,N} c_{j,N} \sum_{l=1}^{N-1} w^{L}_{j,N,l} \varepsilon_{j,N-l,t} + (1-\phi_{j,N}) \varepsilon_{-j,N,t} + \eta_{j,N,t} \end{aligned} \tag{19}$$

Table 5
RMSE over ages summary: Multi-population extension.

Model	$RMSE_{all,16}$	Mean	p-val.	*Std. Dev.*	Q_1	Q_3
Panel A: France						
LL	0.1467	0.1171	0.0000	0.0889	**0.0602**	0.1383
CFDM.W	0.1374	0.1194	0.0000	0.0683	0.0710	0.1568
2-STVAR	0.1123	0.1000	0.0034	0.0514	0.0665	0.1239
2-HSTVAR	**0.1039**	**0.0931**	–	**0.0463**	0.0609	**0.1205**
2-SVAR	0.1402	0.1192	0.0000	0.0742	0.0657	0.1640
Panel B: Spain						
LL	0.2595	0.1907	0.0000	0.1768	**0.0410**	0.3819
CFDM.W	0.2064	0.1621	0.0200	0.1284	0.0554	0.2701
2-STVAR	0.1906	0.1517	0.0000	0.1160	0.0646	0.2076
2-HSTVAR	**0.1825**	**0.1467**	–	**0.1090**	0.0722	**0.1971**
2-SVAR	0.1982	0.1632	0.0380	0.1130	0.0697	0.2535

Note: this table displays the RMSE over age groups for the 16-steps-ahead forecasts of French and Spanish total mortality rates. $RMSE_{all,16}$ is the overall RMSE across all ages and time horizons. Mean, *Std. Dev.*, Q_1 and Q_3 are the sample mean, standard deviation, first quartile and third quartile of the RMSEs over age groups, respectively. Bold numbers represent the smallest RMSEs among the four models. p-val. is the p-value of the corresponding Diebold–Mariano test, which contrasts the forecasting performance of 2-HSTVAR against that of one of the benchmark models (i.e., LL, CFDM.W, 2-STVAR and 2-SVAR) individually. LL, CFDM.W 2-STVAR, 2-HSTVAR and 2-SVAR denote the Li–Lee model, coherent weighted functional demographic model and two-populations extension of the STVAR, HSTVAR and SVAR models, respectively.

where $1 < i < N$, $w^{U}_{j,i,k} = \delta_k(e_j)/\sum_{k=1}^{N-i} \delta_k(e_j)$ and $w^{L}_{j,i,l} = \delta_l(e_j)/\sum_{l=1}^{i-1} \delta_l(e_j)$, with $\delta_k(e_j)$ and $\delta_l(e_j)$ defined in the same ways as in (10). To ensure stationarity of $\varepsilon_{j,i,t}$, we also require that $a_{j,i} \geq 0$, $c_{j,i} \geq 0$, $\phi_{j,i} \geq 0$ and $a_{j,i} + c_{j,i} < 1$ for $j = 1, 2$ and $i = 1, 2, \ldots, N$. To obtain the estimates, we can use the same cross-validation process as described in Section 4.2 to select e_j and the smoothing penalties for each parameter first (four tuning parameter for each population). The PLS procedure described in Section 4.2 can then be employed. Since there are no equality constraints, this is a usual PLS problem and thus has closed-form solutions. PIs of the forecasts are to be produced via simulations, following the steps described in Section 5.3.

6.2. Illustration of the two-population empirical results

We employed the multi-population extension of the LC model, namely the Li–Lee (Li & Lee, 2005) model, the multi-population version of the FDM.W model, namely, the coherent functional demographic model (CFDM.W) investigated in Hyndman et al. (2013), the 2-STVAR model, the 2-HSTVAR model and the 2-SVAR model to fit the French and Spanish data simultaneously over 1950–2000.[7] Out-of-sample forecasts were then produced for 2001–2016. For our 2-HSTVAR model, the selected d_1 and d_2 for French and Spanish data were -0.91 and -0.99, respectively. This suggests the quickly reducing influence of the

[7] The specification, estimation and forecasting of the Li–Lee, CFDM.W and 2-SVAR models can be found in Li and Lee (2005), Hyndman et al. (2013) and Guibert et al. (2019), respectively. We omit descriptions of these for conciseness. CFDM.W also employs an ARIMA forecast method, as in FDM.W.

Please cite this article as: L. Feng, Y. Shi and L. Chang, Forecasting mortality with a hyperbolic spatial temporal VAR model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.05.003.

younger cohorts among the spatial dimension, given the impact of their own lag.

Similar to Table 2, the descriptive statistics of $RMSE_x$ and $RMSE_{all,16}$ are reported in Table 5 for illustration purposes. Consistent with our observations in Section 5, the 2-HSTVAR model had the best overall forecasting performance, indicated by $RMSE_{all,16}$, in all scenarios. Moreover, our proposed approach also resulted in the narrowest spread and smallest average *RMSE* over the age groups. At the 5% level, the DM test suggested that the 2-HSTVAR model significantly outperformed the other four competing models. These results support the superiority of the hyperbolic structure of the cohorts investigated in this paper. Other relevant outputs are largely consistent with those in Tables 2–3 and Figs. 5–8, which are available upon request.

6.3. Multi-population extension

Similar to the 2-HSTVAR model, we can derive the J-HSTVAR model in the case of $J > 2$ populations. Following Li and Lu (2017), this can be realized by the following specification, which is largely consistent with (18).

$$\begin{aligned}
y_{j,1,t} &= m_{j,1} + \sum_{k=1}^{J} \rho_{j,k,1} y_{k,1,t-1} + \varepsilon_{j,1,t} \\
y_{j,2,t} &= m_{j,2} + \rho_{j,j,2}[(1-\beta_{j,2})y_{j,2,t-1} + \beta_{j,2} y_{j,1,t-1}] \\
&\quad + \sum_{k\neq j}^{J} \rho_{j,k,2} y_{k,2,t-1} + \varepsilon_{j,2,t} \\
y_{j,i,t} &= m_{j,i} + \rho_{j,j,i}[(1-\beta_{j,i})y_{j,i,t-1} + \beta_{j,i} y^{w}_{j,i,t-1}] \\
&\quad + \sum_{k\neq j}^{J} \rho_{j,k,i} y_{k,i,t-1} + \varepsilon_{j,i,t}
\end{aligned} \tag{20}$$

where $\sum_{k=1}^{J} \rho_{j,k,i} = 1$ for all $j = 1, 2, \ldots, J$ and each $i = 1, 2, \ldots, N$. $y^{w}_{j,i,t-1}$ is defined in the same way as in (18). Since for each equation, the coefficients of independent variables all sum up to 1, the proof in Appendix A.2 of Li and Lu (2017) directly applies to our case. That is, the difference of any two components of the $(N \times J)$-dimensional process $y_{j,i,t}$ for $j = 1, 2, \ldots, J$ and $i = 1, 2, \ldots, N$ is stationary.

The error structure defined in (19) can then be straightforwardly modified to

$$\begin{aligned}
\varepsilon_{j,1,t} &= \phi_{j,j,1} a_{j,1} \sum_{r=1}^{N-1} w^{U}_{j,1,k} \varepsilon_{j,1+k,t} + \sum_{r\neq j}^{J} \phi_{j,r,1} \varepsilon_{r,1,t} + \eta_{j,1,t} \\
\varepsilon_{j,i,t} &= \phi_{j,j,i}(a_{j,i} \sum_{r=1}^{N-i} w^{U}_{j,i,k} \varepsilon_{j,i+k,t} + c_{j,i} \sum_{l=1}^{i-1} w^{L}_{i,l} \varepsilon_{i-l,t}) \\
&\quad + \sum_{r\neq j}^{J} \phi_{j,r,i} \varepsilon_{r,i,t} + \eta_{j,i,t} \\
\varepsilon_{j,N,t} &= \phi_{j,j,N} c_{j,N} \sum_{l=1}^{N-1} w^{L}_{j,N,l} \varepsilon_{j,N-l,t} + \sum_{r\neq j}^{J} \phi_{j,r,N} \varepsilon_{r,N,t} + \eta_{j,N,t}
\end{aligned} \tag{21}$$

where $\sum_{r=1}^{J} \phi_{j,r,i} = 1$ for all $r = 1, 2, \ldots, J$ and each $i = 1, 2, \ldots, N$. $w^{U}_{j,i,k}$ and $w^{L}_{j,N,l}$ are defined in the same way as in (19).

In terms of the estimation, the same strategy described in Section 6.1 for the 2-HSTVAR model can be directly applied here. There will be four tuning parameters for each population for both (20) and (21), including one hyperbolic parameter (d_j and e_j) as three smoothing penalties ($\lambda_{j,m}$, $\lambda_{j,\beta}$ and $\lambda_{j,\rho}$ for (20) and ($\lambda_{j,a}$, $\lambda_{j,c}$ and $\lambda_{j,\phi}$).[8] These parameters are to be chosen via cross-validation, and PLS can then be performed. Closed-form solutions exist for both (20) and (21), for the same reasons as those described in Section 6.1. At each out-of-sample time-step, forecasts will be produced for all J mortality rates simultaneously, as for the 2-HSTVAR model. PIs can then be simulated from estimated error structures using the multi-Gaussian assumption.

7. Concluding remarks

In this paper, we proposed an effective hyperbolic spatial temporal VAR model (HSTVAR) to investigate and forecast log mortality rates. Three key results can be drawn from our study. First, the proposed HSTVAR model is effective at mortality forecasting. As measured by the RMSE, our HSTVAR model outperformed the famous Lee–Cater model (Lee & Carter, 1992), weighted functional demographic model (or FDM.W) (Shang et al., 2011), spatial–temporal VAR model (Li & Lu, 2017) and sparse VAR model (Guibert et al., 2019). This conclusion consistently held for both French and Spanish total populations when samples from 1950 to 2000 were fitted and forecasts from 2001 to 2016 were produced. Second, the HSTVAR model has a more flexible framework that retains all advantages of the original spatial–temporal VAR (STVAR) model investigated by Li and Lu (2017). Using a hyperbolic memory structure, which is widely employed in financial time series analysis, our HSTVAR model allows for the spatially decayed influence of younger cohorts. Compared with STVAR, almost no additional computational cost is required, and all desirable properties still hold, including stationarity (co-integration), closed-form solutions and a simple error structure. More importantly, with this more flexible spatial structure, the forecasting accuracy improved in all cases over the STVAR model. Finally, in contrast to standard factor models, such as the Lee–Carter model, the extension of STVAR to multi-population mortality modeling is straightforward. Our proposed HSTVAR model outperformed the extended Lee–Carter model (i.e., the Li–Lee model) (Li & Lee, 2005), the extension of the FDM.W (i.e., the weighted coherent functional demographic model) (Hyndman et al., 2013), and the two-population STVAR and sparse VAR models with the French and Spanish data. Theoretical properties still hold for these multi-population extensions, including stationarity, closed-form solutions and a parametric error structure. Thus, our results demonstrate the effectiveness

[8] For simplicity, we only allow one smoothing penalty for $\rho_{j,k,i}$ ($i = 1, 2, \ldots, N$ and $k = 1, 2, \ldots, J$) and all $\phi_{j,r,i}$ ($i = 1, 2, \ldots, N$ and $r = 1, 2, \ldots, J$), respectively.

Please cite this article as: L. Feng, Y. Shi and L. Chang, Forecasting mortality with a hyperbolic spatial temporal VAR model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.05.003.

and usefulness of HSTVAR for mortality modeling and forecasting.

There are also some pathways for future research. First, the effects of older cohorts (e.g., $x+1$ on x) may be allowed for a more flexible structure of the coefficient matrix $\boldsymbol{B}$. Second, for a negative d, it would be of interest to consider only the number of younger cohorts that are significant in mortality forecasting. Third, a variant (e.g., age-specific) hyperbolic parameter may be developed to replace the current constant case. Also, a selection criterion of lags considered in the HSTVAR model could be investigated. In addition, extensions may be developed to consider the potential heteroskedasticity of mortality rates.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank the Jiangxi University of Finance and Economics, Macquarie University and the Australian National University for research support. In particular, the authors would like to thank the Editor (Pierre Pinson) and three anonymous referees for providing valuable and insightful comments on earlier drafts. The usual disclaimer applies.

Appendix

A.1. Proof of Proposition 1

For the smallest age 1, $y_{1,t}$ is clearly a random work with drift m_1.

For age 2, using the expression of $y_{2,t}$ to subtract that of $y_{1,t}$, we have that

$$y_{2,t}-y_{1,t} = m_2-m_1+(1-b_{21})(y_{2,t-1}-y_{1,t-1})+\varepsilon_{2,t}-\varepsilon_{1,t}.$$

Hence, $y_{2,t}-y_{1,t}$ is stationary.

For age 3, in a similar fashion, we have that

$$\begin{aligned} y_{3,t}-y_{2,t} = {} & m_3-m_2+(1-b_{32}-b_{31})(y_{3,t-1}-y_{2,t-1}) \\ & -(b_{31}-b_{21})(y_{2,t-1}-y_{1,t-1})+\varepsilon_{3,t}-\varepsilon_{2,t}. \end{aligned}$$

Hence, $y_{3,t}-y_{2,t}$ is stationary.

Following the same induction, for age i ($3 < i \le N$), we have that

$$\begin{aligned} y_{i,t}-y_{i-1,t} = {} & m_i-m_{i-1}+(1-\sum_{l=1}^{i-1} b_{il})(y_{i,t-1}-y_{i-1,t-1}) \\ & -\sum_{j=2}^{i-1}\left[\sum_{k=1}^{j-1}(b_{i,k}-b_{i-1,k})\right](y_{j,t-1} \\ & -y_{j-1,t-1})+\varepsilon_{i,t}-\varepsilon_{i-1,t}. \end{aligned}$$

and all those $y_{j,t}-y_{j-1,t}$ where $1 < j < i$ are stationary. Hence, $y_{i,t}-y_{i-1,t}$ is also stationary, which completes the proof.

A.2. Proof of Proposition 2

From (14), we have that for $1 < i \le N$,

$$\begin{aligned} y_{i,t} &= m_i+(1-\beta_i)y_{i,t-1}+\sum_{l=1}^{i-1}\beta_i w_{i,i-l}y_{l,t-1}+\varepsilon_{i,t} \\ y_{i,t} &= m_i+(1-\beta_i)y_{i,t-1}+\beta_i y^w_{i,t-1}+\varepsilon_{i,t} \\ y_{i,t}-y_{i,t-1} &= m_i+\beta_i(y^w_{i,t-1}-y_{i,t-1})+\varepsilon_{i,t} \end{aligned}$$

where $y^w_{i,t-1} = \sum_{l=1}^{i-1} w_{i,i-l}y_{l,t-1}$. Hence, the constrained PLS problem of $y_{i,t}$ can be transformed to an equivalent non-constrained PLS problem of $y^w_{i,t-1}-y_{i,t-1}$ regressed against $y_{i,t}-y_{i,t-1}$.

It is then straightforward to rewrite (14) in the following matrix form.

$$LF_1 = (\triangle\mathbf{y}-\boldsymbol{X\theta})'(\triangle\mathbf{y}-\boldsymbol{X\theta})+\lambda_m\boldsymbol{\theta}'\boldsymbol{S_M}\boldsymbol{\theta}+\lambda_\beta\boldsymbol{\theta}'\mathbf{S}_{\boldsymbol{\beta}}\boldsymbol{\theta}$$

where $\triangle\mathbf{y} = (\triangle\mathbf{y_{1,t}}, \triangle\mathbf{y_{2,t}}, \ldots, \triangle\mathbf{y_{N,t}})'_{N(T-1)\times 1}$, $\triangle\mathbf{y_{i,t}} = (y_{i,2}-y_{i,1}, y_{i,3}-y_{i,2}, \ldots, y_{i,T}-y_{i,T-1})'_{(T-1)\times 1}$, $\boldsymbol{\theta} = (M, \boldsymbol{\beta})'_{(2N-1)\times 1}$ and see Box II.

Since both $\boldsymbol{S_M}$ and $\mathbf{S}_{\boldsymbol{\beta}}$ are symmetric and $\boldsymbol{\theta}$ is the only unknown value, we have that

$$\frac{dLF_1}{d\boldsymbol{\theta}} = -2\boldsymbol{X}'\triangle\mathbf{y}+2\boldsymbol{X}'\boldsymbol{X\theta}+2\lambda_m\boldsymbol{S_M\theta}+2\lambda_\beta\mathbf{S}_{\boldsymbol{\beta}}\boldsymbol{\theta}.$$

Setting it to 0, we have the estimated parameter vector

$$\hat{\boldsymbol{\theta}} = [\boldsymbol{X}'\boldsymbol{X}+\lambda_m\boldsymbol{S_M}+\lambda_\beta\mathbf{S}_{\boldsymbol{\beta}}]^{-1}\boldsymbol{X}'\triangle\mathbf{y},$$

which completes the proof.

A.3. Proof of Proposition 3

Following the proof described in Appendix A.1 of Li and Lu (2017), $y_{j,1,t}-y_{-j,1,t}$, $y_{j,2,t}-y_{j,1,t}$, $y_{j,2,t}-y_{-j,1,t}$ and $y_{j,2,t}-y_{-j,2,t}$ are stationary.

Rewrite that

$$\begin{aligned} y_{j,i,t} = {} & m_{j,i}+\rho_{j,i}[(1-\sum_{k=1}^{i-1} b_{j,i,k})y_{j,i,t-1}+\sum_{l=1}^{i-1} b_{j,i,l}y_{j,l,t-1}] \\ & +(1-\rho_{j,i})y_{-j,i,t}+\varepsilon_{j,i,t} \end{aligned}$$

where $b_{j,i,l} = \beta_{j,i}w_{j,i,i-l}$, and therefore $0 < b_{j,i,l} < 1$ for $i = 1, 2, \ldots, N$, $j = 1, 2$ and $l = 1, 2, \ldots, i-1$.

Then,

$$\begin{aligned} y_{j,3,t}-y_{j,2,t} = {} & [\rho_{j,3}(1-b_{j,3,2}-b_{j,3,1})](y_{j,3,t-1}-y_{j,2,t-1}) \\ & +(1-\rho_{j,3})(y_{-j,3,t-1}-y_{-j,2,t-1}) \\ & +(\rho_{j,2}b_{j,2,1}-\rho_{j,3}b_{j,3,1})(y_{j,2,t-1}-y_{j,1,t-1}) \\ & +(\rho_{j,3}-\rho_{j,2})(y_{j,2,t-1}-y_{-j,2,t-1}) \\ & +m_{j,3}-m_{j,2}+\varepsilon_{j,3,t}-\varepsilon_{j,2,t} \end{aligned}$$

Therefore, for the bivariate VAR process ($y_{j,3,t}-y_{j,2,t}$, $y_{-j,3,t}-y_{-j,2,t}$), we need to show that, with the autoregressive matrix

$$\mathbf{A_3} = \begin{bmatrix} \rho_{j,3}(1-b_{j,3,2}-b_{j,3,1}) & 1-\rho_{j,3} \\ 1-\rho_{-j,3} & \rho_{-j,3}(1-b_{-j,3,2}-b_{-j,3,1}) \end{bmatrix},$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 & \cdots & \cdots & \cdots \\ 1 & 0 & \cdots & \cdots & 0 & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 1 & 0 & \cdots & \cdots & y^{w}_{2,1} - y_{2,1} & 0 & \cdots \\ 0 & 1 & 0 & \cdots & \cdots & y^{w}_{2,2} - y_{2,2} & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 1 & 0 & \cdots & \cdots & y^{w}_{N,1} - y_{N,1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 1 & 0 & \cdots & \cdots & y^{w}_{N,T} - y_{N,T} \end{bmatrix}_{N(T-1)\times(2N-1)},$$

$$\boldsymbol{S_M} = \begin{bmatrix} 1 & -1 & 0 & \cdots & \cdots & \cdots & \cdots \\ -1 & 2 & -1 & 0 & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & -1 & 2 & -1 & 0 & \cdots \\ 0 & \cdots & \cdots & -1 & 1 & 0 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{bmatrix}_{(2N-1)\times(2N-1)}$$

and

$$\mathbf{S}_{\boldsymbol{\beta}} = \begin{bmatrix} 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & 1 & -1 & 0 & \cdots \\ 0 & \cdots & \cdots & -1 & 2 & -1 & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & \cdots & \cdots & -1 & 2 & -1 \\ 0 & \cdots & \cdots & \cdots & \cdots & -1 & 1 \end{bmatrix}_{(2N-1)\times(2N-1)}$$

Box II.

the polynomial $\det(\mathbf{A_3} - z\boldsymbol{I})$ has no roots outside the unit circle $|z| \geq 1$. This is satisfied, since

$$\begin{aligned} &|[\rho_{j,3}(1 - b_{j,3,2} - b_{j,3,1}) - z] \\ &\quad \times [\rho_{-j,3}(1 - b_{-j,3,2} - b_{-j,3,1}) - z]| \\ &\geq |[1 - \rho_{j,3}(1 - b_{j,3,2} - b_{j,3,1})] \\ &\quad \times [1 - \rho_{-j,3}(1 - b_{-j,3,2} - b_{-j,3,1})]| \\ &\geq (1 - \rho_{j,3})(1 - \rho_{-j,3}) \end{aligned}$$

for all $|z| \geq 1$.

Following the same induction, it can be verified that for all bivariate VAR processes $(y_{j,k,t} - y_{j,l,t}, y_{-j,k,t} - y_{-j,l,t})$ and $(y_{j,k,t} - y_{-j,l,t}, y_{-j,k,t} - y_{j,l,t})$, where $N \geq k > l \geq 1$, the corresponding autoregressive matrix is

$$\mathbf{A_k} = \begin{bmatrix} \rho_{j,k}(1 - \sum_{r=1}^{k-1} b_{j,k,r}) & 1 - \rho_{j,k} \\ 1 - \rho_{-j,k} & \rho_{-j,k}(1 - \sum_{s=1}^{k-1} b_{-j,k,r}) \end{bmatrix},$$

for which the polynomial $\det(\mathbf{A_k} - z\boldsymbol{I})$ has no roots outside the unit circle $|z| \geq 1$, following the case of $\mathbf{A_3}$ shown above.

As for the process of $y_{j,k,t} - y_{j,k,t}$, where $2 < k \leq N$, we can demonstrate that

$$\begin{aligned} y_{j,k,t} - y_{-j,k,t} = &[\rho_{j,k}(1 - \sum_{r=1}^{k-1} b_{j,k,r}) + \rho_{-j,k} - 1] \\ &\times (y_{j,k,t-1} - y_{-j,k,t-1}) + poly(t-1) \\ &+ m_{j,k} - m_{-j,k} + \varepsilon_{j,k,t} - \varepsilon_{-j,k,t} \end{aligned}$$

where $poly(t-1)$ consists of polynomials of bivariate components of the $((k-1)\times 2)$-dimensional process $y_{j,i,t}$, which are stationary by the induction shown above. Then, $y_{j,k,t} - y_{-j,k,t}$ is clearly stationary, which completes the proof.

A.4. Computational packages

We used the software **R** to perform the computations of all models. SVAR was implemented via the **sparsevar** package. LC, FDM.W and CFDM.W were estimated using the **demography** package. LL was fitted by the **MortalityForecast** package. STVAR and HSTVAR were computed with code written by the authors.

References

Baillie, R. T., Bollerslev, T., & Mikkelsen, H. O. (1996). Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *74*(1), 3–30.

Barrieu, P., Bensusan, H., El Karoui, N., Hillairet, C., Loisel, S., Ravanelli, C., & Salhi, Y. (2012). Understanding, modelling and managing longevity risk: key issues and main challenges. *Scandinavian actuarial journal*, *2012*(3), 203–231.

Biffis, E., & Millossovich, P. (2006). A bidimensional approach to mortality risk. *Decisions in Economics and Finance*, *29*(2), 71–94.

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

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., & Khalaf-Allah, M. (2011). Bayesian stochastic mortality modelling for two populations. *Astin Bulletin*, *41*, 29–59.

Choi, K., Yu, W.-C., & Zivot, E. (2010). Long memory versus structural breaks in modeling and forecasting realized volatility. *Journal of International Money and Finance*, *29*(5), 857–875.
Davidson, J. (2004). Moment and memory properties of linear conditional heteroscedasticity models, and a new model. *Journal of Business & Economic Statistics*, *22*(1), 16–29.
Debón, A., Montes, F., Mateu, J., Porcu, E., & Bevilacqua, M. (2008). Modelling residuals dependence in dynamic life tables: A geostatistical approach. *Computational Statistics & Data Analysis*, *52*(6), 3128–3147.
Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *20*(1), 134–144.
Feng, L., & Shi, Y. (2017). Fractionally integrated garch model with tempered stable distribution: a simulation study. *Journal of Applied Statistics*, *44*(16), 2837–2857.
Feng, L., & Shi, Y. (2018). Forecasting mortality rates: Multivariate or univariate models? *Journal of Population Research*, *35*(3), 289–318.
Gao, G., Ho, K.-Y., & Shi, Y. (2020). Long memory or regime switching in volatility? Evidence from high-frequency returns on the US stock indices. *Pacific-Basin Finance Journal*, *61*, 101059.
Guibert, Q., Lopez, O., & Piette, P. (2019). Forecasting mortality rate improvements with a high-dimensional VAR. *Insurance: Mathematics and Economics*, *88*, 255–272.
Ho, K.-Y., & Shi, Y. (2020). Discussions on the spurious hyperbolic memory in the conditional variance and a new model. *Journal of Empirical Finance*, *55*, 83–103.
Hosking, J. R. M. (1981). Fractional differencing. *Biometrica*, *68*(1), 165–176.
Human Mortality Database (2019). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). http://www.mortality.org.
Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.
Hyndman, R. J., & Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. *International Journal of Forecasting*, *24*(3), 323–342.
Hyndman, R. J., Booth, H., & Yasmeen, F. (2013). Coherent mortality forecasting: the product-ratio method with functional time series models. *Demography*, *50*(1), 261–283.
Hyndman, R., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *27*(3).
Hyndman, R. J., & Shang, H. L. (2009). Forecasting functional time series. *Journal of the Korean Statistical Society*, *38*(3), 199–211.
Hyndman, R. J., & Ullah, M. S. (2007). Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics & Data Analysis*, *51*(10), 4942–4956.
Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American statistical association*, *87*(419), 659–671.
Li, J. S.-H., & Hardy, M. R. (2011). Measuring basis risk in longevity hedges. *North American Actuarial Journal*, *15*(2), 177–200.
Li, N., & Lee, R. (2005). Coherent mortality forecasts for a group of populations: an extension of the Lee-Carter method. *Demography*, *42*(3), 575–594.
Li, H., & Lu, Y. (2017). Coherent forecasting of mortality rates: A sparse vector-autoregression approach. *ASTIN Bulletin: The Journal of the IAA*, *47*(2), 563–600.
Ramsay, J. O., & Silverman, B. W. (2007). *Applied functional data analysis: Methods and case studies*. Springer.
Ramsay, J. O., et al. (1988). Monotone regression splines in action. *Statistical Science*, *3*(4), 425–441.
Renshaw, A. E., & Haberman, S. (2006). A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics*, *38*(3), 556–570.
Shang, H. L., Booth, H., & Hyndman, R. J. (2011). Point and interval forecasts of mortality rates and life expectancy: A comparison of ten principal component methods. *Demographic Research*, *25*, 173–214.
Shang, H. L., & Hyndman, R. J. (2017). Grouped functional time series forecasting: An application to age-specific mortality rates. *Journal of Computational and Graphical Statistics*, *26*(2), 330–343.
Wood, S. N. (1994). Monotonic smoothing splines fitted by cross validation. *SIAM Journal on Scientific Computing*, *15*(5), 1126–1133.

Please cite this article as: L. Feng, Y. Shi and L. Chang, Forecasting mortality with a hyperbolic spatial temporal VAR model. International Journal of Forecasting (2020), https://doi.org/10.1016/j.ijforecast.2020.05.003.